Journal of Science Education and Technology (2022) 31:490–513
https://doi.org/10.1007/s10956-022-09969-w

Bridging the Gap Between Qualitative and Quantitative Assessment in Science Education Research with Machine Learning — A Case for Pretrained Language Models-Based Clustering

Peter Wulff[1] · David Buschhüter[2] · Andrea Westphal[3] · Lukas Mientus[2] · Anna Nowak[2] · Andreas Borowski[2]

Accepted: 2 May 2022 / Published online: 1 June 2022
© The Author(s) 2022

Abstract
Science education researchers typically face a trade-off between more quantitatively oriented confirmatory testing of hypotheses, or more qualitatively oriented exploration of novel hypotheses. More recently, open-ended, constructed response items were used to combine both approaches and advance assessment of complex science-related skills and competencies. For example, research in assessing science teachers' noticing and attention to classroom events benefitted from more open-ended response formats because teachers can present their own accounts. Then, open-ended responses are typically analyzed with some form of content analysis. However, language is noisy, ambiguous, and unsegmented and thus open-ended, constructed responses are complex to analyze. Uncovering patterns in these responses would benefit from more principled and systematic analysis tools. Consequently, computer-based methods with the help of machine learning and natural language processing were argued to be promising means to enhance assessment of noticing skills with constructed response formats. In particular, pretrained language models recently advanced the study of linguistic phenomena and thus could well advance assessment of complex constructs through constructed response items. This study examines potentials and challenges of a pretrained language model-based clustering approach to assess preservice physics teachers' attention to classroom events as elicited through open-ended written descriptions. It was examined to what extent the clustering approach could identify meaningful patterns in the constructed responses, and in what ways textual organization of the responses could be analyzed with the clusters. Preservice physics teachers ($N = 75$) were instructed to describe a standardized, video-recorded teaching situation in physics. The clustering approach was used to group related sentences. Results indicate that the pretrained language model-based clustering approach yields well-interpretable, specific, and robust clusters, which could be mapped to physics-specific and more general contents. Furthermore, the clusters facilitate advanced analysis of the textual organization of the constructed responses. Hence, we argue that machine learning and natural language processing provide science education researchers means to combine exploratory capabilities of qualitative research methods with the systematicity of quantitative methods.

Keywords Attention to classroom events · Noticing · NLP · ML

"You can have data without information, but you cannot have information without data." (Daniel Keys Moran)

Research methods in science education are commonly differentiated into quantitative and qualitative methods (Krüger et al., 2014). The former allow for the confirmatory testing of statistical hypotheses, whereas the latter allow for more exploratory generation of novel hypotheses. This division is artificial and attributes to the imperfect capabilities of modeling complex systems that involve learning processes of humans. It would be desirable to better integrate both methods and conserve the predictive capabilities of quantitative methods and the exploratory capabilities of qualitative methods. It has been suggested that complex algorithmic approaches such as machine learning can better model assessment in science education (Breiman, 2001; Zhai,

✉ Peter Wulff
peter.wulff@ph-heidelberg.de

[1] Physics Education Research Group, Heidelberg University of Education, Heidelberg, Germany

[2] Physics Education Research Group, University of Potsdam, Potsdam, Germany

[3] Department of Educational Research, University of Greifswald, Greifswald, Germany

2021) and eventually provide a new methods paradigm. "Machine learning is about inductively solving problems by machines, i.e., computers." (Rauf, 2021, p.8). Inductive learning requires appropriate data for the machines to improve on relevant tasks. Given advances in data storage and accessibility, machine learning (ML) models dramatically improved their performance on many tasks such as image classification, or spoken and written language analytics (Goodfellow et al., 2016; Goldberg, 2017). Scholars in the fields of education and discipline-based educational research also argued that ML methods can advance educational research (Singer, 2019; Baig et al., 2020), even "revolutionize" assessments (Zhai et al., 2020). Among others, the education sector presents a field where datasets of unprecedented size become available (Baig et al., 2020).

ML methods have been utilized in science education research in different contexts. Mostly, science education researchers employed supervised ML methods where a model is trained to map responses to predefined outputs (Zhai et al., 2020). However, oftentimes problems in science education research are less well defined and only small datasets can be collected with reasonable effort. For example, in research on university-based teacher education such as noticing and attention to classroom events, typically small samples are available (Chan et al., 2021; Wilson et al., 2019). Noticing, among others, comprises the careful observation of events in a teaching situation. In science education research it has been highlighted that preservice science and mathematics teachers attend to many different events and contents in a teaching situation (Talanquer et al., 2015). To capture the complexity of noticing, science education researchers therefore used open-ended, constructed response formats to assess noticing and attention to classroom events (Barth-Cohen et al., 2018; Luna et al., 2018; Chan et al., 2021). The responses are then analyzed with some form of content analysis. However, not only do the differences in attention between the teachers yield to the complexity of assessing noticing and attention processes, but also the teachers' use of language in constructed-response items. Language use was characterized to be "noisy, ambiguous, und unsegmented" (Jurafsky, 2003, p.39). Hence, probabilistic approaches are required to analyze language-related processes and products. A probabilistic approach that also captures complexity is ML. The application of ML-based modeling could provide researchers means to gain novel insights into these complex constructs (Zhai et al., 2020). Yet, it is not clear in what ways ML-based approaches can be utilized to identify meaningful patterns in teachers' constructed responses with respect to noticing and attention to classroom events.

In the present study we therefore evaluate potentials and challenges of using a pretrained language model-based clustering approach to analyze preservice physics teachers' open-ended, constructed responses in the context of describing a standardized teaching situation. We critically examine to what extent the application of ML in our research context can bridge the divide between quantitative and qualitative methods and provide a more integrative approach.

Utilizing NLP and ML to Model Complex Dataset

Applications with ML and natural language processing (NLP) attracted a lot of interest in the field of science education research (Zhai et al., 2020). ML refers to computers' inductive problem solving based on data (Zhai, 2021; Rauf, 2021). Two major types of ML are supervised and unsupervised ML (Jordan & Mitchell, 2015). In supervised ML, human-annotated data are provided for the models to learn a mapping from input to output in order to classify or predict unseen data (Marsland, 2015). Unsupervised ML, on the other hand, encompasses algorithms to reduce complex datasets and extract patterns in them. Both types of ML can be used to analyze natural language. The study of natural language by means of computers is called NLP. NLP refers to the systematic and structured processing of natural language data. Natural language can be contrasted with artificial language such as programming languages or mathematics which are more aligned with formal logic. The attribute "natural" relates to the fact that this form of language can be characterized to be "noisy, ambiguous, and unsegmented" (Jurafsky, 2003). It has been argued that it is not possible to specify clear-cut rules for natural language (i.e., a grammar) that explain phenomena of language comprehension and production: "we can't reduce what we want to say to the free combination of a few abstract primitives" (Halevy et al., 2009, p. 9). Hence, probabilistic approaches such as ML methods are increasingly incorporated into NLP research in addition to rule-based approaches, given the capacity of probabilistic approaches such as ML to systematically process complex language data, extract patterns in it and classify instances of language use (Goldberg, 2017).

ML research experienced a new spring with the successful application of deep neural networks to learn input-output mappings that outperformed more simple (shallow) ML models in most tasks in image and language analysis (Goodfellow et al., 2016). A heuristic in ML research states that problems which are easy for humans are difficult for machines to solve such as character recognition or speech perception (Goodfellow et al., 2016). Simple ML algorithms like logistic regression excelled in problems where the input representation through features is particularly informative, e.g., the age of a student. The selection and engineering of inputs typically requires efforts for the human researcher, because data are not typically represented in this aggregated form in real-world contexts. Simple ML models would lose

performance when more complex data such as images or language form the input (Goodfellow et al., 2016). Deep neural networks have been found to be capable of representing the input as part of the modeling, which allowed ML and NLP researchers to apply these models to problems where complex data has to be represented in the first place. Thus, human feature selection and engineering is partly replaced by automated feature representation in the deep neural network approaches oftentimes with the loss of interpretability of the model decisions.

A major facilitator for the deep learning revolution in the last decades was the availability of annotated data. For once, researchers spend tremendous efforts to annotate data manually in order to train deep neural networks that are capable of language comprehension and production, or image classification. For the now famous ImageNet competition, researchers manually labeled over three million images in two years with the help of crowdsourcing (Mitchell, 2020). Similar efforts have been undertaken in NLP. To advance language translation, ML researchers were fortunate to find annotated datasets from the cold war where translations were important for intelligence or from the European Parliament that consists of many different nations (Mitchell, 2020). However, curating and annotating these datasets captures resources that are not widely available such as money and compute time. Consequently, for most researchers in domains like science education no such well-developed datasets will be available for their specific research questions.

However, the ML paradigm of transfer learning that became important with increasingly complex deep neural networks (Devlin et al., 2018) might solve this problem. Transfer learning enables sharing of previously trained ML models for different tasks (Ruder, 2019). Much as humans learn language from experiences, feedback and reinforcement (Bruner, 1985) and build on learned structures (Rumelhart et al., 1986), the paradigm of transfer learning posits that prior trained weights in a given context can be further used to improve model performance in different contexts/domains and with different tasks (Ruder, 2019). NLP researchers used transfer learning in the context of language modeling. While in image processing models are oftentimes pretrained on the ImageNet dataset to improve downstream performance (Devlin et al., 2018), language models in NLP research can be trained on corpora such as the Internet or Wikipedia (Devlin et al., 2018; Ruder, 2019). NLP researchers then pretrain language models that are capable of representing language in a way that researchers can use in downstream tasks (Mikolov et al., 2013). Typically, these language models are trained with the objective to simply predict context words. The pretrained language models can then be used to generate an informative representation of language to enhance task performance (Mikolov et al., 2013; Mikolov et al., 2013; Devlin et al., 2018).

Modeling Unstructured Data in Science Education Research with NLP and ML

"Perhaps when it comes to natural language processing and related fields [that model human behavior], we're doomed to complex theories that will never have the elegance of physics equations" (Halevy et al., 2009, p.8). The "unreasonable effectiveness" (Wigner, 1960) of mathematics has been recognized for physics; however, educational sciences are far from having theories in this elegant formulation—given the complexity of the involved problems. In this context, NLP and ML have probably much to offer for these fields where complex theories prevail. Yet, especially more sophisticated NLP and ML applications such as deep neural networks might pose unfulfillable requirements on required size of training datasets and model implementation to be useful for science education research. The size of the training dataset should be judged against the complexity of the task and the complexity of the ML model. While the review by Zhai et al. (2020) shows that typical applications of NLP and ML in science education research comprise fewer than 30k training samples, the reviewed studies exclusively focus on simpler ML models such as logistic regression, support-vector machines, or naive Bayes. More generally, data collection in domains such as science education is costly and time-consuming, because large coordination efforts are necessary to recruit enough subjects. There are literally no studies in science education where millions of subjects have been collected that comprise datasets that seem to be required to train more general-purpose deep learning models. Does this imply that particularly the more complex ML methods are not applicable for science education researchers?

For supervised ML this hypothesis has been refuted in some science education research contexts. Wulff et al. (2022) could show that pretrained language models improve classification performance for discourse elements with preservice physics teachers' written reflections. The findings in this study suggest that complex ML models that are trained from scratch can reach classification accuracy of simpler ML models. Furthermore, the authors show that utilizing pretrained weights for the complex models enhances classification accuracy and generalizability further. Carpenter et al. (2020) showed that deep contextualized embeddings from pretrained language models could improve prediction of students' reflective depth in a biology learning context. These findings buttress the applicability of complex ML models such as deep neural networks as facilitators for supervised ML. These studies, however, do not suggest that training the more performant deep learning models from scratch is possible with the available science education datasets. Furthermore, it is not clear from these studies to what extent pretrained language models could be used to extract patterns in the datasets.

Prior research on pattern extraction from unstructured data with simpler unsupervised ML models and larger datasets in education and science education contexts focused on standardized documents such as dissertation or conference abstracts. Munoz-Najar Galvez et al. (2020) established a data-driven way to systematically analyze the field of education research. They identified paradigm shifts in education research on the basis of 137,024 dissertation abstracts, reconstructing a shift from an outcome-oriented paradigm to an interpretative paradigm. In science education research, Odden et al. (2020) used latent Dirichlet allocation (LDA), a generative probabilistic topic model, to analyze all papers that were extracted from the Physics Education Research Conference Proceedings from 2001 to 2018 (overall 1,302 papers). They outline shifts in the paper's topics in the conference over time. Despite the potentials of LDA to summarize occurring research topics and trends over time, the authors recognize some shortcomings with this algorithm. For example, the LDA model groups together segments that use similar vocabulary. However, the segments might differ in meaning anyways (see also: Odden et al., 2021). Other researchers could show that simpler unsupervised ML methods could also be used to explore patterns in comparably smaller datasets in science education. Sherin (2013) used a vector space model and a hierarchical agglomerative clustering algorithm to identify students' science explanations in interview transcripts. He showed the general applicability of these NLP-based methods in this context, but contends that the algorithms could not account for word ordering effects. He also mentions the desire to more systematically extract the number of topics that are likely present in the data (see also: Xing et al., 2020). Also Rosenberg and Krist (2020) successfully applied an unsupervised clustering algorithm to assess students' considerations of generality in science (see also: Xing et al., 2020; Zehner et al., 2016).

A domain of research in science education where NLP and ML in unsupervised contexts has not yet been applied widely is university-based science teacher education. In fact, no reviewed study in Zhai et al. (2020) engaged in university-based educational research. Besides supervised ML approaches in university-based science teacher education that have been occasionally applied (Wulff et al., 2020), unsupervised approaches could facilitate researchers and instructors novel insights into relevant constructs because they can explore patterns in unstructured data (Halevy et al., 2009; Hao, 2019).

Science Teachers' Noticing of Classroom Events

Teachers face the challenge to professionally act in uncertain situations (Clifton & Roberts, 1993; von Aufschnaiter et al., 2019; Chan et al., 2021). Learning to professionally act in uncertain situations requires teachers to develop the capacity to reflect on their teaching experiences (Korthagen, 1999). An important part of reflective competencies are noticing skills that relate to perceptual and cognitive thinking processes (Chan et al., 2021). In particular, noticing comprises observation, interpretation, and reasoning about learning-relevant events in classrooms (Sherin & van Es, 2009; van Es & Sherin, 2002a; Chan et al., 2021; Furtak, 2012). Van Es & Sherin (2002) define noticing with regard to three key aspects: "(a) identifying what is important or noteworthy about a classroom situation; (b) making connections between the specifics of classroom interactions and the broader principles of teaching and learning they represent; and (c) using what one knows about the context to reason about classroom interactions." (p. 573) Noticing research has documented the difficulties that novice and even expert teachers have to direct their attention and notice relevant classroom events (Sherin & Han, 2004; Chan et al., 2021; Talanquer et al., 2015; Levin et al., 2009; Roth et al., 2011). For example, novice science and mathematics teachers struggle to attend to student thinking and the substance of what they are saying (Sherin & Han, 2004; Hammer & van Zee, 2006), and tend to strive for quick and conclusive inferences that are right or wrong, rather than tentative interpretations (Crespo, 2000). This strand of research also showed that science teachers provide more general evaluations as compared to more specific accounts of student understanding (Hammer & van Zee, 2006). Mathematics and science education scholars generally highlighted the complexity of the noticing construct (Chan et al., 2021; Talanquer et al., 2015). Talanquer et al. (2015) summarize the noticing foci of teachers as: "the object of noticing (e.g., student actions, student thinking), the noticing stance (e.g., evaluative, interpretive), the specificity of noticing (e.g., specific student, whole class), and the noticing focus (e.g., specific concept, general topic)" (p. 587). To design authentic learning opportunities for mathematics and science teachers to enhance noticing skills, valid, reliable, and scalable assessment of attention to classroom events and noticing is necessary.

To assess noticing and attention to classroom events, science education researchers increasingly embraced constructed response items, e.g., open-ended, free-recall written responses (Barth-Cohen et al., 2018; Luna et al., 2018; Talanquer et al., 2015; Chan et al., 2021). Open response items have been argued to allow a more authentic examination of teachers' professional competencies as compared to more closed-form questions (Nehm et al., 2012; Zhai, 2021). Many of the noticing research then seeks to analyze inductively what teachers are noticing (Chan et al., 2021). However, the mere linguistic complexity of the constructed responses (noisy, ambiguous, and unsegmented) and the

complexity of the noticing construct make it challenging to integrate all information in the responses and infer the noticing skills. From their review on teacher noticing research in science education, Chan et al. (2021) conclude that "methodological trade-offs between different ways of investigating teacher noticing need to be better explored" (p. 37). We suggest that ML-based methods can provide novel means to analyze teachers' responses inductively "to understand what teachers notice" (Chan et al., 2021, p.34). Thus, ML methods potentially help researchers to gather 'knowledge of teachers' (Fenstermacher, 1994). We also concur with Lamb et al. (2021) that ML models are powerful tools to advance algorithmic understanding of relevant underlying cognitive processes that can explain the process and products of writing. Zhai et al. (2020) argued that ML models can particularly advance understanding and assessment of complex constructs such as noticing and provide means to automate assessment and feedback. Consequently, this study examines potentials and challenges of an ML-based clustering approach when applied in the context of assessing noticing of classroom events for preservice science teachers.

Research Questions

Noticing or directing attention to relevant classroom events is highly relevant for mathematics and science teachers and, thus, plays an important role in science education research. Attention to classroom events and contents played a particularly important role in mathematics and science education research. Star and Strickland (2008) suggested that noticing research should focus particularly on what catches teachers' attention and what is missed. 25 of the 26 science education studies reviewed by Chan et al. (2021) considered attention to classroom events as an essential aspect of noticing; 11 studies even restricted noticing to attention. Attention to classroom events has often been studied through video clips that present teachers with a standardized teaching situation and are typically followed by some form of eliciting teachers' observations (Zhai, 2021; van Es & Sherin, 2002a; Seidel & Stürmer, 2014; Putnam & Borko, 2000; Darling-Hammond, 2000; Kleinknecht & Gröschner, 2016; Sherin & van Es, 2009).

Noticing research can be characterized as a context where it seems to be notoriously difficult to recruit large sample sizes, rendering quantitative research methods difficult to apply. Reviews suggest that studies typically comprise small samples of up to 241 teachers (Wilson et al., 2019; Chan et al., 2021). This restricts researchers to using mostly qualitative methods with some form of content analysis (Wilson et al., 2019; Chan et al., 2021; Talanquer et al., 2015). As such, it is important to examine to what extent ML-based approaches can be utilized in this context as a means to advance quantifiable hypotheses. Particularly, pretrained language models can improve the ML methods to be more robust with small samples. Hence, we ask the following overarching research question: To what extent and in what ways can a pretrained language model-based clustering approach extract meaningful patterns in preservice physics teachers' written descriptions of a teaching situation?

In the context of RQ1, we analyze the validity of the extracted clusters:

- RQ1: To what extend can a pretrained language model-based clustering approach extract interpretable (RQ1a), specific (RQ1b), and robust (RQ1c) clusters in the preservice physics teachers' written descriptions of a teaching situation?

We then examined ways in which these clusters provide insights into the composition of the written descriptions. van Es and Sherin (2002) used the concept of analytical chunks in their noticing research, referring to experts' tendency to organize their essays more coherently in reference to teaching and learning principles. Based on this concept of analytical chunks, we hypothesize that the analysis of interconnections between the clusters in the teachers' written descriptions provides tools to develop a more quantitative understanding of chunks in the writing. To analyze the organization of the teachers' written descriptions based on the extracted clusters, we explored dependencies among clusters:

- RQ2: What kinds of dependencies with respect to textual organization can be analyzed based on the extracted clusters?

Method

Written Descriptions of a Video-Recorded Teaching Situation in Physics

In the present study preservice physics teachers' were given the instruction to describe, evaluate and reason about a video-recorded lesson which presented the teachers an authentic teaching situation in a 9th grade physics classroom held by an in-service physics teacher. Overall, the teaching goal of the observed lesson was to introduce influencing factors on the movement of falling objects and the definition of free fall. Table 2 outlines the chronological order of events in the teaching situation. The teaching situation can be broadly divided into two phases. In the first phase, the teacher performed several experiments with falling objects (two masses, and a vacuum tube with screw

Table 1 Sample description

Experience	Term	Place	Seminar type	*N*	*M* (age)	*Md*	*SD*	prop female
Bachelor	Winter 2020/21	University C	Physics edu. seminar	5	22.6	23.0	2.5	0.40
Bachelor/Master	Summer 2020	University A	Physics edu. seminar	31	24.7	23.0	3.7	0.13
Master	Summer 2020	University B	Teaching internsh. (Master)	7	24.0	24.0	2.4	0.14
Master	Winter 2019/20	University B	Teaching internsh. (Master)	13	25.1	23.0	3.9	0.23
Master	Summer 2021	University B	Teaching internsh. (Master)	7	25.3	23.0	6.4	0.29
Master	Winter 2020/21	University B	Teaching internsh. (Master)	12	25.8	25.0	6.6	0.50
Master	Winter 2020/21	University B	Teaching internsh. (Master)	8	26.4	25.0	7.5	0.62
Bachelor	Winter 2020/21	University B	Teaching internsh. (Bachelor)	3	26.3	27.0	3.1	0.33

and feather). The students posed hypotheses on the outcome of the experiments (e.g., which of the two masses of different weight will hit the floor first. In the second phase, the teacher provided the definition of free fall and students devised experiments to investigate what type of movement free fall is. This video-recorded teaching situation was chosen because it presents preservice physics teachers a complex and authentic teaching situation where many different noticing-relevant general and subject-specific issues could be identified. Teachers could describe mere surface-level, general issues such as that the students were noisy at several occasions, or more deep-level, subject-specific issues such as that several students raised concerns with the experimental setup (e.g., missing control of variables) or conceptual difficulties (e.g., whether an ever-accelerating object reaches the speed of light). Following the classification rubric for noticing research in science education by Chan et al. (2021), our approach was meant to characterize teacher noticing (purpose) as assessed through observation of other teachers' teaching (teaching context), where the observing teachers could not control what happened (role of teacher) and the noticing-relevant events were predetermined (what to notice) and selected by the researchers (selection of probes) with open-ended prompts (nature of prompt) and divergent answers without correct answer (type of teacher responses).

The video is about 17 minutes long. The preservice physics teachers were allowed to watch the video only once, without rewinding the recording, in order to simulate in-the-moment pressures of decision-making (Chan et al., 2021). It was an authentic lesson that was recorded in a German grade 9 high school physics classroom as part of a post-university physics teacher preparation program. In Germany, after the university-based teacher training teacher trainees are required to pass a one- to two-year program, run by federal states, that will approve if they are finally allowed to teach in public schools. Using a recorded lesson from this post-university teacher preparation program presents a lesson that is proximal to what the preservice teachers will do in their future careers. Overall, N=75 preservice physics teachers participated in the study who produced 86 written descriptions (sometimes preservice teachers produced two texts, pre and post to a seminar). The teachers varied in their teaching experience and came from three different universities throughout Germany (see Table 1). Preservice teachers spent approximately one hour on the entire questionnaire of the online video-vignette. The text production took approximately 20 minutes (independently of another 17 minutes video observation and another 20 minutes answering further questions). Preservice physics teachers were instructed to first describe what happened in the teaching situation. Afterwards, they should evaluate the situation, devise alternative modes of action, and formulate consequences for their own teaching.

Given that preservice physics teachers either described, evaluated, and reasoned about the observed teaching situation, the sentences that count as descriptions were extracted through an ML-based classifier. The ML-based classifier automatically retrieved descriptive sentences based on a classification algorithm that was described elsewhere (Wulff et al., 2022). This classifier annotated each sentence with one of the following labels: "circumstances", "description", "evaluation", "alternatives", and "consequences." Using sentences as the segmentation units was found to be a reasonable strategy in similar contexts of writing analytics (Ullmann, 2019). The descriptive sentences were further filtered to a length greater than four words to remove headlines and similar non-informative sentences. 98% of sentences of the original descriptive sentences remained (1537 sentences in total). The preservice teachers wrote on average 16.0 (SD = 7.9, min: 4, max: 59) words in a descriptive sentence. In descriptive sentences, the preservice teachers wrote in various ways about the events in the lesson as outlined in Table 2. A randomly drawn sentence from a preservice physics teacher reads as follows: "The observations [from the students] and differences [to the hypotheses] were collected and summarized by the teacher as free falling movement is independent of the mass." This sentence and all words and sentences in the following were translated from German to English by the authors who are familiar

Table 2 Sequencing of the lesson

Sequence	Description	Teaching Goal
S1	Introduction: Teacher reminds students of a problem from the last lesson where they compared a race between hare and hedgehog - Teacher: "Today we are going to observe a different race, between a feather and a screw."	Motivation for the topic, focus on the movement of free falling objects
S2	Teacher shows both objects and asks which lands on the floor first if the teacher drops them; students hypothesize: 1. screw because of gravitational pull, 2. feather has higher air resistance; Teacher conducts the experiment (feather sticks to his hand at first)	Demonstration that feather falls slower compared to screw
S3	Teacher: How can we change the experiment to determine influencing factors on free falling movement?; Students offer: wider screw (in order to raise air resistance), thicker feather (to make it comparable in mass to screw)	Transition to experiment with two equally shaped objects that differed in their masses
S4	Teacher drops a 100g and 50g mass object; given similar movements, teacher concludes that the mass has no influence; Student suggests to probe the experiment in vacuum, which the teachers takes as a "perfect" transition to the next step	Demonstrate that movement was independent of mass of objects; collection of students' hypotheses
S5	Teacher shows vacuum tube with feather and mass inside; Teacher asks for hypotheses about result; One student first suggests that both arrive at the same time, and then refines his answer to hypothesize that the heavier object arrives first, another student suggests that both arrive at the same time because the air resistance is removed	Demonstrate dependence of free falling movement on air resistance
S6	Teacher opens valve in vacuum tube to let air flow into it and repeats the experiment. Initially the feather sticks to the glass. Afterwards, the objects move at different rates.	Demonstrate that initial results (dependence of movement on air resistance) can be replicated with the vacuum tube
S7	Teacher summarizes the experimental findings; he asks what influences the falling movement of the objects. The students reply air resistance; teacher asks what does not influence the movement and the students reply the weight.	Conclusions, summary
S8	Teacher introduces movement without air resistance and downward as free fall; Student asks why a jump by a parachute jumper is called a free fall: because it is without a parachute in the initial phase or because there is no air resistance? Teacher postpones the question until later	Defines free fall as movement without air resistance
S9	Teacher asks students to devise a new experiment that can test the type of this movement; Student suggests to measure time at certain waymakers in space; Another student suggests to measure distance at certain time measurement points; Further student asserts objects would reach speed of light without bottom and without air resistance, in accelerated movement; Discussion which experimental setup is more practical	Devise new experiment to check the type of free falling movement (i.e., constantly accelerated movement)

with English language, in particular specialized vocabulary in physics. Some intricacies emerged with the translations. For example, German language has many specific abbreviations in educational contexts, e.g., "SuS" ("Schülerinnen und Schüler") for female and male students or "LK" ("Lehrkraft") as an inclusive word for teacher that have no equivalent in English. We tried to highlight those issues when they occur. Furthermore, German language is well known for its compound nouns that can become very long (e.g., "Fallröhrendemonstrationsexperiment", which can be translated to "demonstration experiment with drop tube"). In German, compound nouns may count as one word in the vocabulary, whereas in English many different words would be added. Consequently, the German vocabulary in terms of distinct words is larger compared to the English vocabulary.[1]

Clustering Sentences of the Written Descriptions

ML methods that extract patterns in unstructured data such as the constructed responses are categorized as unsupervised

[1] Researchers who would like to adopt the presented clustering approach with English language data would have to implement the English language model which is readily available (see description of technical implementation in the supplement).

ML. Unsupervised ML typically include some form of dimensionality reduction and clustering oftentimes with the purpose to make high-dimensional data human-interpretable. Clustering approaches that were not based on pretrained language models enabled science education scholars to identify emergent topics in conferences or students' writing (Odden et al., 2020; Sherin, 2013), however, they also oftentimes require involved preprocessing of the data (Angelov, 2020; Odden et al., 2020; Zehner et al., 2016). Most often, researchers needed to remove frequent words (stopwords), lower-case all words (which might be disadvantageous in German where upper-case letters can differentiate word senses), or transform words into their base form to reduce vocabulary size (Odden et al., 2020; Rosenberg & Krist, 2020). Furthermore, researchers noted the difficulty in determining the number of clusters that should be extracted in these approaches (Sherin, 2013) and these approaches oftentimes assume that word order in the sentences is irrelevant (bag-of-words assumption). Finally, these approaches are ignorant of ambiguous word senses. No prior information on the words is incorporated in these approaches such that the word "bank" in the phrases "river bank" and "bank robbery" might be treated as the same word even though the meaning differs substantially. Recently, however, advances in NLP and ML research provided pretrained language models that provide contextualized embeddings for language data that help to cope with some of the aforementioned challenges. These contextualized embeddings potentially enable researchers to model constructed responses in a more language-sensitive way that is able to preserve word ordering and word sense disambiguation as features.

Pretrained language models can generate contextualized embeddings for language input that enhances modeling of the language data (Mikolov et al., 2013; Sherin, 2013; Taher Pilehvar & Camacho-Collados, 2020). Essentially, words are mapped to a position in high-dimensional vector space, called a distributed representation in the form of embeddings (Taher Pilehvar & Camacho-Collados, 2020). Vector space models thus encode word similarity and efficiently represent words. Given the claim that one understands a word by the company it keeps (Jurafsky & Martin, 2014), word embeddings can be learned through ML approaches, where model weights are optimized with the goal that a word embedding for a given word predicts the context words (Mikolov et al., 2013). More advanced approaches utilize pretrained language models that result in embeddings that also account for the context (contextualized embeddings) and the position in a segment that a word occurs in (Taher Pilehvar & Camacho-Collados, 2020). Pretrained language models are typically trained on large unstructured datasets (e.g., the Internet, Wikipedia). Training tasks involve prediction of context words (Devlin et al., 2018). For practical purposes the vocabulary is often restricted to some 30,000 tokens, where unknown words can be built from the 30,000 tokens. Linguists have estimated that 30,000 words are sufficient to understand many general English texts well (Mitchell, 2020). If a sentence is input into a pretrained language model, typically embeddings for each word in the sentence (given the position and context words) is the output. To generate a contextualized embedding for a sentence, the word embeddings can be pooled.

As an illustrative example for sentence embeddings based on pretrained language models, the following physics-related and general sentences should be considered (some noise data points were added which will be motivated later on): 'Earth exerts a force', 'The force acts on', 'The force on earth', 'We force her', 'They force him', 'How to force him', 'Grass is green', 'The sunset can be red', 'Green is grass' (called Segment 1 to 9 respectively). Force in the first three sentences relates to the physics meaning (given as a noun). In the following three "force" is included as a verb that encapsulates a certain kind of rather aggressive behavior. The final three sentences are included as sentences that are entirely different in meaning. "Force" in the former sentences has a different word sense compared to the sentences 4 to 6 and should be distinguished in a clustering approach. In Fig. 1(a) a two-dimensional representation of the sentence embeddings gleaned from a pretrained language model is depicted. As can be seen from the separation of datapoints in space, pretrained language model's word embeddings can in fact disentangle the senses to a certain degree. To further inspect the embedding space, a clustering approach can now determine which sentences are likely related to each other (Angelov, 2020).

Extracting clusters from contextualized embeddings can be done with Hierarchical density-based spatial clustering of applications with noise (HDBSCAN) (Campello et al., 2013). HDBSCAN is a way to calculate the number of dense volumes (i.e., clusters) in the embedding space. Density-based clustering methods consider the probability density of a collection of data points (Kriegel et al., 2011). In Fig. 1(b) the probability density distribution for the data points in Fig. 1(a) is depicted. To extract clusters, an imaginary water level can be introduced into the probability space. The water level represents a threshold for cluster extraction. Emerging islands, i.e., regions above the water level, represent clusters. If water level rises, less probability mass lies above the water level, and thus fewer clusters are extracted. A suitable water level has to be chosen in order to extract an appropriate amount of clusters.

To perform the actual clustering the nearest neighbors for each data point will be determined and the closest distance between nearest neighbors will be highlighted as edges in a graph, i.e., the minimal spanning tree (see Fig. 1(c)). A threshold parameter (i.e., the minimal distance) is then

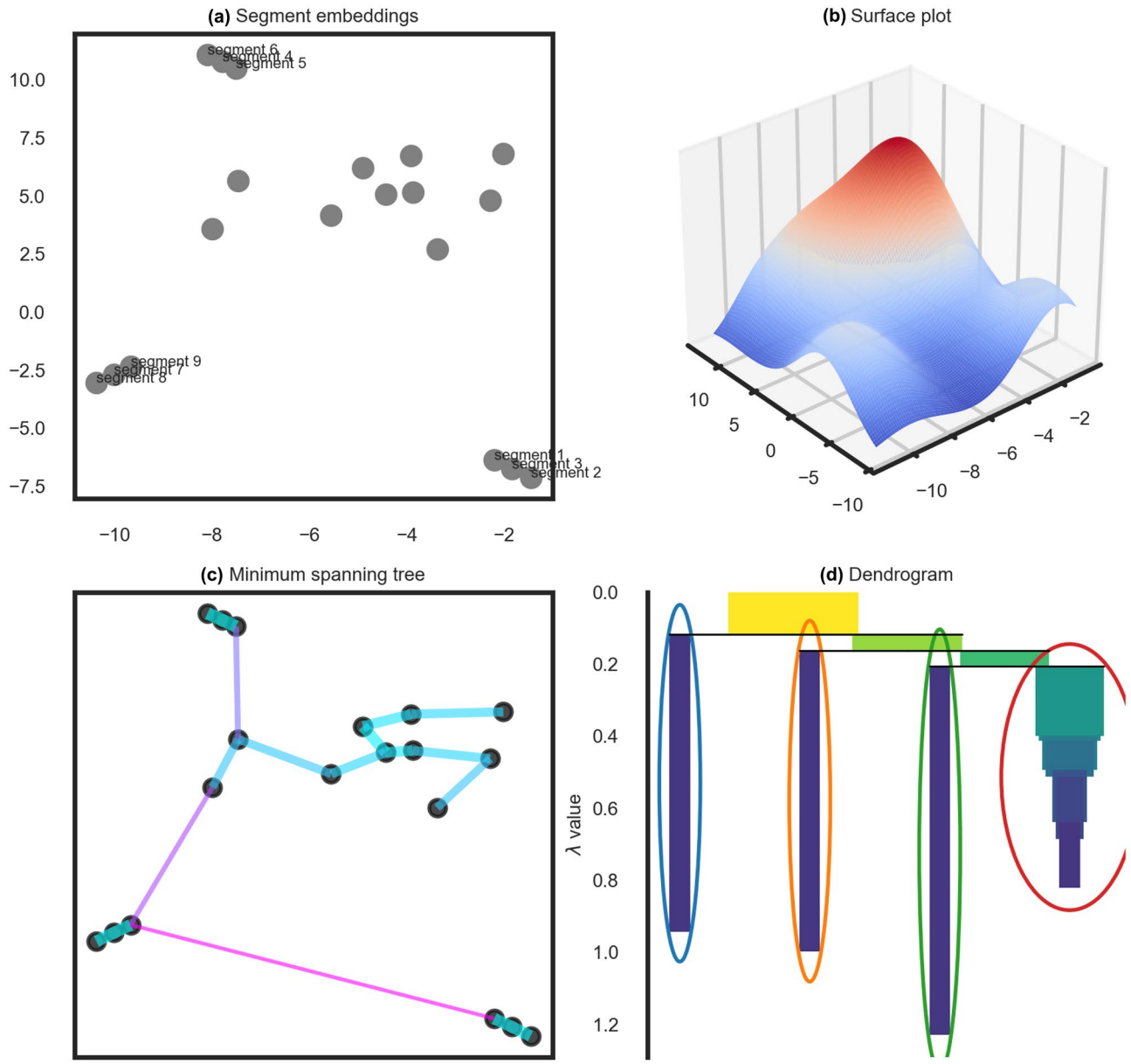

Fig. 1 **a** Two-dimensional representation of the example segments and noise. **b** Surface plot of probability density of the data points. **c** Minimal spanning tree with data points as nodes (colors indicate the mutual reachability distance). **d** Dendrogram of clusters for varying density values (colored circles indicate clusters)

varied where edges that surpass the threshold are removed from the graph. Finally, the minimal spanning tree is mapped into a condensed tree representation (see Fig. 1(d)). The condensed tree depicts the number of data points in a cluster (width of the branches) with varying densities (λ). A way to extract clusters from the condensed tree is by defining a minimal cluster size and examining the stability of the branches over different density values (moving up and down in Fig. 1(d)). It is desirable to have clusters that persist over varying density-levels. The stability of a cluster basically relates to the regions of maximum area in the condensed tree Kriegel et al. (2011), Campello et al. (2013). The algorithm thus determines a number of clusters by examining properties of the clusters. From the illustrative example, the resulting clusters based on this clustering approach (HDBSCAN combined with pretrained language models) are depicted as blue, orange, and green ovoids in Fig. 1(d). The red-shaped ovoid cluster could be considered as noise, given the instability over density values in Fig. 1(d). If the sentence embedding points in Fig. 1(a) were to be colored, the closely aligned sentences would in fact be colored with the same colors, respectively.

Analysis Procedures

Interpretability of Clusters (RQ1a) In order to evaluate if the pretrained language model-based clustering approach[2] outputs represent interpretable clusters, the most representative words for each cluster were considered, and a definition was derived. Visual inspection of the two-dimensional embedding space and the condensed tree representation helped to determine similarities and differences of the clusters. If the five most representative words could be mapped to distinct sections in the observed teaching situation (see Table 2) and were coherent, then we considered this as evidence of a meaningful cluster, because clusters were anticipated to attend to localizable events (e.g., experiments) or actions (e.g., devising hypotheses). We also assessed to what extent the clusters related to physics ideas that were implicitly or explicitly relevant in the observed teaching situation, and what ideas or events were not clustered.

Specificity of Clusters (RQ1b) Then it was evaluated to what extent physics-savvy human raters could use the extracted clusters to manually annotate the video-recorded teaching situation. If human raters struggled to annotate a certain cluster in the video recording, this would provide evidence of unspecific focus of a cluster. To annotate the teaching situation on the basis of the extracted clusters, three independent raters with physics background (one postdoc, two PhD students) who were familiar with the observed teaching situation annotated the entire video sequence based on 10 second intervals. All the information they received were the five most representative words for the respective clusters (coding 1) with no further instruction. In a second iteration (coding 2), the human raters discussed and agreed on some coding rules, e.g., that the entire process of an experiment should be annotated if relevant words of a cluster occurred only at the beginning. To evaluate the reliability of this annotation, we first examined a graphical representation of the annotations over time to evaluate interrater agreement. We considered each cluster separately. To evaluate interrater agreement, Krippendorff's α for each cluster was calculated because Krippendorff's α is more appropriate than Cohen's κ for three raters. A Krippendorff's α value of 1 refers to perfect reliability and a value of 0 to absence of reliability. Values between .667 and .800 are usually considered to allow researchers to draw tentative conclusions, i.e., consider the agreements as non-random (Krippendorff, 2004).

Robustness of Clusters (RQ1c) To analyze robustness of clusters, the clustering approach was applied to smaller subsets of the dataset. To test if small sample sizes are enough, subsets of N=43 randomly chosen pre-service teachers and $N = 8$ randomly chosen pre-service teachers were considered. The extent to which similar clusters emerge was examined. If meaningful clusters could be identified in these subsets, then we considered the algorithm robust with sample size variations which could be beneficial for science education researchers who oftentimes only have small samples at their disposal. Furthermore, we compared the outputs of the pretrained language model-based clustering algorithms with a clustering approach that was not based on pretrained language models, but was successfully applied in a science education research context before. We therefore adopted the topic modeling approach outlined by Sherin (2013). He devised an accessible approach for extracting clusters in interview transcripts. He started by segmenting texts into chunks of 100 words (with overlap). Afterwards, a normalized term-document matrix was formed. To circumvent the problem of similar topics (low levels of variability in the data), deviation vectors were calculated. Based on the deviation vectors, hierarchical agglomerative clustering yielded a distribution of topics, depending on the number of topics. Finally, the ten most representative words were found as the highest ranking words in the centroid vectors for the respective topics. With parameter values adapted to our research context, we extracted clusters from our descriptions based on this approach. Based on the comparison from the ten most representative words for each topic, we evaluate to what extent both clustering approaches yield similar topics. This would yield evidence that the pretrained language model-based approach could also be successfully employed in science education research contexts.

Advanced Textual Analytics Based on the Clusters (RQ2) The applicability of the pretrained language model-based clustering for analytics of the constructed responses was evaluated through exploratory analysis of the textual organization of the constructed responses. Based on episodic memory theory it can be expected that the preservice teachers provide a chronologically ordered text organization. Hence, the temporal progression of the clusters within the teachers' written descriptions was analyzed. To depict the temporal progression of the clusters within the written descriptions, the sentences were mapped to their relative position in reference to the other descriptive sentences for each teacher (see similar analysis in: Sherin, 2013). Mapping the sentences to their relative position was supposed to produce certain peaks where clusters are most prevalent in the descriptions. For example, it could be expected that mentioning the introduction with hedgehog and hare or the teacher experiments precedes other clusters such as the discussion of the type of movement, because these descriptions appeared first in the observed teaching situation and teachers are expected to describe the teaching situation chronologically. Distinctiveness in temporal progression would indicate that the extracted clusters in fact captured different aspects of the teaching situation. To further analyze

[2] Please find details on the technical implementation in the supplement.

Table 3 Number, share (i.e., number of segments) and top five words of the extracted topics. *M*, *Md* are mean and median number of sentences for a cluster in a written description, respectively. *SD* is the standard deviation; range is minimum and maximum number of sentences

Topic	Share	Top five words	Definition	*M*	*Md*	*SD*	range
-1	760	students, student, experiment, teacher (formal), teacher (informal)	-	8.7	7.0	6.5	1 - 34
0	38	floor, piece of mass, simultaneous, pieces of mass, students	Pieces of mass dropped on floor and touch floor simultaneously.	1.5	1.0	0.9	1 - 5
1	56	pieces of mass, masses, same, shape, fall	Pieces of mass have the same shape.	1.1	1.0	0.4	1 - 2
2	246	feather, screw, tube, vacuum, air	Experiment in vacuum tube with feather and screw, without air.	3.4	3.0	2.4	1 - 13
3	10	respond, feedback, again, summarize, guides	Teacher summarizes answers, responds and gives feedback.	1.7	1.5	0.8	1 - 3
4	42	students, raise arms, raised arms, respond, same	Students raise their arms and respond to the teachers' questions	1.8	1.0	1.9	1 - 8
5	11	summarize, main hypotheses, main theses, claims, hypotheses	Teacher summarizes the main hypotheses/theses/claims.	1.2	1.0	0.7	1 - 3
6	56	claims, hypotheses, validate, students, asks	Students validate their claims/hypotheses through experiments.	1.4	1.0	1.0	1 - 6
7	25	summarize, claims, guides, teacher (formal), teacher (informal)	Teacher summarizes the claims of the students.	1.4	1.0	0.6	1 - 3
8	49	type of movement, what kind, teacher (informal), asks, type	Teacher asks what type of movement it is.	1.6	1.0	0.9	1 - 4
9	54	measure, movement, distance, steady motion, certain	Students propose to measure distance at certain times and respond steady motion.	1.8	1.5	1.0	1 - 5
10	19	screw, weight force, air resistance, higher, different	Students say screw has a higher weight force, and feather higher air resistance.	1.3	1.0	0.8	1 - 4
11	64	air resistance, mass, influence, speed of fall, dependent	Air resistance has influence on speed of fall, also dependent on mass.	1.6	1.0	0.7	1 - 3
12	79	free, fall, folder, definition, topic	Teacher writes the definition of free fall, students copy it in their folders.	1.6	1.0	1.2	1 - 8
13	28	hedgehog, lesson, lesson, hare, last	Teacher reminds students of last lesson with race between hedgehog and hare.	1.5	1.0	0.9	1 - 4

textual organization, we employed a network-analytical approach to calculate the centrality of different clusters and a vector-field approach where the movements through cluster space can be characterized. In both approaches we will evaluate to what extent the respective empirical distributions, i.e., the directed network of clusters and the vector-field representation, are better captured by random processes or more deterministic processes. If teacher's written descriptions can be characterized by more deterministic processes, we can conclude that the presented clustering approach can yield insights into textual organization.

Findings

Validity of the Clustering Approach (RQ1)

Interpretability of the Extracted Clusters (RQ1a)

To evaluate the interpretability of the extracted clusters, contextualized embeddings of the preservice physics teachers' descriptive sentences were generated with the pretrained language models and clusters were extracted with the HDBSCAN algorithm. This approach yielded a number of 14 clusters and a noise cluster (cluster -1). The absolute sizes (# of sentences in a cluster) are depicted in Table 3. We also provided a definition of the clusters based on the most representative words for each cluster, and we determined how many sentences per written description on average were categorized into each cluster (see Table 3). The largest share of sentences was coded as -1.[3] The graphical representation of the embedding space with clusters highlighted in colors can be seen in Fig. 2. The embedding space can be fundamentally separated into two overarching groups (indicated

[3] This means that these sentences were not close enough to any of the cluster centroids. Column three of Table 3 indicates that very general words fall into this cluster, such as "SuS" (which is the unisex abbreviation for "students" in German) or "lehrer" (engl.: "teacher"). Seemingly, this cluster encapsulated descriptive sentences that were too general or that might belong to multiple clusters such that they average out.

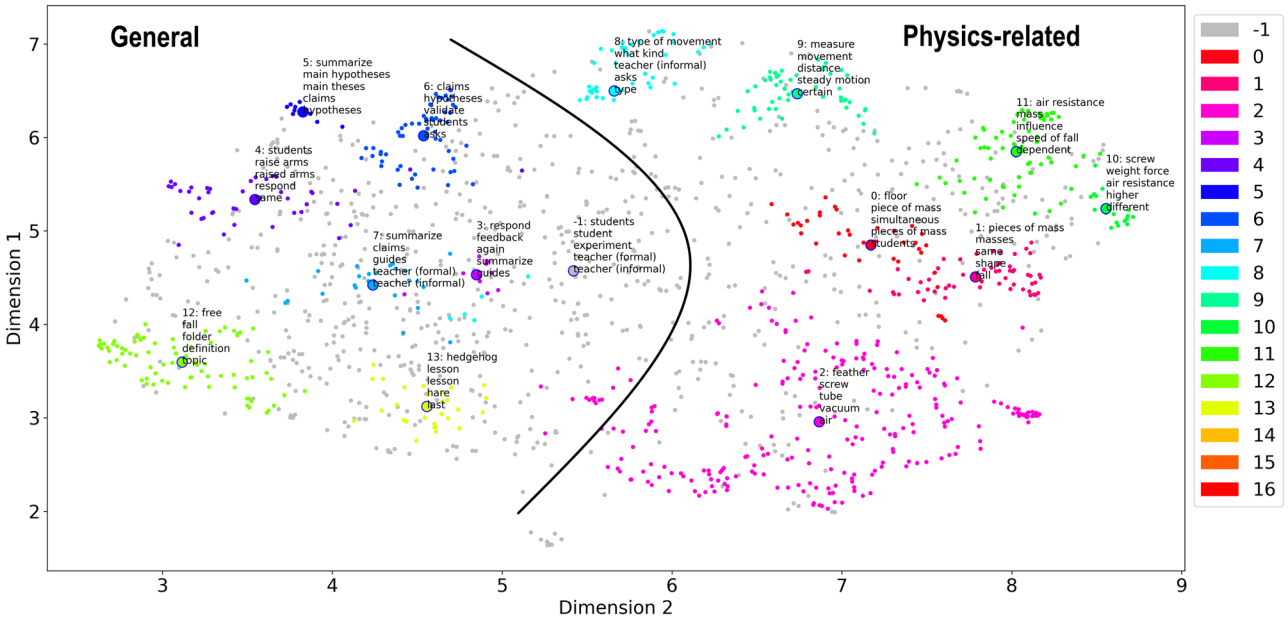

Fig. 2 Two-dimensional representation of clusters. A point represents the projection of a sentence embedding into the two dimensions. Colors represent belonging to a cluster. Gray points represent "noise", i.e., not belonging to any cluster. Larger points indicate cluster centroids

by the black line): (1) clusters that relate to physics-related events or topics that occurred during the teaching situation and (2) clusters that encapsulate general actions, and specific, non-subject-related events. In group 1, cluster 2 thematizes the central experiment of the lesson where a feather and screw are observed falling in a vacuum tube. Cluster 2 had the second largest share of sentences in the descriptions (see Table 1). Relatedly, cluster 10 likely represents the students' hypotheses that the screw has a higher weight, whereas the feather has a high air resistance. Clusters 0 and 1 represent the other experiment, in which two mass pieces (equal shape, different mass) are dropped simultaneously to deduce that free fall is independent of mass. Clusters 8 and 9 refer to the teacher's question about which type of movement a free fall is and how this type of movement can be experimentally determined.

On the other hand, in group 2, clusters 6 and 7 represent teachers' and students' actions of summarizing and posing hypotheses/claims respectively. Given the similarity of clusters 6 and 7, they were also close in embedding space. Cluster 6 was related to posing hypotheses by the students, whereas cluster 5 was related to the process of summing up the hypotheses by the teacher. In fact, this was a recurrent thread in the lesson: the teacher asked the students to hypothesize about the results in advance of an experiment which is why the cluster was coded at several points. Cluster 3 also refers to the teachers' responding to students' answers. Cluster 4 represents the students' action of raising arms and responding to the teachers' questions. Cluster 13 captured the beginning of the lesson where the teacher reminds the students of the former lesson regarding the race between hedgehog and hare. Finally, cluster 12 referred to the instruction by the teacher that the students may copy the definition of free fall into their folders.

In sum, the clusters encapsulate both short and rather specific events in the teaching situation (e.g., writing the definition of free fall in the folder) and more abstract ideas such as summarizing hypotheses which occurred more than once in the scene. They also include more general clusters (summarizing students' hypotheses, e.g., cluster 5) and more physics-related contents (characterization of the type of movement, e.g., clusters 8 and 9). Preservice physics teachers wrote on average 3.4 sentences on cluster 2, which comprised the largest share (after the noise cluster), followed by cluster 1 and 11 with 1.8 sentences on average. Thus, physics-specific clusters were more extensively included in the written descriptions. However, the overall low average counts of one sentence for a cluster could indicate that oftentimes the preservice teachers only briefly elaborated on an event. It is also noteworthy that some important events in the teaching situations are not captured in a cluster. During the lesson the students asked for example: "Why is it called free fall for a parachute jumper?", "Would an infinitely accelerating mass surpass the speed of light?", or "Would two plates, one made of cardboard the other made of metal, actually arrive on the floor at the same time?"

Specificity of the Extracted Cluster (RQ1b)

To examine to what extent the extracted clusters map to discernible events and topics in the teaching situation, human raters used the clusters as represented through the most informative words to annotate the video recording of the teaching situation (RQ3). Figure 5 depicts all codings from three independent annotators separated by cluster over time. To estimate human interrater agreement, we calculated the Krippendorff α values for the clusters. After the first round of rating the video-recorded teaching situation (coding 1), the Krippendorff α's indicate that some clusters (e.g., 0, 1, 2, 8, 12, and 13) could be identified with good reliability given only the five most representative words and no annotator training. Cluster 12 related to the introduction of the definition of free fall by the teacher. This, apparently, was a localizable event in the teaching situation. Cluster 0 related to the experiment with two masses (similar reasoning for cluster 1). The teacher used two masses only once as an experiment, hence, this formed a recognizable event for the human raters. Cluster 13 related to the very beginning of the lesson. The words "hedgehog" and "hare" are unique for this event. The human annotators reached poor reliability on clusters with more general words (e.g., 3, 7, and 11). The words "respond", "feedback", "summarize", and "teacher" could be applied to many different events in the teaching situation. They represent high-inferential categories, because the teacher and students did not specifically say that they "responded" or "summarized" ideas.

After coding 1, the three annotators made their coding rules more explicit and discussed them. On this basis, the video-recording of the teaching situation was annotated again by all three annotators (coding 2). Some improvements could be seen after the discussion. Most notably, clusters 1, 2, 4, and 9 substantially improved in interrater agreement (see Table 4). Cluster 9 made the most substantial improvement. This cluster related to the measurement and determination of the type of movement. The raters agreed to include all student suggestions at the ending of the teaching situation because this represented a coherent phase, which caused the improvements in agreement. However, other clusters (3, 6, 7, 10, 11) seemed to remain too vague to be annotated based on the five most representative words.

Robustness of the Extracted Cluster (RQ1c)

To evaluate the robustness of the extracted clusters, we probed to what extent the clustering algorithm would still yield interpretable and comparable clusters for smaller sample sizes. The baseline for comparison formed the extracted clusters based on the entire dataset (see Fig. 2). As sample sizes in noticing research in science education are typically smaller, subsets of N=43 and N=8 were drawn. The entire clustering approach was performed for these subsets of the data. The resulting cluster embeddings and condensed trees can be seen in Fig. 4. We particularly mapped the extracted clusters based on the top five words to the baseline clusters as extracted with the entire dataset. It is noteworthy that the spatial outline and the actual extracted clusters can be mapped well onto each other. This is even possible for a sample size of only N=8 teachers. The two overarching groups (general and physics-specific) could be identified for the subsamples as well. Based on the condensed trees, some similarities in cluster evolution over different density values can be inferred as well. For example, clusters 8 and 9 seem related in all condensed trees as they evolve from a

Table 4 Values for interrater agreement as measured through Krippendorff's α for each cluster

Clusters	Top Words	α (Coding 1)	α (Coding 2)
Cluster 12	free, fall, folder, definition, topic	**Cluster 12 (0.87)**	**Cluster 12 (0.97)**
Cluster 13	hedgehog, lesson, lesson, hare, last	Cluster 13 (0.66)	**Cluster 9 (0.93)**
Cluster 0	floor, piece of mass, simultaneous, pieces of mass, students	Cluster 0 (0.65)	**Cluster 2 (0.85)**
Cluster 8	type of movement, what kind, teacher (informal), asks, type	Cluster 8 (0.65)	Cluster 1 (0.74)
Cluster 2	feather, screw, tube, vacuum, air	Cluster 2 (0.55)	Cluster 8 (0.73)
Cluster 1	pieces of mass, masses, same, shape, fall	Cluster 1 (0.54)	Cluster 4 (0.68)
Cluster 5	summarize, main hypotheses, main theses, claims, hypotheses	Cluster 5 (0.49)	Cluster 13 (0.66)
Cluster 9	measure, movement, distance, steady motion, certain	Cluster 9 (0.47)	Cluster 0 (0.65)
Cluster 10	screw, weight force, air resistance, higher, different	Cluster 10 (0.29)	Cluster 5 (0.49)
Cluster 4	students, raise arms, raised arms, respond, same	Cluster 4 (0.28)	Cluster 10 (0.47)
Cluster 6	claims, hypotheses, validate, students, asks	Cluster 6 (0.23)	Cluster 11 (0.29)
Cluster 11	air resistance, mass, influence, speed of fall, dependent	Cluster 11 (0.17)	Cluster 7 (0.16)
Cluster 7	summarize, claims, guides, teacher (formal), teacher (informal)	Cluster 7 (-0.05)	Cluster 6 (0.11)
Cluster 3	respond, feedback, again, summarize, guides	Cluster 3 (-0.13)	Cluster 3 (-0.13)

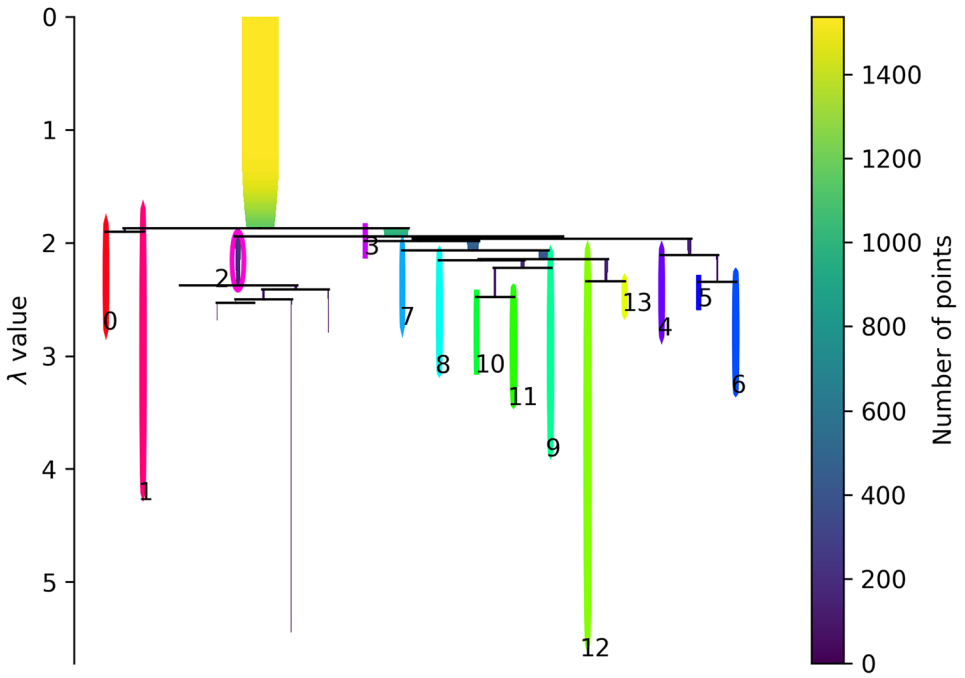

Fig. 3 Condensed tree representation of the extracted clusters

common branch. Both clusters comprise sentences on type of movement which are physics-specific. Interestingly, in Fig. 3, also clusters 10 and 11 fall on the same branch as 8 and 9. This might be attributed to the fact that in clusters 10 and 11 the influence of air resistance on free fall is considered which is closely related to movement as well. While clusters 0 and 1 are linked in Fig. 3 (both include the vacuum tube experiment), this link does not exist in Fig. 4. For these clusters, probably the five most representative words are not informative enough to allow for clear distinction. Clusters 4, 5, and 6 relate to the students' and teachers' actions of posing hypotheses (see Fig. 3). While they neatly evolve from one parent branch in Fig. 3, only one of the respective clusters was present in the smaller samples. However, they also separate early (at low densities) from the other clusters (see Fig. 4).

Further evidence for robustness of the presented clustering approach based on pretrained language models can be gleaned by comparison with a formerly successfully employed clustering approach in science education research that was not based on pretrained language models. To implement a clustering approach based on hierarchical agglomerative clustering, a similar protocol as outlined in Sherin (2013) was followed. However, we did not segment our texts into 100-word chunks, but rather into the sentences that were used as smallest segments. We considered this useful, because we expected the grain size of our clusters (i.e., discernable events in the teaching situation) to be smaller compared to the grain-size of the clusters in Sherin (2013), i.e., explanations. Our overall vocabulary was 2,786 unique words in German language. 232 stopwords were removed. This enabled us to calculate deviation vectors and apply clustering. A number of 14 clusters were found to be reasonable for our data (see Supplementary Material for detailed Table).

Table 5 depicts the resulting clusters with the most representative words for each cluster vis-á-vis the clusters from the pretrained language model-based clustering approach. Most of the resulting clusters can be mapped to the clusters that were extracted based on the pretrained language model-based clustering approach. Cluster 0S[4] thematizes students' formulating hypotheses and summarization by the teacher. This relates to clusters 3, 5, and 7. Clusters 1S and 2S relate to the vacuum tube experiment, where cluster 1S focusses on the execution and cluster 2S on the observation and results. This maps to cluster 2. Cluster 3S relates to the dependency of air resistance and fall velocity, and possibly relates to clusters 10 and 11. Cluster 4S is not entirely clear, and cluster 5S deals with the teacher repeating the experiment, which has no apparent equivalent cluster. Cluster 6S focusses on students' raising their arms and responding, which could be mapped to cluster 4. Cluster 7S relates to the writing down of the definition of free fall, which can be linked to cluster 12. Cluster 8S likely mixes the response of one female student and the remark of another male student, to what extent the speed of light would be reached by a falling object. No apparent link can be made to the pretrained language model-based clusters. Cluster 9S relates

[4] For comparison purposes all clusters which were extracted from the approach by Sherin (2013) were appended with an 'S', e.g., cluster 0S.

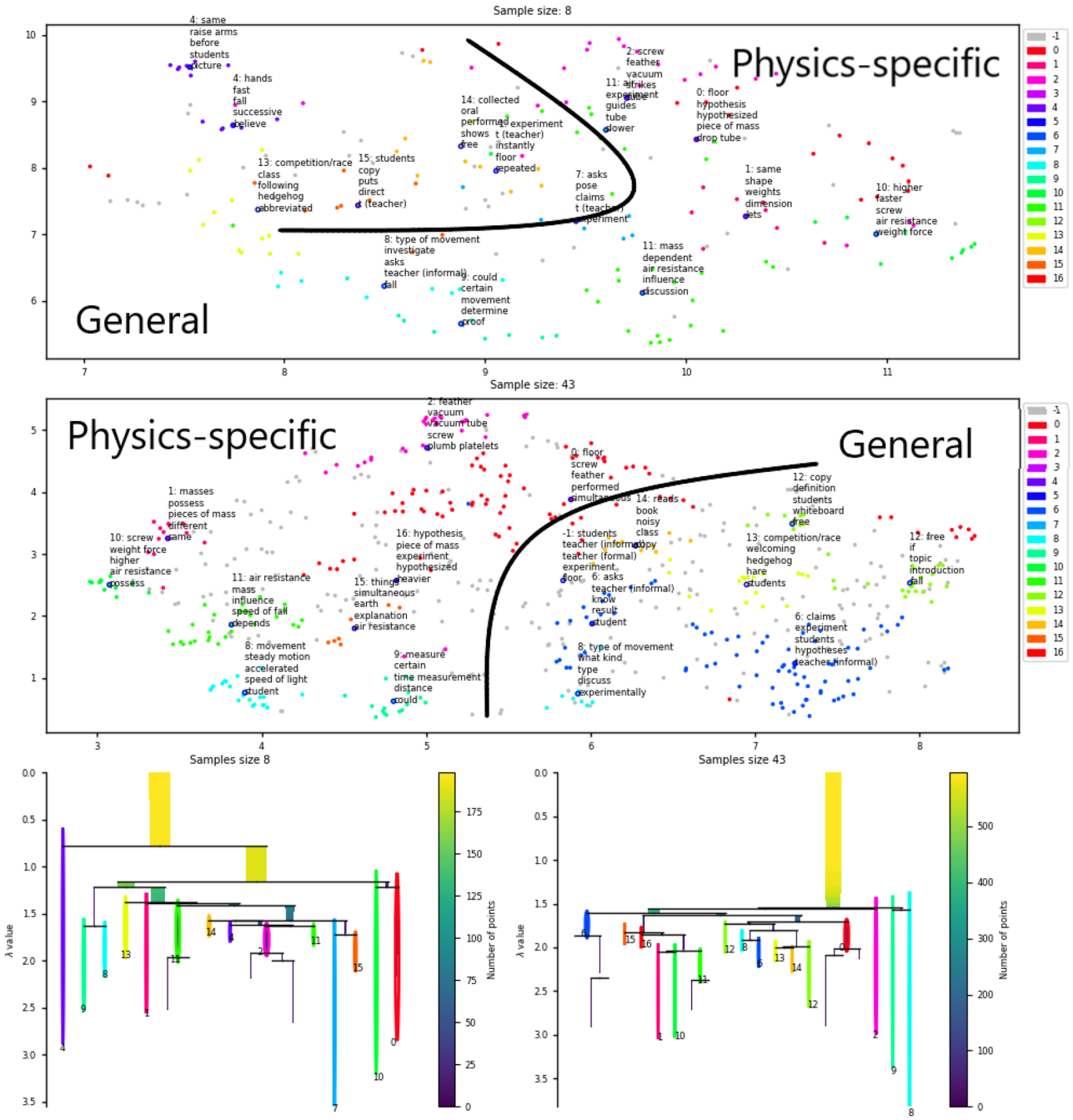

Fig. 4 Scatter plots and condensed trees for cluster evaluation of smaller samples ($N = 8$ and $N = 43$ teachers)

to the experiment with two masses that would most likely map to clusters 0 and 1. Cluster 10S addresses the transition from introduction of the experiments with no apparent corresponding cluster. Cluster 11S, again, deals with the experiment with two masses and links to clusters 0 and 1. Cluster 12S addresses a students' answer to the question about what kind of movement the free fall is. The closest resemblance is with cluster 8. Finally, cluster 13S addresses the vacuum tube experiment, in particular the repetition of the same. No apparent equivalent exists in the pretrained language model-based clustering approach. Finally, we calculated the proportion of sentences in each cluster from the approach by Sherin (2013) that were classified as noise in the pretrained language model-based clustering approach. The respective proportions for each cluster were: 0.46 (0S), 0.28 (1S), 0.45 (2S), 0.40 (3S), 0.60 (4S), 0.60 (5S), 0.48 (6S), 0.38 (7S), 0.62 (8S), 0.35 (9S), 0.61 (10S), 0.32 (11S), 0.34 (12S), and 0.09 (13S). Clusters 4S, 5S, 8S, and 10S

Table 5 Comparison of clusters extracted from the pretrained language model-based clustering approach and the clustering approach that was adopted from Sherin (2013), and the respective mapping

Pretrained language model approach				Approach adopted from Sherin (2013)		
Clusters	**Share**	**Top Words**		**Clusters**	**Share**	**Top Words**
Cluster 0	38	floor, piece of mass, simultaneous, pieces of mass, students		9S	49	let, pieces of mass, objects, arrive, floor, equally, fast, fall, both, simultaneous
Cluster 1	56	pieces of mass, masses, same, shape, fall		11S	57	pieces of mass, lets, shape, experiment, same, different, masses, fall, pieces of mass, two
Cluster 2	246	feather, screw, tube, vacuum, air		1S	144	weight, coin, down, slower, vacuum tube, fall, falls, faster, screw, feather
Cluster 3	10	respond, feedback, again, summarize, guides		2S	77	touches, object, down, body, piece of mass, feather, arrives, screw, floor, first
Cluster 4	42	students, raise arms, raised arms, respond, same		0S	69	students, were, set up, teacher (informal), oral, pose, collected, summarize, together, claims
Cluster 5	11	summarize, main hypotheses, main theses, claims, hypotheses		6S	110	possibilities, pose, asked, same, different, raised arms, raise arms, assumptions, respond, students
Cluster 7	25	summarize, claims, guides, teacher (formal), teacher (informal)		12S	32	says, student, measure, be, could, type, what kind, steady motion, accelerated, movement
Cluster 8	49	type of movement, what kind, teacher (informal), asks, type		3S	81	plays, called, role, depends, influence, speed of fall, on what, dependent, air resistance, mass
Cluster 10	19	screw, weight force, air resistance, higher, different		7S	84	copy, folder, what kind, free, free, definition, type of movement, if, free, fall
Cluster 11	64	air resistance, mass, influence, speed of fall, dependent		4S	555	these, teacher, followed by, discussion, question, time, class, t (teacher), experiment, teacher (formal)
Cluster 12	79	free, fall, folder, definition, topic		5S	65	teacher (formal), two, again, subsequently, was, several, repeated, performed, guides, experiment
Cluster 6	56	claims, hypotheses, validate, students, asks		8S	85	consider, must, observe, means, student (female), answers, shall, speed of light, students (female), student
Cluster 9	54	measure, movement, distance, steady motion, certain	Assignments unclear	10S	97	last, thereupon, directs, beginning, topic, know, experiment, lesson, asks, teacher (informal)
Cluster 13	28	hedgehog, lesson, lesson, hare, last		13S	32	distinct, difference, dropped it, located, vacuum, was, repeated, feather, air, tube

had a particularly large shares of noise-clustered sentences. Interestingly, these clusters could not be easily mapped to the clusters from the pretrained language model-based clustering approach (however, cluster 13S with a particularly low proportion could also not be assigned). They also consistently included generic words (e.g., teacher or students), which were attributed with the noise cluster in the pretrained language model-based clustering approach Fig. 5.

Exploring Textual Organization with the Extracted Clusters (RQ2)

To evaluate to what extent the extracted clusters provide quantifiable information on the textual organization of the written descriptions, we first plot the occurrence of clusters throughout the written descriptions, examine the non-random organization of the clusters, and examine properties of the cluster embeddings. Occurrence of clusters throughout the written descriptions is depicted in Fig. 6. The vertical bars indicate the textual position for the respective maximum occurrence of a certain cluster. The textual positions of the maxima are equally distributed throughout the written descriptions, so that all parts of the written descriptions are attributed with a cluster. Furthermore, the cluster occur at expected positions, given the events in the teaching situation. For example, cluster 13 addressed the beginning of the lesson and it occurred most frequently at the very beginning of the written descriptions (see Fig. 6). In the observed teaching situation, three experiments were carried out one after the other: Free fall of a screw and a spring (cluster 10), free fall of two masses of the same size but different weights (cluster 1) and, finally, free fall in a vacuum tube (cluster 2). Cluster 10 appeared at the beginning of the texts. Cluster 1, in contrast, appeared somewhat later, which maps to the temporal sequence of events in the observed teaching situation, since both experiments that were referenced in these clusters were carried out shortly after each other in the first half of the video. Cluster 2 was addressed frequently and extensively throughout the descriptions. In fact, cluster 2 relates to the most noteworthy experiment (vacuum tube) in the entire teaching situation, which might explain the preponderance in the written descriptions.

A problem (cluster 0) occurred in the second experiment (cluster 1). The shapes of the curves for cluster 0 and 1 match well (as it is also evident in Fig. 3). Before the first experiment, the teacher summarized the "main hypotheses"; the corresponding cluster 5 for this event also occurred chronologically at the beginning. The other actions, i.e., the formulation and discussion of hypotheses (clusters 6 and 7), the reaction to pupils' answers (cluster 3) and the pupils' answers (cluster 4) occurred throughout the teaching situation, which is reflected in the considerably high frequency throughout the first half of the written descriptions in Fig. 6. Cluster 11 related to the discussion of the connection between air resistance, mass and fall velocity. This was also

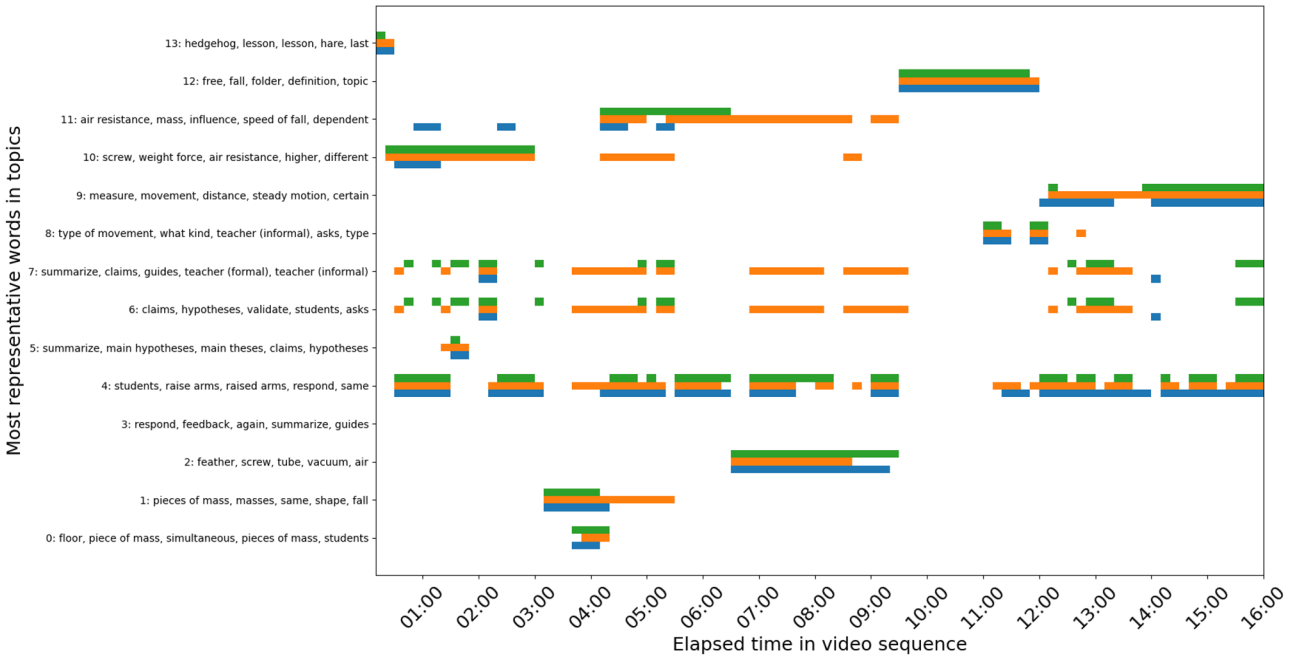

Fig. 5 Codings of video sequence (coding 2) with identified clusters based for three independent raters

related to the experiments seen (observations were described and interpreted; hypotheses regarding the connection were posed and tested). The temporal progression was appropriate, less at the beginning, more towards the middle of the texts. Cluster 12 addressed summarizing the findings of the three experiments. It occurred quite often at the beginning of the descriptions, which does not correspond to the chronological sequence of events. The reason for this could be that some preservice physics teachers began the descriptions with what the goal/result of the sequence was. Otherwise, cluster 12 had its second peak before clusters 8 and 9, which again fits the temporal sequencing of events in the teaching situation. At the end of the sequence, the teacher asked what kind of movement the free fall is. The corresponding clusters were the question itself (cluster 8) and the discussion about it (cluster 9). They occurred most often in the middle of the texts, which corresponded to the end of the written descriptions. The noise cluster (cluster -1) occurred almost equally distributed throughout the written descriptions. The respective counts for each relative position were: 57 (0.0), 71 (0.1), 91 (0.2), 73 (0.3), 79 (0.4), 71 (0.5), 66 (0.6), 68 (0.7), 88 (0.8), 76 (0.9), 20 (1.0). This provides evidence that no particular position in the written descriptions was prone to include more noise sentences compared to other positions. The lower counts at the beginning and end positions resulted from the calculation of the relative position index.

To analyze the sequential interdependence of the clusters, directed network graphs were generated based on the incoming and outgoing connections for each cluster (see Fig. 7). A connection between clusters was established when one cluster occurred in the preceding or receding sentence of another cluster's sentence. Edges (i.e., the interconnections between two clusters) in the networks were weighted by the cluster sizes to highlight connections that appeared often irrespective of the cluster size. The edges with the largest values for the connections were labeled with the respective values (see the small numbers on the edges in Fig. 7(a)). The empirical network graph highlights that certain clusters are central in the network (see Fig. 7(a)). The greatest importance in the network had clusters -1, 2, 4, 6, and 11. In particular, cluster 2 represents the vacuum tube experiment, and cluster 4 the general cluster that students raise their arms and respond. Hence, both physics-specific and general clusters were highly interconnected in the physics teachers' written descriptions.

By analyzing interconnections between two nodes, it appears that clusters -1, 2, 9, 3, and 10 were self-referenced particularly often. Except for clusters -1 and 3, these clusters related to physics-specific events such as the vacuum tube experiment, the type of movement, and the weight and air resistance. Moreover, clusters 8 and 9, clusters 0 and 1, and 1 and 6 are interconnected particularly often. The former two connections directly attribute to the close connection of these clusters in meaning. The connection of cluster 1 (experiment with two masses) and cluster 6 (students' hypotheses) can be explained by the fact that the teacher linked this experiment with posing hypotheses.

Finally, movements of the preservice physics teachers through embedding space by means of addressing specific

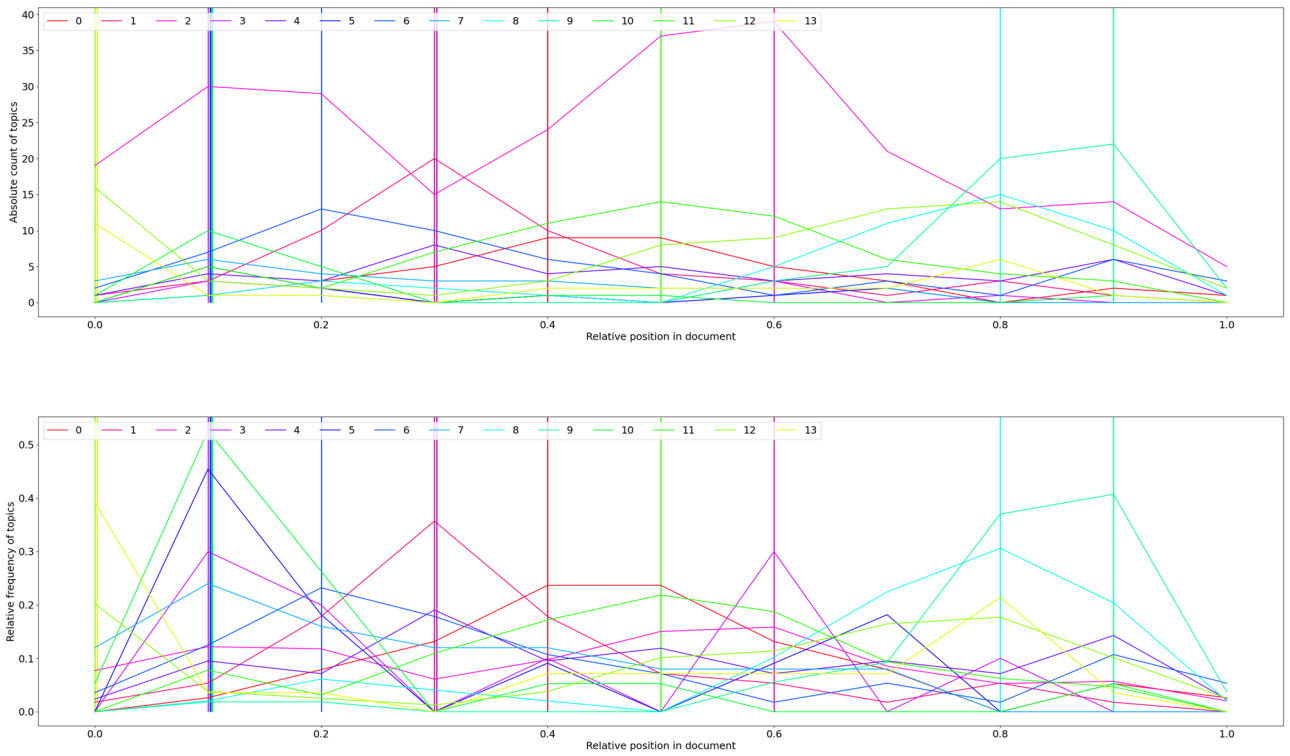

Fig. 6 Progression of extracted clusters relative to other descriptive sentences in the documents. Top: absolute count of occurrence for a cluster at a given document position. Bottom: relative frequency for a cluster at a given document position. Vertical lines indicate the overall peaks in occurrence for each cluster

clusters in their texts should be analyzed with streamline plots (see Fig. 7(b)-(d)). Streamline plots are vector field representations. We define a connecting vector between two sentences that belong to any of the clusters as a "velocity" vector, indicating the movement through cluster embedding space. The resulting vector field is represented in Fig. 7 (b). A tendency to "move" through cluster embedding space in center direction can be verified, because the streamlines direct toward the center. By comparing Fig. 7(b) with (c), which represents a vector field where every velocity magnitude and direction were chosen at random, it is evident that Fig. 7(b) does not represent a random vector field. When positional information is added generate the velocity vector direction (see Fig. 7(d)), the resulting vector field resembles the empirical vector field. The entropies[5] for comparing velocities in plots (b) with (c), and (b) with (d) in x- and y-direction, respectively, were .45 and .28, and .03 and .10. This indicates that the vector field in Fig. 7(d) better approximates the empirical vector field. Thus, the preservice physics teachers do not randomly walk through the cluster embedding space, but rather deliberately compose their texts by attending to the different clusters that were extracted with the pretrained language model-based clustering approach.

Discussion

Attention to learning-relevant classroom events and students' thinking is an important skill for teachers to implement a student-centered pedagogy (van Es & Sherin, 2002b; Chan et al., 2021; Levin et al., 2009). However, assessment of teachers' attention to classroom events is complex, because either the uncertainty of teaching situations is oftentimes related to the inherent complexity of ongoing processes, and describing one's attention processes is intricately tied to teaching knowledge and other filters (Chan et al., 2021). Constructed response formats have been argued to facilitate more authentic assessment of attention processes, and computer-based analytical tools such as ML methods have been found to provide promising means to further our understanding and assessment of complex constructs such as attending to classroom events (Lamb et al., 2021; Zhai et al., 2020). In this paper we sought to examine potentials and challenges of a pretrained language model-based clustering approach for the purpose of extracting patterns, i.e., clusters, in preservice physics teachers' written descriptions of an observed

[5] Note that higher entropies indicate higher mismatch of two distributions.

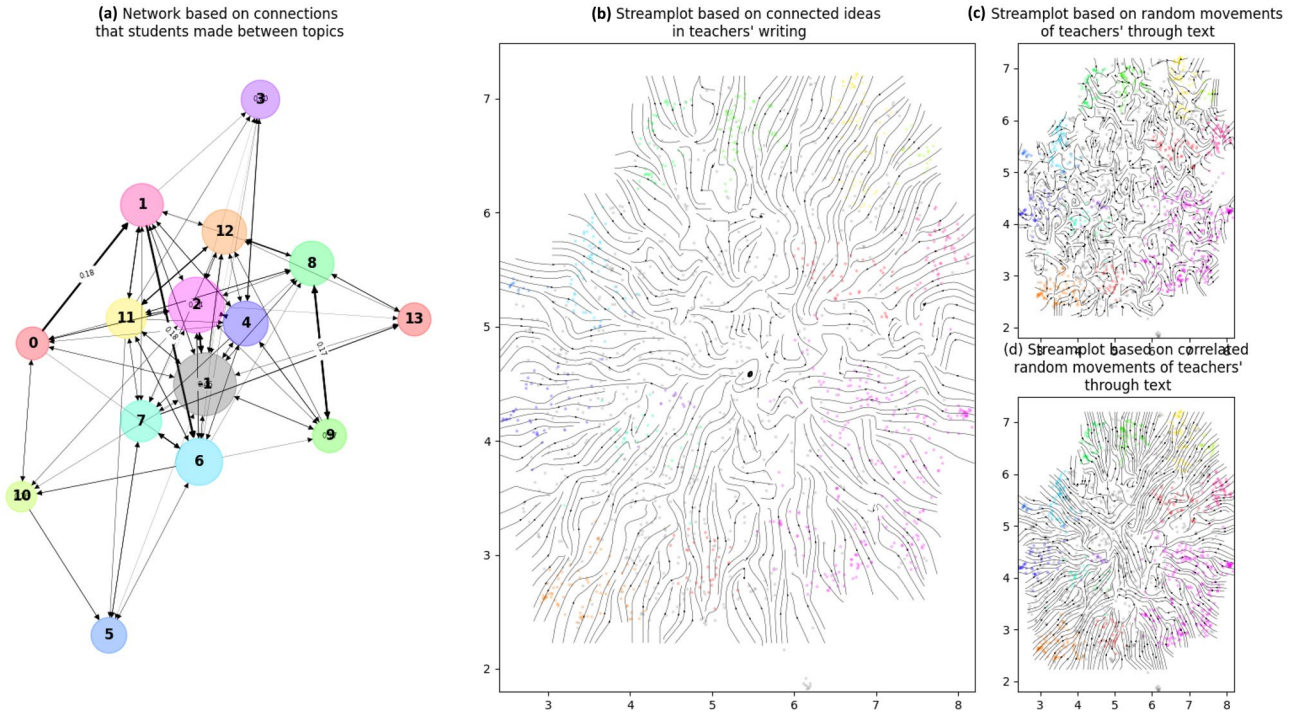

Fig. 7 Directed network graphs of clusters and streamline plots of cluster embeddings: **a** Empirical directed network based on the actual connections between clusters present in the written descriptions; **b** Streamline plot of actual connections between clusters; **c** Streamline plot with randomly distributed directions; **d** Streamline plot where directions are sampled from pool of existing connections.

teaching situation. We examined the validity of the extracted clusters (RQ1) and explored novel ways in which the clusters enable textual analytics that allow to examine quantitative hypotheses on textual organization (RQ2).

To assess the validity of the extracted clusters, the interpretability (RQ1a), the specificity (RQ1b), and the robustness (RQ1c) of the extracted clusters from the pretrained language model-based clustering approach were evaluated. The clustering approach identified a number of 14 clusters that can be grouped into physics-specific and more general clusters. With regard to the contents of the clusters, all clusters could be related to distinct events in the teaching situation. The clusters encapsulated short, concrete events (recapitulating the last lesson), and more abstract ideas (summarizing hypotheses). We found that more specific, event-related clusters could be reliably coded by the raters. However, the more general clusters (related to posing and summarizing hypotheses) that were applicable to several parts of the teaching situation yielded lower reliability scores, and are thus more inferential. The extracted clusters were also robust to variation in sample size and clustering method. A sample size of only N=8 preservice physics teachers' written descriptions yielded a similar distribution of clusters. This likely resulted from grounding the clustering with embeddings from the pretrained language model. A further indication of robustness resulted from the comparison with a previously employed clustering approach in science education research (Sherin, 2013). We found that many of the extracted clusters from the pretrained language model-based clustering approach mapped to the clusters that resulted from the application of the clustering approach by Sherin (2013).

Given that the clusters were well interpretable and could be mapped to the teaching situation, we conclude that the algorithm identified meaningful and distinguishable clusters in the preservice teachers' descriptions. The variety of different foci and abstractness in the extracted clusters is well represented within the different foci of noticing that were summarized by Talanquer et al. (2015). Moreover, the differentiation of more general clusters and physics-specific clusters resonates with the well-established construct of teachers' knowledge, in particular the notions of general pedagogical knowledge and content knowledge (Shulman, 1986; Carlson et al., 2019). The pedagogical content knowledge as an "amalgam of content and pedagogy" (Hume, 2009) might be conceptualized as the relevant knowledge to connect the clusters and discuss pedagogical implications of the physics-specific, and more general clusters. The pretrained language model provides the relevant structures to classify sentences along this dimension. The contextualized embeddings from the pretrained language model facilitate science education researchers means to extract robust clusters in their datasets.

Furthermore, the pretrained language model-based clustering approach integrates the data preprocessing into the modeling and introduces a novel criterion for cluster extraction (stability of clusters over density variation) that provides the human analyst another important measure of appropriate cluster selection.

The findings in the context of RQ1 also indicate that the preservice physics teachers included very general clusters and a comparably large amount of noise clustered sentences. This observation might relate to the finding that novice teachers tend to include broad and general statements in their observations, merely as placeholders (Mena-Marcos et al., 2013). Mena-Marcos et al. (2013) found that more knowledgeable teachers also include more precise statements in their reflections. Furthermore, the preservice physics teachers tended to include only few sentences on each cluster. This indicates that, on average, not much space is spent to describe an event in detail. This might relate to the finding that novice mathematics and science teachers in particular struggle to attend to the specific contents of what was said (Sherin & Han, 2004; Levin et al., 2009; Roth et al., 2011). Rather than describing the concrete hypotheses that the students uttered, many teachers might abstract from the specific contents and simply note that the students posed hypotheses. Yet, developing noticing skills would require the preservice physics teachers to detail the concrete ideas of the students and teacher in order to make an informed evaluation on the substance of the classroom interactions (Levin et al., 2009). However, the unspecific contents might relate to our instructional approach. For example, it should be tested if pre-service teachers can attend to specific events if they can watch the video multiple times and take notes for themselves.

In the context of RQ2 we evaluated to what extent the extracted clusters could be used to assess the textual organization of the written descriptions. The absolute and relative frequency of sentences in certain clusters with regard to their relative position in the written descriptions were analyzed through visual means. We found that the maximum counts for the clusters well matched their expected positions in the teaching situation. This suggests that the preservice physics teachers, on average, compose their written descriptions according to the chronological occurrence of the events in the teaching situation. This finding resonates with episodic memory theory which suggests that free recall of events occurs in temporal order (Conway, 2009; Kahana et al., 2008). Further evaluation of textual organization of clusters by means of network graphs enabled us to document that certain clusters are cued together more closely as would be expected by chance and cluster size. This means that clusters that were semantically or chronologically related were linked by the preservice physics teachers more often. This relates to the contiguity effect, namely that neighboring items (here: events in a teaching situation) are recalled successively (Kahana et al., 2008). Furthermore, streamplot analyses revealed that the preservice physics teachers' movement through cluster embedding space was non-random and dependent on the position in this space. On a local scale, the position in cluster space thus determines the propensity with which the preservice physics teachers' move in a certain direction in this space. Analysis of textual organization can extend assessment of analytical chunks as outlined by van Es and Sherin (2002). van Es and Sherin (2002) differentiate expertise in noticing in a trajectory where experts include more interconnections among their evidences (here: clusters and interconnections between them in the descriptions). The extracted clusters alongside with the network representation directly would yield a quantification of noted events and thus provide a tool to diagnose expertise levels in noticing.

Limitations

Even though the utilization of a pretrained language model allowed us to integrate data preprocessing into the ML-based modeling, there are assumptions on the pretrained language models that have to be critically examined. For example, the resulting contextualized embeddings are determined by the choice of the pretrained language model and cannot be easily adjusted. Problems with the pretrained embeddings have also been reported. Given that they are trained on the Internet, certain biases related to gender or ethnicity are present in the embeddings (Caliskan et al., 2017; Bhardwaj et al., 2020). As such, it has to be critically examined to what extent these biases might be propagated into educational assessments which can be disadvantageous.

Another feature of the pretrained language model-based clustering approach was the algorithm-derived extraction of the number of clusters present in the data. Even though the means to extract the clusters based on the stability over density variation might be an additional tool for researchers to use in order to determine a viable number of clusters, there are still many hyperparameters that can be tuned which yield different numbers of clusters. Given the scope of this paper, we did not systematically vary the hyperparameters to find a final number of clusters. We rather sought to establish that the proposed number of clusters was well interpretable in reference to the observed teaching situation. However, the large proportion of noise datapoints also indicates that a large share of the data is not accounted for in the clustering.

With regard to the contents of the clusters, it was noticeable that the clustering approach did not capture some relevant students' questions from the observed teaching situation into a distinct cluster even though some pre-service teachers included them in their descriptions. Attending to these student questions in the teaching situation required physics knowledge. One student asked whether the different movement of feather and screw (the feather was zig-zagging

whereas the screw moved straight to the ground) could explain the differences in falling time. This is a relevant question that hints at the missing control of variables in the experiment. Some preservice physics teachers included this question in their descriptions, however, no separate cluster appeared to capture it. This is a consequence of the instability and scarcity of this observation as represented in the preservice physics teachers' written descriptions. Omitting contents from the clusters is in fact a goal for unsupervised ML approaches that seek to reduce a complex dataset (Jordan & Mitchell, 2015). For the purpose of assessing skills related to attention to classroom events, adjustments in the clustering procedure should be made to allow more clusters to occur, because the identification of this student question demonstrates close attention to student thinking and an understanding of the problematic aspects of the teaching situation and would be considered to correspond to high levels of noticing skills.

Conclusions

Many domains such as physics embraced ML methods to extract information from unstructured data, e.g., to sift through collider data (unfeasible for humans) to detect outliers (i.e., noise-clustered datapoints) with even the same clustering approach that has been applied in this study (Arpaia et al., 2021). Given the novel potentials to extract information from unstructured data and the increasing availability of this data, science education researchers should critically examine potentials and challenges of these novel ML-based methods in their research contexts as well. This study could show that a pretrained language model-based clustering approach could be used as an assessment tool to analytically induce what teachers attended to in an observed teaching situation and evaluate the potentials of ML for analyzing open-ended responses. We suggest that the applied pretrained language model-based clustering approach can be enhanced by further fine-tuning the pretrained language model weights to science-specific language. This will enable more involved language analytics such as analogical reasoning or synonym detection (Mikolov et al., 2013). It has been shown that pretrained language models capture some knowledge about quantities (e.g., the magnitude of weight of a prototypical dog), or some knowledge graphs about entities (e.g., "Bob Dylan is a songwriter") (Wang et al., 2020; Zhang et al., 2020). In fact, representing natural language into vector spaces can enable novel research approaches to answer research questions in science education research (Sherin, 2013). Once the pretrained language models are trained and publicly available, advanced analytics of written descriptions will be enabled. The presented clustering approach could be applied as a recommender tool to automatically feedback to the teachers which events and contents they addressed and which they missed to pay attention to.

The pretrained language models enabled an informed contextualized representation through embeddings of the language data. Representation of language data through embeddings will also enable researchers to map language to other modalities such as graphical/visual data or mathematical expressions (see: Krstovski & Blei, 2018). Multiple representations and translating between different representations has been considered a constitutive feature for scientific literacy (Brookes & Etkina, 2009). However, it will be necessary to develop theoretically grounded ontologies and epistemologies of what preservice science teachers can observe and how they reason about it (Brookes & Etkina, 2009). Once pretrained language models are developed and ontologies and epistemologies can guide analyses, the presented clustering approach in conjunction with these models can help to make analyses more comparable, scalable, and robust.

With the help of the clustering approach in this study quantitative hypotheses on text composition could be explored. For example, we suspect that preservice physics teachers include general and specific language statements in their written descriptions, are scarce to describe a particular cluster, and compose their texts in chronological order of the appearance of the events. Writing a sentence that can be classified into a specific cluster, to a certain extent, predisposes the teachers to move through the cluster embedding space in certain directions, and noticing certain events predisposes them to also include temporally related events. These hypotheses need to be more systematically tested, because they can enhance assessment of noticing-related cognitive mechanisms such as careful observation and attention to classroom events. We even wonder to what extent mapping the teachers' trajectories through the embedding space can be captured by more physics-involved concepts such as movement through a potential where equations of motion and conservation laws determine the teachers' writing. We are not aware that these hypotheses have been tested. ML-based methods will enable these analyses.

In line with the argument put forth by Singer (2019), we encourage science education researchers to adopt more observational studies that are grounded in data science, assessment and measurement (Singer, 2019). Insights in physics today also come from simulation studies and observational (non-manipulable) experiments. The recent Nobel price of 2021 on complex systems' behavior or insights in astrophysics are testimony to this. We believe that science education researchers can gain novel insights on studied phenomena through ML-based, computational approaches such as the one presented in this study where an unstructured body of textual data is analyzed. Zhai et al. (2020) and Lamb et al. (2021) argued that ML-based computational models can capture the complexity of cognitive

processes and "revolutionize" science assessment. We concur with these arguments and emphasize the necessity to develop an understanding in the science education research community for unsupervised ML approaches and pretrained language models in particular, given the preponderance of observational data that is available in educational contexts. Unsupervised ML methods have thus great potentials to bridge the gap between quantitative and qualitative methods in science education. Pretrained language models, more particularly, capture human-like semantics as measured through implicit association tests and thus represent cognitive structures of humans (Caliskan et al., 2017). Hence, pretrained language models arguably are most promising candidates to model language-based processes. Given that, in our case, the ML-based approach scaled seamlessly (neither human annotations nor preprocessing of the textual data was necessary to extract clusters) and is publicly available to researchers, it would be desirable to increase efforts to share data and models in order to make the most use of the available resources.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10956-022-09969-w.

Funding Open Access funding enabled and organized by Projekt DEAL. This project is part of the "Qualitätsoffensive Lehrerbildung", a joint initiative of the Federal Government and the Länder which aims to improve the quality of teacher training. The program is funded by the Federal Ministry of Education and Research. The authors are responsible for the content of this publication.

Data Availability Please send requests to corresponding author.

Code Availability Please send requests to corresponding author.

Declarations

Ethical Statement All procedures followed were in accordance with ethical standards for research with human subjects, as outlined for example by the American Psychological Association.

Consent Statement Informed consent with all participants was assured.

Disclosure of Potential Conflicts of Interest The authors are not aware of any potential conflict of interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. arXiv.

Arpaia, P., Azzopardi, G., Blanc, F., Bregliozzi, G., Buffat, X., Coyle, L., et al. (2021). Machine learning for beam dynamics studies at the CERN Large Hadron Collider. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 985,* 164652. https://doi.org/10.1016/j.nima.2020.164652

Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: a state of the art, limitations, and future research directions. *International Journal of Educational Technology in Higher Education*, *17*(1). https://doi.org/10.1186/s41239-020-00223-0

Barth-Cohen, L. A., Little, A. J., & Abrahamson, D. (2018). Building Reflective Practices in a Pre-service Math and Science Teacher Education Course That Focuses on Qualitative Video Analysis. *Journal of Science Teacher Education, 29*(2), 83–101. https://doi.org/10.1080/1046560X.2018.1423837

Bhardwaj, R., Majumder, N., & Poria, S. (2020). Investigating Gender Bias in BERT. arXiv.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*(3), 199–231.

Brookes, D. T., & Etkina, E. (2009). "Force,'' ontology, and language. *Physical Review Special Topics - Physics Education Research, 5*(1), 643. https://doi.org/10.1103/PhysRevSTPER.5.010110

Bruner, J. S. (1985). *Child's talk: Learning to use language*. New York, London: W.W. Norton & Company.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science (New York, NY), 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 160–172). Heidelberg: Springer, Berlin Heidelberg, Berlin.

Carlson, J., Daehler, K., Alonzo, A., Barendsen, E., Berry, A., Borowski, A., et al. (2019). The Refined Consensus Model of Pedagogical Content Knowledge. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning Pedagogical Content Knowledge in Teachers' Professional Knowledge*. Singapore: Springer.

Carpenter, D., Geden, M., Rowe, J., Azevedo, R., & Lester, J. (2020). Automated Analysis of Middle School Students' Written Reflections During Game-Based Learning. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 67–78). Cham: Springer International Publishing.

Chan, K. K. H., Xu, L., Cooper, R., Berry, A., & van Driel, J. H. (2021). Teacher noticing in science education: do you see what I see? *Studies in Science Education, 57*(1), 1–44. https://doi.org/10.1080/03057267.2020.1755803

Clifton, R. A., & Roberts, L. W. (1993). *Authority in classrooms*. Scarborough, ON: Prentice-Hall.

Conway, M. A.(2009). Episodic memories. *Neuropsychologia*, *47*(11), 2305–2313. Retrieved from https://www.sciencedirect.com/science/article/pii/S0028393209000645. https://doi.org/10.1016/j.neuropsychologia.2009.02.003

Crespo, S. (2000). Seeing More Than Right and Wrong Answers: Prospective Teachers' Interpretations of Students' Mathematical Work. *Journal of Mathematics Teacher Education, 3,* 155–181.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Education Policy Analysis, 8*(1), 1–44.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv 1810.04805*.

Fenstermacher, G. (1994). Chapter 1: The Knower and the Known: The Nature of Knowledge in Research on Teaching. *Review of Research in Education*, *20*.

Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching, 49*(9), 1181–1210. https://doi.org/10.1002/tea.21054

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Morgan and Claypool: Synthesis Lectures on Human Language Technologies.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press, Cambridge, Massachusetts and London, England. Retrieved from http://www.deeplearningbook.org/

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, pp 8–12.

Hammer, D., & van Zee, E. (2006). *Seeing the science in children's thinking: Case studies of student inquiry in physical science*. Portsmouth, NH: Heinemann Educational Books.

Hao, K. (2019). The AI technique that could imbue machines with the ability to reason: Yann LeCun, Facebook's chief AI scientist, believes unsupervised learning will bring about the next AI revolution: MIT Technology Review.

Hume, A. (2009). Promoting higher levels of reflective writing in student journals. *Higher Education Research & Development, 28*(3), 247–260.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science (New York, NY), 349*(6245), 255–260. https://doi.org/10.1126/science.aac4520

Jurafsky, D. (2003). Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In J. Hay, R. Bod, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–95). Cambridge, MA: MIT Press.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (2nd ed.). Pearson Education, Harlow: Always learning.

Kahana, M. J., Howard, M. W., & Polyn, S. M. (2008). Associative Retrieval Processes in Episodic Memory. *Psychology*, *3*.

Kleinknecht, M., & Gröschner, A. (2016). Fostering preservice teachers' noticing with structured video feedback: Results of an online- and video-based intervention study. *Teaching and Teacher Education, 59,* 45–56. https://doi.org/10.1016/j.tate.2016.05.020

Korthagen, F. A. (1999). Linking Reflection and Technical Competence: the logbook as an instrument in teacher education. *European Journal of Teacher Education, 22*(2–3), 191–207. https://doi.org/10.1080/0261976899020191

Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *WIREs Data Mining and Knowledge Discovery, 1*(3), 231–240. https://doi.org/10.1002/widm.30

Krippendorff, K. (2004). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research, 30*(3), 411–433.

Krstovski, K., & Blei, D. M. (2018). Equation Embeddings. arXiv.

Krüger, D., Parchmann, I., & Schecker, H. (Eds.). (2014). *Methoden in der naturwissenschaftsdidaktischen Forschung*. Berlin and Heidelberg: Springer Spektrum.

Lamb, R., Hand, B., & Kavner, A. (2021). Computational Modeling of the Effects of the Science Writing Heuristic on Student Critical Thinking in Science Using Machine Learning. *Journal of Science Education and Technology, 30*(2), 283–297. https://doi.org/10.1007/s10956-020-09871-3

Levin, D. M., Hammer, D., & Coffey, J. E. (2009). Novice Teachers' Attention to Student Thinking. *Journal of Teacher Education, 60*(2), 142–154. https://doi.org/10.1177/0022487108330245

Luna, M. J., Selmer, S. J., & Rye, J. A. (2018). Teachers' Noticing of Students' Thinking in Science Through Classroom Artifacts: In What Ways Are Science and Engineering Practices Evident? *Journal of Science Teacher Education*, *29*(2), 148–172. https://doi.org/10.1080/1046560X.2018.1427418

Marsland, S. (2015). Machine learning: An algorithmic perspective, second edition edn. Chapman & Hall / CRC machine learning & pattern recognition series, CRC Press, Boca Raton, FL. Retrieved from http://proquest.tech.safaribooksonline.de/9781466583283

Mena-Marcos, J., García-Rodríguez, M. L., & Tillema, H. (2013). Student teacher reflective writing: what does it reveal? *European Journal of Teacher Education, 36*(2), 147–163. https://doi.org/10.1080/02619768.2012.713933

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv (1301.3781v3).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *NIPS, 13,* 3111–3119.

Mitchell, M. (2020). Artificial Intelligence: A guide for thinking humans. Pelican Books.

Munoz-Najar Galvez, S., Heiberger, R., & McFarland, D. (2020). Paradigm Wars Revisited: A Cartography of Graduate Research in the Field of Education (1980–2010). *American Educational Research Journal, 57*(2), 612–652. https://doi.org/10.3102/0002831219860511

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming Biology Assessment with Machine Learning: Automated Scoring of Written Evolutionary Explanations. *Journal of Science Education and Technology, 21*(1), 183–196. https://doi.org/10.1007/s10956-011-9300-9

Odden, T. O. B., Marin, A., & Caballero, M. D. (2020). Thematic analysis of 18 years of physics education research conference proceedings using natural language processing. *Physical Review Physics Education Research*, *16*(1). https://doi.org/10.1103/PhysRevPhysEducRes.16.010142

Odden, T. O. B., Marin, A., & Rudolph, J. L. (2021). How has Science Education changed over the last 100 years? An analysis using natural language processing. *Science Education, 105*(4), 653–680. https://doi.org/10.1002/sce.21623

Putnam, R. T., & Borko, H. (2000). What Do New Views of Knowledge and Thinking Have to Say about Research on Teacher Learning? *Educational Researcher, 29*(1), 4–15.

Rauf, I. A. (2021). Physics of Data Science and Machine Learning. *CRC Press, Boca Raton,*. https://doi.org/10.1201/9781003206743

Rosenberg, J. M., & Krist, C. (2020). Combining Machine Learning and Qualitative Methods to Elaborate Students' Ideas About the Generality of their Model-Based Explanations. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09862-4

Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. Z. (2011). Videobased lesson analysis: Effective science PD for teacher and student learning. *Journal of Research in Science Teaching, 48*(2), 117–148.

Ruder, S. (2019). *Neural Transfer Learning for Natural Language Processing: Dissertation*. Ireland: National University of Ireland.

Rumelhart, D. E., Hinton, G., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323,* 533–536.

Seidel, T., & Stürmer, K. (2014). Modeling and Measuring the Structure of Professional Vision in Preservice Teachers. *American Educational Research Journal, 51*(4), 739–771. https://doi.org/10.3102/0002831214531321

Sherin, B. (2013). A Computational Study of Commonsense Science: An Exploration in the Automated Analysis of Clinical Interview Data. *Journal of the Learning Sciences, 22*(4), 600–638. https://doi.org/10.1080/10508406.2013.836654

Sherin, M. G., & Han, S. Y. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education, 20*(2), 163–183. https://doi.org/10.1016/j.tate.2003.08.001

Sherin, M. G., & van Es, E. A. (2009). Effects of Video Club Participation on Teachers' Professional Vision. *Journal of Teacher Education, 60*(1), 20–37. https://doi.org/10.1177/0022487108328155

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher, 15*(2), 4–14.

Singer, J. D. (2019). Reshaping the Arc of Quantitative Educational Research: It's Time to Broaden Our Paradigm. *Journal of Research on Educational Effectiveness, 12*(4), 570–593. https://doi.org/10.1080/19345747.2019.1658835

Star, J. R., & Strickland, S. K. (2008). Learning to observe: using video to improve preservice mathematics teachers' ability to notice. *Journal of Mathematics Teacher Education, 11*(2), 107–125. https://doi.org/10.1007/s10857-007-9063-7

Taher Pilehvar, M., & Camacho-Collados, J. (2020). Embeddings in Natural Language Processing: Theory and Advances in Vector Representation of Meaning. Morgan and Claypool.

Talanquer, V., Bolger, M., & Tomanek, D. (2015). Exploring prospective teachers' assessment practices: Noticing and interpreting student understanding in the assessment of written work. *Journal of Research in Science Teaching, 52*(5), 585–609. https://doi.org/10.1002/tea.21209

Ullmann, T. D. (2019). Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches. *International Journal of Artificial Intelligence in Education, 29*(2), 217–257. https://doi.org/10.1007/s40593-019-00174-2

van Es, E., & Sherin, M. G. (2002a). Learning to notice: scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education*, *10*(4), 571–596.

van Es, E., & Sherin, M. G. (2002b). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education*, *10*(4), 571–596.

von Aufschnaiter, C., Fraij, A., & Kost, D. (2019). Reflexion und Reflexivität in der Lehrerbildung: 144-159 Seiten / Herausforderung Lehrer_innenbildung - Zeitschrift zur Konzeption, Gestaltung und Diskussion, Bd. 2 Nr. 1 (2019): Herausforderung Lehrer_innenbildung - Ausgabe 2. https://doi.org/10.4119/UNIBI/HLZ-144

Wang, C., Liu, X., & Song, D. (2020). Language Models are Open Knowledge Graphs. arXiv.

Wigner, E. P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. Richard courant lecture in mathematical sciences delivered at New York University, May 11, 1959. *Communications on Pure and Applied Mathematics*, *13*(1),1–14. https://doi.org/10.1002/cpa.3160130102

Wilson, C. D., Borowski, A., & van Driel, J. H. (2019). Perspectives on the Future of PCK Research in Science Education and Beyond. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning Pedagogical Content Knowledge in Teachers' Professional Knowledge* (pp. 289–300). Singapore: Springer.

Wulff, P., Buschhüter, D., Nowak, A., Westphal, A., Becker, L., Robalino, H., et al. (2020). Computer-Based Classification of Preservice Physics Teachers' Written Reflections. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09865-1

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2022). Utilizing a Pretrained Language Model (BERT) to Classify Preservice Physics Teachers' Written Reflections. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-022-00290-6

Xing, W., Lee, H. S., & Shibani, A. (2020). Identifying patterns in students' scientific argumentation: content analysis through text mining using Latent Dirichlet Allocation. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-020-09761-w

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic Coding of Short Text Responses via Clustering in Educational Assessment. *Educational and Psychological Measurement, 76*(2), 280–303. https://doi.org/10.1177/0013164415590022

Zhai, X. (2021). Practices and Theories: How Can Machine Learning Assist in Innovative Assessment Practices in Science Education. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-021-09901-8

Zhai, X., Haudek, K., Shi, L., Nehm, R., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459. https://doi.org/10.1002/tea.21658

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

Zhang, X., Ramachandran, D., Tenney, I., Elazar, Y., & Roth, D. (2020). Do Language Embeddings Capture Scales? arXiv.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.