Technology, Knowledge and Learning
https://doi.org/10.1007/s10758-023-09682-6

ORIGINAL RESEARCH

Annotation Protocol for Textbook Enrichment with Prerequisite Knowledge Graph

Chiara Alzetta[1] · **Ilaria Torre**[2] · **Frosina Koceva**[3]

Accepted: 1 August 2023

Abstract
Extracting and formally representing the knowledge embedded in textbooks, such as the concepts explained and the relations between them, can support the provision of advanced knowledge-based services for learning environments and digital libraries. In this paper, we consider a specific type of relation in textbooks referred to as *prerequisite relations* (PR). PRs represent precedence relations between concepts aimed to provide the reader with the knowledge needed to understand a further concept(s). Their annotation in educational texts produces datasets that can be represented as a graph of concepts connected by PRs. However, building good-quality and reliable datasets of PRs from a textbook is still an open issue, not just for automated annotation methods but even for manual annotation. In turn, the lack of good-quality datasets and well-defined criteria to identify PRs affect the development and validation of automated methods for prerequisite identification. As a contribution to this issue, in this paper, we propose PREAP, a protocol for the annotation of prerequisite relations in textbooks aimed at obtaining reliable annotated data that can be shared, compared, and reused in the research community. PREAP defines a novel textbook-driven annotation method aimed to capture the structure of prerequisites underlying the text. The protocol has been evaluated against baseline methods for manual and automatic annotation. The findings show that PREAP enables the creation of prerequisite knowledge graphs that have higher inter-annotator agreement, accuracy, and alignment with text than the baseline methods. This suggests that the protocol is able to accurately capture the PRs expressed in the text. Furthermore, the findings show that the time required to complete the annotation using PREAP are significantly shorter than with the other manual baseline methods. The paper includes also guidelines for using PREAP in three annotation scenarios, experimentally tested. We also provide example datasets and a user interface that we developed to support prerequisite annotation.

Keywords Text annotation · Annotation protocol · Knowledge engineering · Educational textbooks

Extended author information available on the last page of the article

Published online: 21 September 2023

1 Introduction

Textbooks play a central role in the learning process despite the recent worldwide growth of distant learning, possibly because they provide deep knowledge about a subject and help consolidate learning outcomes (Carvalho et al., 2018). Their availability in academic digital libraries and repositories provides learners with the opportunity to access them at lower cost (Eighmy-Brown et al., 2017) and exploit further services (Atkinson, 2020). Regarding the latter, recent advances in artificial intelligence and natural language processing have opened up possibilities for automating the extraction of knowledge embedded in educational textbooks. Specifically, our focus is on extracting concepts and prerequisite relations (PRs) between them. Together, these components form a graph that represents the content structure (Wang et al., 2016; Lu et al., 2019). We refer to it as the prerequisite knowledge graph (PR graph, for short), i.e., a graph composed of concepts as nodes and PRs as edges.

The availability of datasets annotated with PRs can support the development of supervised methods for prerequisite learning and can also support semi-supervised approaches and the evaluation of non-machine learning methods. However, the existing literature on prerequisite relations lacks high-quality resources and well-defined annotation criteria. As a result, the datasets generated are difficult to compare and reuse, and often show low inter-annotator agreement scores (Chaplot et al., 2016; Gordon et al., 2016; Fabbri et al., 2018). This issue could be addressed by adopting annotation protocols for PRs since they would provide specifications on the annotation criteria and rationale, along with guidelines for their application (Fort et al., 2011; Pustejovsky and Stubbs, 2012). Furthermore, we intend to address the lack of approaches that rely solely on the text for the annotation of RPs, without relying on the prior knowledge of the annotators. Such an approach would allow for annotations that faithfully reflect the content of the textbook, as current methods rely heavily on the annotators' knowledge of the subject matter.

Our research aims to tackle these challenges by designing an annotation protocol that addresses the following *goals*:

1. Designing a knowledge engineering procedure for the annotation of prerequisites and the creation of PR datasets, with the aim of reducing the ambiguity of the annotation task and thus achieving more reliable and consistent datasets;
2. Implementing a *textbook-driven annotation* procedure aimed to annotate concepts and prerequisites based solely on the content of the text, rather than relying on the annotator's domain knowledge. By adopting an in-context annotation approach, we seek to explicitly identify the instructional design principles that underlie the organization of content in the textbook, specifically identifying which concepts serve as prerequisites for others.

To achieve these goals, we designed PREAP (PRErequisite Annotation Protocol) using an iterative design methodology. We evaluated the final version of the protocol in a mixed quantitative-qualitative study involving education experts. The study aimed to answer the following Research Questions (RQs):

- RQ1: to what extent PREAP succeeds in obtaining PR-annotated datasets that are reliable in terms of completeness and accuracy (the former intended as the extent to which the annotations cover the relevant information and the latter as the correctness and precision of PRs);

- RQ2: to what extent PREAP succeeds in obtaining textbook-driven annotations.

The contribution we make in this paper lies in the following points:

(1) A knowledge engineering procedure for prerequisite annotations that led to an increased agreement between annotators and higher accuracy compared to existing methods in the literature;
(2) A novel methodology that binds PR annotation to the textbook in order to make explicit and annotate not only the content of the textbook but also the underlying structure of prerequisites.

In addition to these methodological contributions, we also provide resources that are publicly available on GitHub: the dataset resulted from a case study annotation project described in this paper, a tool for the annotation and analysis of PRs, and the recommendations for applying the protocol in different annotation scenarios.

The remainder of the paper is organised as follows. In Sect. 2 we review related works on prerequisite annotation. Section 3 introduces PREAP protocol, focusing on its design process and annotation principles, while Sect. 4 presents PREAP evaluation. Section 5 describes the application of the protocol in an annotation project case study. Section 6 extends the case study, comparing the datasets produced with three options of the protocol and using them to train a machine learning (ML) system for automatic prerequisite extraction. Section 7 concludes the paper and Sect. 8 describes the datasets and the other shared resources.

2 Related Work and Background

The content of educational texts such as textbooks is typically structured and presented according to instructional design principles that authors intuitively or deliberately apply (Gagne, 1962; Ruiz-Primo, 2000; Council, 2000). For example, arithmetic and algebra textbooks typically introduce the concept of "addition" before explaining "multiplication" as it is useful to refer to the former when introducing the latter. Thus, "addition" can be said a prerequisite of "multiplication" from a teaching point of view.

As in (Liang et al., 2017), we define a *prerequisite relation (PR)* as a *binary dependency relation connecting a prerequisite and a target concept where the former is a concept that has to be known in order to understand the latter*. In other words, the prerequisite concept provides the prior knowledge required to understand the target concept. The set of PRs in a textbook can be represented as a knowledge graph, resembling Educational Concept Maps (Novak et al., 2008) as sketched in Fig. 1, where concepts are nodes and edges represent prerequisite relations between them. The edge in the graph, for instance, between concept B (e.g., "addition") and C (e.g., "multiplication"), is read as B is prerequisite of C ("addition is prerequisite of multiplication", $B \prec C$).

2.1 Concepts and Prerequisite Relations

The term "*concept*" refers in general terms to an abstract and general idea conceived in the mind (Carey, 2009). Given such a broad definition, the nature of concepts is a matter of debate in many fields. We refer to "*concepts*" similarly to other works in the literature on prerequisite annotation (Talukdar and Cohen, 2012; Wang et al., 2016; Liang et al.,

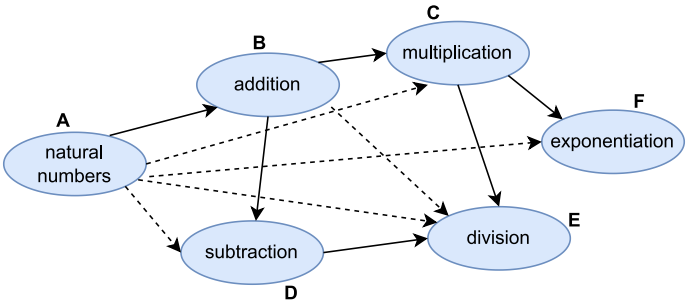


Fig. 1 Example of PR knowledge graph representing arithmetic concepts as nodes and their PRs as edges. Dashed edges represent transitive PRs. The label of edges is "prerequisite", e.g., A is prerequisite of B, A is prerequisite of D

2017; Pan et al., 2017a; Zhou & Xiao, 2019; Adorni et al., 2019; Alzetta et al., 2019; Limongelli et al., 2015; Xiao et al., 2022) that basically associate concepts to terms, intended as lexical units composed of single or multiple words with unambiguous sense in the given context. Similarly to (Chau et al., 2020b; Wang et al., 2021), we identify terms representing concepts from an educational text as a subset of words therein (more precisely, noun phrases) that convey a domain-specific meaning Cabré (1999). This perspective borrows from the approaches in terminology research, according to which the terminology of a domain provides as many lexical units as there are concepts in its subspace (Sager, 1990), and also from information extraction, which addresses automatic keyword extraction (Augenstein et al., 2017; Shen et al., 2014; Martinez-Rodriguez et al., 2020). Computational linguistics and natural language processing specifically tackle keyword extraction from unstructured resources, that is, text, defining either (i) pattern-based linguistic approaches, which employ syntactic parsing to identify domain terms among short noun phrases in the text Faure and Nedellec (1999); Hippisley et al. (2005); Golik et al. (2013), or (ii) statistical approaches, that assign a *termhood* degree to words by relying on distributional properties Suresu and Elamparithi (2016); Rani et al. (2017); Zhao and Zhang (2018) or on sentence-level contextual information Cimiano and Völker (2005); Velardi et al. (2013); Dell'Orletta et al. (2014). Only a few works have considered textbooks as a source for extracting concepts Wang et al. (2015); Labutov et al. (2017).

In a *prerequisite relation*, the concepts involved are referred to as the prerequisite concept and target concepts respectively, meaning that the prerequisite concept must be understood before the target concept (Liang et al., 2017, 2019). According to Hübscher (2001), the term "prerequisite" has at least two meanings. First, it signifies a pedagogical relationship between two elements that a student should learn. Secondly, it indicates a formal mechanism that can be used to partially order two instructional units (such as concepts, pages, exercises, or similar) into a sequence. Early studies in instructional design (Gagne, 1962; Ausubel et al., 1968; Carey, 1999; Merrill, 2002) emphasized the significance of prior knowledge in the process of learning new concepts. These studies proposed that learning occurs in a sequential manner, building upon existing knowledge.

This pedagogical perspective paved the way for representing educational content in the form of graph and concept map structures. Graph structures inherently represent interlinked concepts and are easily exploited for computer-based applications (Novak, 1990; Gruber, 1993). For example, in automatic lesson plan generation, graph structures enable the inclusion of multiple paths between components to accommodate students' needs and interests (Brusilovsky & Vassileva, 2003; Yu et al., 2021).

2.2 Prerequisite Annotated Datasets

A prerequisite annotated dataset is a collection of concept pairs where the information concerning the presence or absence of a PR is explicitly indicated by assigning a 'prerequisite' or 'non-prerequisite' label to each pair (Wang et al., 2016; Chaplot et al., 2016; Gordon et al., 2016). PR graphs usually display only "prerequisite" edges, as in Fig. 1. These PR-annotated datasets serve two main purposes: training and testing ML algorithms (Gasparetti et al., 2018; Liang et al., 2018; Li et al., 2019) and evaluating PR extraction methods against a gold dataset (Liang et al., 2015; Adorni et al., 2019). Ultimately, the aim of PR datasets is to serve as knowledge bases for developing advanced services (Talukdar and Cohen, 2012; Liang et al., 2019; Changuel et al., 2015). This demands reliable and quality PR-annotated datasets. However, the availability of high-quality datasets annotated with PRs between educational concepts is limited, due to the insufficient accuracy of automatically created ones, the high effort required for their manual construction, and the shortage of reliable and systematic annotation procedures. Even more critical is the fact that existing datasets vary with respect to the annotated items and the annotation principles. In fact, PRs can concern prerequisite relations between university courses (Yang et al., 2015; Liang et al., 2017; Li et al., 2019), MOOCs (Chaplot et al., 2016; Pan et al., 2017a; Roy et al., 2019; Zhao et al., 2020), MOOC videos (Pan et al., 2017c; Huang et al., 2021; Wen et al., 2021; Xiao et al., 2021), learning objects (Gasparetti, 2022), scientific databases (Gordon et al., 2017) or Wikipedia pages (Talukdar and Cohen, 2012; Gasparetti et al., 2018; Miaschi et al., 2019; Zhou & Xiao, 2019; Sayyadiharikandeh et al., 2019; Bai et al., 2021; Hu et al., 2021), all represented as PR relations between concept pairs. Alternatively, concepts can be relevant domain terms acquired from a text, as in (Wang et al., 2016; Lu et al., 2019; Adorni et al., 2019; Alzetta et al., 2019; Chau et al., 2020b; Wang et al., 2021), and in our approach.

2.3 PR Annotation

Automated methods. The most used methods for the automatic identification of PRs are based on relational metrics (Liang et al., 2015; Adorni et al., 2019) and machine learning approaches. (Talukdar and Cohen, 2012; Liang et al., 2019; Manrique et al., 2018; Gasparetti, 2022; Xiao et al., 2021). Among ML approaches, we distinguish between approaches exploiting link-based features (Gasparetti et al., 2018; Wen et al., 2021), text-based features (Miaschi et al., 2019; Alzetta et al., 2019), or a combination of the two (Liang et al., 2018; Hu et al., 2021). The former refers to ML approaches that exploit the structure of the source text provided by links, in the sense of connections, between concepts and portions of contents (e.g., Wikipedia graph of categories, DBpedia links, organization in sections

and paragraphs, etc.), while the approaches exploiting text-based features use only features from the raw text (e.g., bag-of-words and word embeddings).

The most widely used and effective methods, such as RefD (Liang et al., 2015), rely on external knowledge. Recently, the task has been addressed employing neural language models (Angel et al., 2020; Li et al., 2021; Bai et al., 2021). However, automatic methods for concepts and PR extraction are generally still not good enough to be used in knowledge-based services for learning support (Chau et al., 2020b) and still need gold datasets for evaluation, thus manual annotation is still a crucial task in this field.

Manual methods. Manual PR annotation is commonly carried out by recruiting domain experts (Liang et al., 2015, 2018; Fabbri et al., 2018) or graduate students (Wang et al., 2015; Pan et al., 2017b; Zhou & Xiao, 2019) to annotate all pairwise combinations of predefined concepts (Chaplot et al., 2016; Wang et al., 2016; Li et al., 2019; Zhou & Xiao, 2019) or a random sample of that set (Pan et al., 2017c; Gordon et al., 2017; Gasparetti et al., 2018). Asking annotators to autonomously create concept pairs based on their domain knowledge, as in Lu et al. (2019), is less common. These strategies aim to identify PRs between concepts in the given domain without accounting for concept organization in the text. In fact, annotators generally rely on their prior domain knowledge (Talukdar and Cohen, 2012; Chaplot et al., 2016; Wang et al., 2016; Li et al., 2019), at most checking dubious cases on a given collection of documents (Gordon et al., 2016), and not on a specific text. In PREAP we use an approach called *textbook-driven* in-context annotation of PR relations between pairs of concepts, that takes into account how concepts are organized in the annotated text. Differently from the approaches that annotate PR relations in a given domain unbounded from a specific text, this approach does not fit the goal of developing intelligent tutoring systems (ITS) that can be used regardless of the textbook chosen. Conversely, textbook-driven annotation is thought to produce training and testing datasets for NLP tools that are mostly used to extract information from corpora, since it is essential to feed these models with training examples that can be associated with a text passage written in natural language. The two approaches can be said to be complementary. In Sect. 4 we will compare datasets produced by using the PREAP approach against datasets produced through pairwise combinations of predefined concepts, showing that the former approach not only better expresses the content explained in the text, as expected, but also improves the coherence and consistency between annotations produced by different annotators.

Indeed, PR-annotated datasets frequently report low annotation agreement and performance variability of systems trained on such data (Chaplot et al., 2016; Gordon et al., 2016; Fabbri et al., 2018; Alzetta et al., 2020), possibly due to the lack of reproducible procedures for creating them. In fact, although properly defining an annotation task is vital to reduce annotation inconsistencies (Ide & Pustejovsky, 2017), the available PR-annotated datasets are mostly poorly documented, and *annotation guidelines* tend to be absent or fairly basic, mostly relying on a naive definition of prerequisite relation. PREAP tries to fill this gap in PR literature as it defines a systematic procedure for annotating educational texts: we could not find any other knowledge engineering procedure for prerequisite annotation and PR dataset creation, while methods exist for the mere task of concept annotation, including a recent one from (Wang et al., 2021).

Additionally, to improve the documentation of the released PR datasets, PREAP recommends that they are published and described following the principles of the Linked Data paradigm, a W3C standard for sharing machine-readable interlinked data on the Web using standard vocabularies and unique identification of resources (URI/IRI). The linked data approach has not been used much in PR annotation, while it is very common in other types of annotation. We freely distribute our datasets described accordingly.

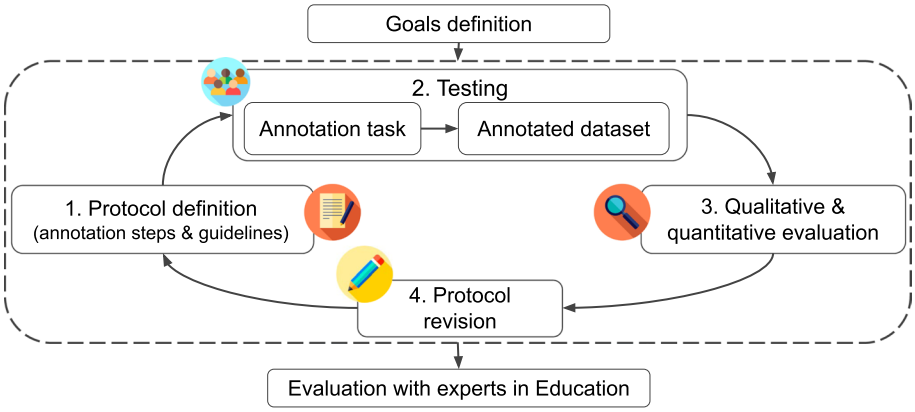


Fig. 2 Iterative design of PREAP annotation protocol

3 PREAP Protocol: Design and Description

3.1 Design of PREAP Protocol

The design of the PREAP protocol for manual annotation of prerequisite relations has been guided by the MATTER framework (Pustejovsky, 2006; Ide & Pustejovsky, 2017), which defines an iterative methodology to obtain annotated corpora for machine learning tasks. We took into account in particular the recommendations for model and annotation definition.

Figure 2 shows the process that led to the definition of the PREAP protocol. The *Goals definition* block in the figure represents the initial input for the overall iterative design of the protocol. The goals have been defined through the experience and groundwork (Adorni & Koceva, 2016; Alzetta et al., 2018; Alzetta et al., 2019; Adorni et al., 2019; Alzetta et al., 2020a, 2020b) that guided toward the identification of the goals stated in the Introduction in Sect. 1.

The central part of the figure shows the four-step cycle for the definition of the protocol: *definition-testing-evaluation-revision*. The first step, *protocol definition*, includes input decisions for that cycle (i.e., annotation and revision methods). It is followed by *testing*, which involves the *annotation task* performed by annotators according to the annotation protocol and the resulting *annotated dataset*. The third step is *evaluation*, where both the annotation process and the datasets are evaluated using quantitative (inter-agreement and dataset analysis) and qualitative (focus group with annotators) methods in order to identify unclear instructions. The outcome of the evaluation drove the *revision* of the protocol and the start of a new cycle. For consistency, the annotation tasks for each cycle were performed on the same introductory computer science textbook (Brookshear & Brylow, 2015). The annotators involved in the annotation tasks were four master's students in Computer Science, different in each cycle.

The current version of the PREAP protocol, presented in Sect. 3.2, is the result of three iterative cycles that lasted about two years. As a final step, PREAP underwent an *Evaluation with experts in Education* reported in Sect. 4.

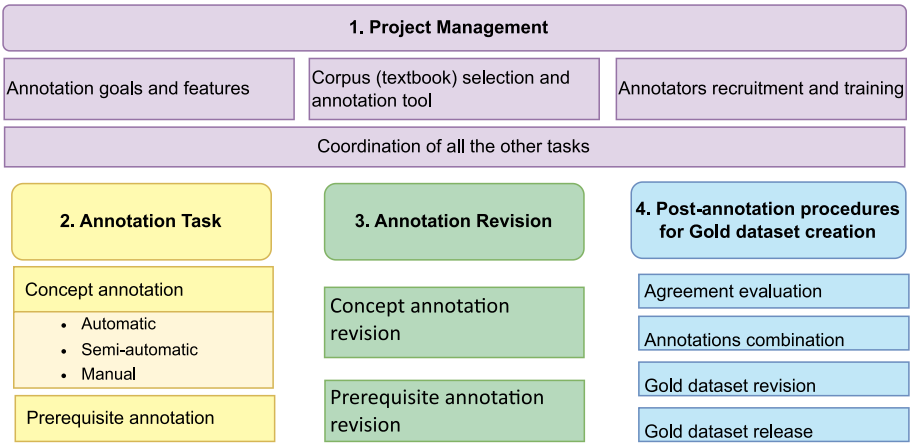


Fig. 3 PREAP tasks to carry out a PR Annotation Project

To support the testing and evaluation phases, we built a tool, PRAT, described in Sect. 5. PRAT provides an interface for manual annotation of prerequisites and facilities for quantitative and visual analysis.

3.2 Description of PREAP Protocol

The main principle addressed in PREAP is the *textbook-driven* annotation approach: annotations are anchored to the text portion where a PR relation occurs between concepts. As a result, the application of PREAP results in the creation of a gold-PR dataset (or *gold dataset*, for short). This dataset is annotated with PR relations following the systematic annotation procedure defined by PREAP. The dataset can be represented as a PR graph whose nodes are the concepts explained in the textbook and the edges are the prerequisite relations between the concepts expressed in the text. The dataset can be directly employed in services for augmented textbooks that demand high-quality manual annotation, or as ground truth data for the development and evaluation of automated methods for prerequisite extraction.

To attain a gold dataset, the person or the team *managing the annotation project* must set up and coordinate the set of tasks shown in Fig. 3. PREAP provides specifications for dealing with all the tasks: (1) Project Management, (2) Annotation Task, (3) Annotation Revision, and (4) Post–annotation procedures for gold dataset creation.

The *Project Management task* concerns supervising the whole project and making decisions especially regarding task (4). Tasks (2) and (3) are performed by the *annotators* recruited for the specific annotation project. Below we describe the main points of each task.

3.2.1 Project Management

The design and setup of an annotation project are handled by the manager(s) of the annotation project (Fort et al., 2011). As shown in Fig. 3, the decisions concern: *(i) annotation goals and features* (what the annotation project is intended for, domain and language); *(ii) corpus (textbook) selection and tool* to be used for the annotation, if any; *(iii) annotators recruitment and training* (selecting annotators with adequate expertise to properly understand the textbook content; setting up a trial task to train the annotators and then assess their understanding of the guidelines); *(iv) coordination of all the other tasks* described below.

3.2.2 Annotation Task

The text annotation phase, performed by annotators and supervised by the annotation manager, is the core part of a PR annotation project. The annotation recommendations are systematised within the *Annotation Manual* which comprises two complementary resources: the *Annotation Guidelines* (AG), describing how the annotation process should be carried on, and a list of *Knowledge Elicitation Questions* (KEQ), aimed at clarifying dubious cases through questions and examples. Prior to the actual annotation, both AG and KEQ should be given to annotators in a trial annotation task where the manager(s) of the project can check whether the annotators interpreted the instructions correctly. Training annotators is recommended in annotation projects to reduce the biases caused by annotators' background knowledge and subjective interpretation of the task instructions (Hovy & Lavid, 2010). The manual remains accessible to the annotators throughout the entire annotation process.

As shown in Fig. 3, the Annotation task encompasses the *concept* and *prerequisite annotation* subtasks.

(i) Concept annotation. The Annotation Manual provides a definition of what should be regarded as concept in the annotation task and it also provides examples in order to increase the reliability of the identification: therein, concepts are described as the building blocks of learning, namely what a student should understand in order to comprehend a subject matter. Depending on the topic and detail level of the given textbook, concepts can be general (e.g., algebra, geometry, mathematics etc.) or very specific (e.g., radius, integer multiplication, fraction denominator). Either way, they are domain terms represented in texts as lexical entities (more precisely, noun phrases) constituted by a single or multi-word term.

In PREAP, the identification of domain concepts in the text (see Fig. 3) can be carried out in two ways: *autonomously* or *simultaneously* with the prerequisite annotation task, based on the project management decisions. In the former approach, the list of concepts, i.e., the *terminology*, can be obtained through manual extraction or (semi)automated extraction approaches. In these cases, the work of Chau et al. (2020a) proved the benefit of including the evaluation of a domain expert to refine the list of concepts.[1] Alternatively, in the simultaneous approach, the identification of domain concepts is performed by the annotators alongside the task of prerequisite annotation. This option seems appealing for saving time. However, it is likely to result in less consistent annotations and lower agreement, as shown in our experimental tests reported in Sect. 6, which thus demand heavy revision and

[1] Sect. 5 presents an Annotation Project case study where more details on this scenario are provided.

time consumption. Hence, our recommendation is to adopt this option carefully, e.g., when obtaining a rich although less consistent annotation complies with the project goal.

(ii) Prerequisite annotation. The *in-context annotation* approach of PREAP requires annotators to perform the annotation of prerequisite relations while reading the educational text. This implies identifying PRs based on the explanations provided by the textbook rather than relying on the annotator's background knowledge about the topic. Differently from existing PR datasets (ref. Sect. 2.2), PREAP aims to capture the view of the textbook's author on which concepts should be presented, and how they should be presented, to allow students to understand the target concepts. This approach is referred to as *textbook-driven* annotation, as it aligns with the content and organization of the textbook itself.

The specific properties of PRs, as intended in PREAP, should be preserved in the annotation to avoid invalid relations from a structural and semantic point of view. Specifically, PRs are binary relations characterised by the following properties: (i) irreflexive: if A and B show a PR relation, A must be different from B; (ii) asymmetry: if $A \prec B$, the opposite cannot be true (e.g., if *network* $\prec$ *internet*, *internet* $\prec$ *network* can't be true); iii) transitivity: for every A, B, and C, if $A \prec B$ and $B \prec C$, then $A \prec C$ (e.g., if *computer* $\prec$ *network* and *network* $\prec$ *internet*, then *computer* $\prec$ *internet*).

Note that, differently from (Chaplot et al., 2016; Wang et al., 2016; Li et al., 2019; Zhou & Xiao, 2019), annotators are not required to explicitly annotate non-PR pairs. In the proposed *textbook-driven* annotation approach, non-annotated transitive relations (dashed edges in Fig. 1) remain implicit, but they can be inferred using PR properties. Specifically, transitivity allows retrieving PRs that derive from paths involving intermediate concepts; in addition, asymmetry can be used to infer those non-PRs represented by inverse relations.

Considering the semantic properties of the relation, an extension of PREAP would be accounting for different strengths of PR as a weight assigned by the annotator to each detected relation. Consistently with the PR annotation approach, a *strong* weight should be assigned if the prerequisite is described in the textbook as absolutely necessary to understand the target concept, while a *weak* weight could be used to indicate that the prerequisite is useful for a deeper comprehension of the concept but not strictly necessary.

To guide the annotation of PRs, KEQs offer examples of lexical taxonomic relations that can easily subtend PR, such as hyponyms, hypernyms and meronyms, or semantic relations like causal or temporal relations. In fact, the goal of KEQ is to provide examples in order to build a shared understanding of the PR interpretation. The instructions in KEQ for assigning PR weights, are a first draft whose results are still under evaluation and possibly subject to future refinements.

3.2.3 Annotation Revision

Manual annotation is known to be error-prone, as well-recognised in the literature (Fort et al., 2011; Dickinson, 2015; Wang et al., 2021) and also studied in our own work (Alzetta et al. 2020a). Therefore, PREAP recommends a revision phase (*Annotation revision* task in Fig. 3) after the annotation task: searching for errors and inconsistencies is aimed at improving the reliability and consistency of the annotations (Plank et al., 2014).

In line with the Annotation task, the Revision phase of PREAP consists of two subtasks: *Concept annotation revision* and *Prerequisite annotation revision*. For both subtasks, PREAP recommends "in-context revision" in order to comply with the *textbook-driven* annotation approach.

(i) Concept annotation revision. When concept annotation is conducted autonomously using semi-automatic or automatic extraction tools, it is recommended the support of experts to review the set of concepts. According to Chau et al. (2020a), domain experts are best suited for this task as they provide high-quality annotations, are less burdened by difficult annotation instances and are more capable to spot erroneous automatic annotations than non-experts Lee et al. (2022). The manager of the annotation project provides both AG and KEQ to the experts so that they can revise the semi(automatically) extracted concepts based on the examples and definitions of the PREAP manual. The validated set can be then provided to the annotators for the annotation of prerequisite relations.

When concept annotation is simultaneous to PR annotation, annotators who earlier identified and annotated the concepts should revise the set using the approach for PR revision that will be explained below.

(ii) Prerequisite annotation revision. To comply with the in-context annotation approach, annotators are required to read again the portion of text where they found a PR relation before making the final decision of approving, excluding or modifying the relation. While reading the textual context, each annotator reconsiders her/his own annotations and checks if the inserted pairs comply with the formal and semantic requirements of prerequisite relations described in the annotation manual. Note that, like PR annotation, PR revision is carried out by each annotator individually.

Since revision is a time-consuming process, a convenience approach to balance the benefit of revision and its cost might be revising only a subset of annotations, specifically PR pairs that are more likely to contain annotation errors, i.e., those with lower agreement. This is because the highest chance of finding errors lies in phenomena that are rarely annotated (Eskin, 2000). In this case, the criteria for selecting the PR sample to be checked should not be shared with annotators to avoid biased revisions. The same approach can be used also in the case of simultaneous annotation of concepts and PRs. However, if incorrect concepts are identified, it is necessary to revise all the direct and indirect PRs related to those concepts.

3.2.4 Post-annotation Procedures for Gold Dataset Creation

Once the revision task is completed, the manager(s) of the annotation project has to undertake actions toward the creation of the gold dataset as a result of the combination of the revised annotations. The main actions are shown in Fig. 3 and explained in the following.

(i) *Agreement evaluation*, using agreement metrics to assess the homogeneity and consistency of annotations produced by different annotators;
(ii) *Annotations combination*, using appropriate combination criteria;
(iii) *Gold dataset revision* after annotations combination (e.g., looking for loops in the resulted PR graph);
(iv) *Gold dataset release*: meta-annotation and documentation, to enable sharing and reuse of the resulted PR graph.

The first three actions are unnecessary if only one annotator has been recruited, although this is generally not recommended to minimize errors.

The use of agreement metrics is recommended to quantify the consistency and homogeneity of annotations produced by different annotators:[2] while disagreement can be due to multiple factors, as long studied in the literature (Bayerl & Paul, 2011), high agreement is generally assumed as an indicator of common understanding of the annotation specifications as well as of the specific phenomenon to annotate (Artstein & Poesio, 2008). Thus, in case of low agreement, the annotation manager should check the annotators' understanding of the annotation specifications and investigate any possible issues with the annotation instructions (Di Eugenio & Glass, 2004). Among agreement metrics, pairwise Cohen's Kappa coefficient (k) (Cohen, 1960) is a *de facto* standard for manual annotation evaluation. However, it presents some weaknesses, particularly when dealing with skewed distributions of the phenomena within the annotated set (Di Eugenio & Glass, 2004; Byrt et al., 1993). Moreover, as traditionally employed, k only accounts for the match between the labels assigned by two annotators to the same item. This means that it does not account for 'implicit agreement', i.e. agreement given by the transitive property, specifically relevant to PR annotation. Hence, it is necessary to process the dataset in a way that allows applying k properly. To this aim, we assume that two annotators agree on the PR $A \prec C$ in both the following cases: (i) both annotators manually created the pair $A \prec C$; (ii) one annotator created the pair $A \prec C$ and the other created the pairs $A \prec B$ and $B \prec C$.[3] Then, the k metric can be computed as follows: given the terminology T of concepts used during annotation, consider as total items of the annotation task the list P of each pairs-wise combination p of concepts in T, including both $A \prec B$ and $B \prec A$ in P. For each annotator, consider as positive PR each p that is either manually created by the annotator or that can be derived for the transitive property. Consider p as non-PR otherwise. Then, compute k for each pair of annotators using equation 1.

$$k = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where:

- P_o probability that a concept pair is annotated as PR or non-PR by both annotators, i.e. the number of concept pairs annotated in the same way in both annotations over all possible concept pairs
- P_e probability of agreement occurring by chance, i.e. the probability that a pair is annotated and not annotated as PR.

For the whole group of annotators, use Fleiss' variant of Cohen k (Fleiss, 1971).

Depending on the obtained agreement and the project goals, more or less inclusive annotations combination methods can be chosen. At the two ends, taking the *Union* $\cup$ of PRs means including all the PRs identified by the annotators, while taking their *Intersection* $\cap$ means including only shared PRs (i.e., PRs detected by all the annotators). In

[2] This is a consolidated practice for evaluating the reliability of manually produced annotations. Refer to Artstein (2017) for an overview of inter-annotator agreement measures and their use.

[3] Note that, in this case, annotators are regarded as agreeing on the annotation of the pair $A \prec C$, but not on that of $A \prec B$ and $B \prec C$.

general, when the goal of the PR project is to analyze every case where annotators claim to encounter a relation, it is advisable to use more inclusive combination approaches such as the union. This is particularly relevant when the goal is to discover linguistic patterns in the textual realizations of PRs or when the annotators' judgments are highly reliable due to their strong domain expertise, assuming that annotation revisions have been carried out. On the other hand, this approach is not recommended with low-experienced annotators and when the annotations revision has not been performed. Less inclusive combination approaches offer higher certainty and guarantee higher consensus about the relations included in the gold dataset. However, they result in more limited datasets, particularly when there is low agreement among annotators.

It is worth noting that, when possible, a good practice consists in discussing among annotators about disagreement cases in order to converge toward an agreed PR graph, as suggested in Wang et al. (2021) for concept annotation.

The final phases of *Revision* and *Release* of the gold dataset will be detailed in Sect. 5 through the description of an annotation project and its meta-annotation using a standard vocabulary, following Linked Data principles.

4 Evaluation of the PREAP Protocol

In this section, we present the final evaluation (lower block of Fig. 2, *Evaluation with education experts*) that we carried out on different domains by comparing five datasets produced using PREAP against datasets obtained through alternative PR annotation methods.

To evaluate if PREAP succeeds in reaching the goals stated in the introduction, we formulated the following Research Questions (RQ).

RQ1: to what extent does PREAP succeed in obtaining PR-annotated datasets that are reliable? Specifically:

- RQ1.1: to what extent are PR relations *consistently annotated* by the annotators?
- RQ1.2: to what extent is the gold-PR dataset resulting from the combination of individual PR datasets *complete and accurate*?

RQ2: to what extent does PREAP succeed in obtaining textbook-driven annotations, i.e., PR-annotated datasets that *match the text* in terms of prerequisite concepts used by the textbook's author to make the reader understand the target concepts?

4.1 Methods

We conducted a mixed-method study based on quantitative and qualitative dimensions for data quality assessment (Zaveri et al., 2013), detailed below. These were used to compare the datasets produced using PREAP against datasets annotated by employing alternative approaches, referred to as *baseline methods*.

To answer RQ1.1 concerning *consistency*, i.e., the extent to which the dataset does not report conflicting annotations for similar phenomena (Mendes et al., 2012), we exploited agreement metrics between manually produced annotations as usual in such cases (Artstein & Poesio, 2008; Artstein, 2017; Hripcsak & Wilcox, 2002).

To answer RQ1.2 concerning *completeness* (the extent to which the annotations cover the relevant information of the data (Mendes et al., 2012; Zaveri et al., 2013)) and *accuracy* (the degree of correctness and precision with which the annotation represents information (Zaveri et al., 2013)) we performed an evaluation where education experts, i.e. teachers in the respective domains and a pedagogist, were asked to evaluate the annotated PR datasets represented as graphs and face-to-face interviewed to discuss the answers.

To answer RQ2 teachers were asked to assess the match between text and PR annotations in their respective domain, by evaluating the *adherence* between the annotation and the content of the source text, focusing on the way concepts are presented, and *relevancy*, i.e. the extent to which the annotated data are applicable and helpful for a task at hand (Zaveri et al., 2013), in our case learning support. The assessment was followed by a face-to-face interview.

Additionally, in order to obtain a comprehensive comparison of PREAP against the baseline methods, we computed the average *completion time* required to perform each annotation.

4.1.1 Baseline Annotation Methods

Four PR-annotation methods were used as baselines.

Manual Methods (MMs):

- MMP, a Manual Method for concept Pairs annotation of PRs (Li et al., 2019). In this method, annotators annotate if a PR exists between all possible pairwise combinations of pre-defined concepts using their background knowledge.
- MMT is an adaptation from MMP since we could not find Textbook-driven approaches in the literature. Instead of relying on their background knowledge, annotators are given a text to check if a PR exists between pairs of concepts therein.

Automated Methods (AMs):

– RefD (Liang et al., 2015), a widely adopted method for PR identification (cf Sect. 2), which exploits knowledge external to the text: basically, a PR is found between concepts that result associated from the analysis of links between their corresponding Wikipedia pages;
– Burst-based method (Adorni et al., 2019) annotates PRs based on the text content. Specifically, it uses Burst Analysis to identify portions of texts where each concept is estimated as relevant and then exploits temporal patterns between them to find concept pairs showing a PR.

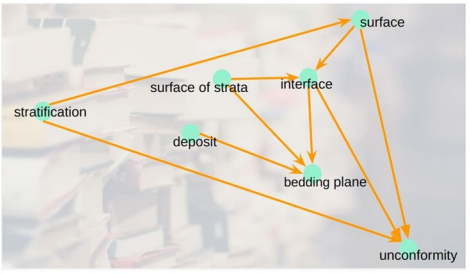
4.1.2 Source Texts and Participants

For the annotation task, we used five source texts from three domain areas: two texts in mathematics (algebra, statistics), two texts in natural science (biology, biochemistry) and

Table 1 Datasets statistics: domain of the source texts, text size in tokens and number of concepts

Source text domain	Tokens in text	Num. of concepts	PR PREAP	PR MMP	PR MMT	PR RefD	PR Burst
Algebra	253	23	22	287	161	68	28
Statistics	243	14	25	73	93	23	9
Biology	135	23	24	124	95	86	27
Biochemistry	286	16	29	108	97	37	49
Archaeology	208	18	25	115	114	0	17

The last five columns report the number of PRs obtained by annotating each source text according to PREAP and each baseline method



There are thus two main types of interface: those which are the surfaces of strata and those which are only surfaces, formed by the removal of existing stratification. In geology, these types are referred to as bedding planes and unconformities. The surfaces of strata are bedding planes, and 'mark successive positions of the surface, perhaps a sea floor or a lake bottom or a desert , on which material that now forms rocks was deposited' (Kirkaldy 1963: 21). Bedding planes are equal to the horizontal spread of a deposit and are contemporary with the cessation of its formation. Unconformities are surfaces which mark the levels at which existing stratification has been destroyed by erosion.

Harris, E. *Principles of archaeological stratigraphy*. p.54, Elsevier, 2014.

Fig. 4 Example of a PR graph from PRAT user interface and corresponding textbook portion (Archaeology domain)

one text in archaeology. Each text was acquired from a textbook targeting undergraduate students not majoring in the field of the book.[4]

We recruited six annotators, two for each domain area (post-graduate level expertise, age range between 25 and 49, AVG=29.8, SD=9.8). For the evaluation, we recruited 12 university teachers, grouped for domain area (age range between 32 and 65, AVG=45.4, SD=12.9), and one pedagogist (senior researcher in Education, age 47).

4.1.3 Study Setup and Procedures

1. Creation of the PR annotated datasets. To ensure a consistent experimental setting, human annotators and automatic methods were provided with the same set of concepts extracted from the source texts as in the semi-automatic autonomous option of PREAP.

– PR datasets creation through MMs: the annotators were asked to perform the annotation task using MMP, MMT and PREAP, following the respective annotation procedures, but varying their order to avoid biases. This resulted in 30 individual PR data-

[4] Jarboui A, et al. (2016) Fundamentals of Algebra, Magnum Publishing. Tabak J (2009) Probability and statistics: The science of uncertainty, W.H. Freeman & Co. Bartee L, et al. (2019) General Biology I, Open Oregon Educational Resources. Molnar C, et al. (2013) Concepts of Biology, OpenStax College. Harris E (2014) Principles of archaeological stratigraphy, Elsevier.

Table 2 The table reports, for each manual annotation method: the inter-annotator agreement obtained for each dataset (left side); the average time employed for performing the annotation and the standard deviation between annotators (right side)

	Inter-annotator agreement			Average annotation time in minutes		
Domain	PREAP	MMP	MMT	PREAP	MMP	MMT
Algebra	0.60	0.17	0.23	20 (± 2.83)	31.5 (± 9.20)	24 (± 1.41)
Statistics	0.61	0.31	0.01	15 (± 2.83)	18.5 (± 4.95)	20 (± 0)
Biology	0.71	0.60	0.27	13 (± 2.83)	30 (± 7.07)	23 (± 2.83)
Biochemistry	0.45	0.16	0.22	19 (± 1.41)	35.5 (± 3.53)	26.5 (± 2.12)
Archaeology	0.53	0.22	0.22	21.5 (± 0.71)	38.5 (± 2.12)	31 (± 1.41)

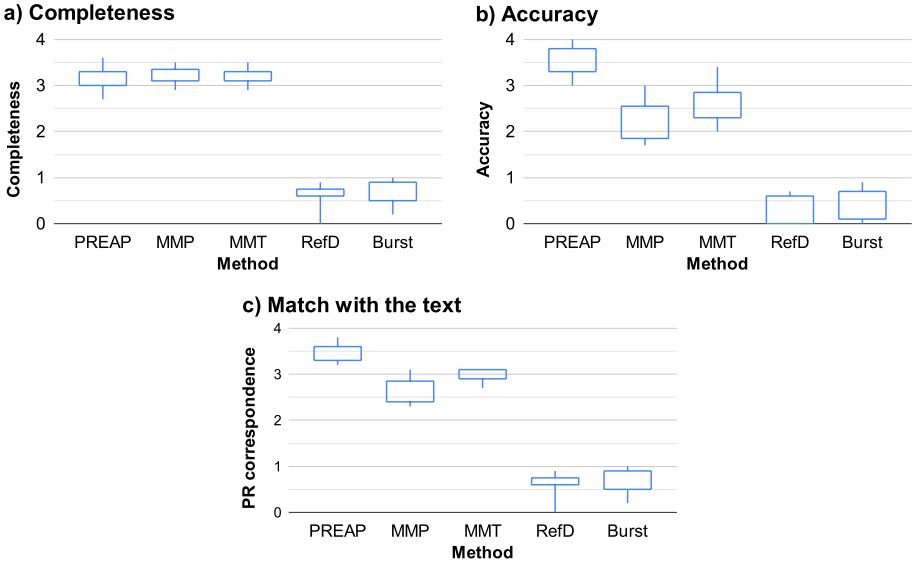


Fig. 5 Results for **a** completeness, **b** accuracy, and **c** match with the text (PR correspondence)

sets (annotators*methods*domains), then combined using the union option to obtain 15 gold datasets (i.e., one for each method for each domain).

– PR datasets creation through Automated Methods (AMs): we implemented the RefD and Burst-based methods as described in the cited references, then we generated the PR datasets using them. This resulted in 10 PR datasets (methods*domains: one dataset for each method for each domain).

Table 1 provides details about the resulting gold datasets. Figure 4 shows a portion of one of the datasets visualized as a *PR graph* on the PRAT tool.

2. Evaluation with education experts. We organized individual face-to-face meetings with the teachers and the pedagogist. After general instructions, teachers were provided with the set of concepts and the PR graphs obtained using MMs and AMs in their area of

expertise. They were given about 1 h, or more when required, to analyze and evaluate the graphs. Later, they were asked to read the source texts and evaluate the graphs according to the dimensions introduced in Sect. 4.1. Finally, we discussed the answers in an open-ended interview. The average time of each meeting with teachers was 130 min. In a final meeting, all the results were discussed with a pedagogist, commenting on the use of the PR graphs for educational purposes.

4.2 Results

Annotations consistency. To investigate the effect of PREAP on annotations' consistency (RQ1.1), we measured the *inter-annotator agreement* using the approach described in Sect. 3.2.4 between the individual PR datasets produced with PREAP and the manual baseline methods. Results show better performance of PREAP compared to the manual methods MMP (AVG +0.98) and MMT (AVG +2.05) (Table 2, left side).

Completeness and accuracy of the combined PR datasets. To investigate RQ1.2, we relied on the evaluations of teachers. Specifically, *completeness* is evaluated by detecting the number of PR pairs in common between the datasets produced by the annotation methods and the PRs identified by the teachers. To this aim, teachers were asked to identify the PRs for each concept as as if they had to explain them to a student, drawing a concept map of prerequisites. In this process, they were free to look at the graphs under evaluation and to modify their identified PRs in order to produce their optimal map as in a process of ground truth creation. Then, for each prerequisite in their map, teachers were asked to confirm its presence in the graph being evaluated. In detail: a 'good' score is given if a direct or indirect PR exists in the evaluated graph, while an 'acceptable' score is given if the two concepts are not linked but their PRs are consistent with the graph. No scores otherwise. Labels are then converted to numbers and combined. The result is normalized by the total number of PRs identified by the teacher and mapped to a five-point scale. Summary results are reported in Fig. 5a, and detailed data are reported in Appendix. We used the Kruskal–Wallis non-parametric test to check if any significant difference exists among the completeness scores of the six methods, finding that there is a significant difference among the groups ($X^2(4)= 53.98$, $p < .001$). Then we used the Post-Hoc Mann Whitney U test for pairwise comparison. We did not find any significant difference between the Manual Methods pairs, while we found that the difference between the MMs and the AMs is significant with $p < .001$ for each MM-AM pair. The Bonferroni correction ($\alpha = 0.005$) did not change the statistical significance of any of the outcomes above, since all of these have p values $< .001$.

Accuracy, as defined in Sect. 4, is measured by asking teachers to evaluate the correctness of a set of randomly extracted paths of three nodes form each graph. Evaluating paths instead of single pairs is coherent with the definition of PRs characterized by transitivity, and thus relevant to assess accuracy. In detail, if both the PRs in the path are correct then, a 'very good' score is given; if one of the PRs is correct and the other weakly wrong, but consistent with the graph, then a 'good' score is given; 'bad' score is given otherwise. As with the completeness score, labels are converted to numbers, combined, normalized, and mapped to a five-point scale. Results are reported in Fig. 5b, details on each evaluation test are in Appendix.

By performing the same statistical analysis, we found a significant difference among the six methods, according to the Kruskal-Wallis test ($X^2(4)= 64.43$, $p < .001$). The

Mann Whitney test showed a significant difference in the mean ranks of each Manual Method compared to each Automated Method. Moreover, it revealed a significant difference in PREAP accuracy compared against both MMP and MMT ($p < .001$ in both cases) and a difference between MMP and MMT ($P = 0.0140$), while there was no significant difference between the two AMs. Bonferroni correction ($\alpha = 0.005$) did not modify the results except for the difference between MMP and MMT that became not significant.

Match with text. To investigate the effect of PREAP on the correspondence of annotations to the prerequisites expressed in the text (RQ2), teachers were asked to read the source texts of their domain area and write down, for each concept, the prerequisite concepts used to explain it. Then, they were asked to repeat the procedure for completeness assessment, but checking only the first condition, i.e. the existence of the same PRs in the evaluated graphs. Results are reported in Fig. 5c and in Appendix. We performed the same statistical analysis as above, finding statistical differences among the groups using the Kruskal–Wallis test ($X^2(4)= 65.63$, $p < .001$). As well as for accuracy, the pairwise comparison of the methods in terms of 'match with the text' showed a significant difference in the mean ranks of each Manual Method compared to each Automated Method. Moreover, it revealed a significant difference in PREAP accuracy compared against both MMP and MMT ($p < .001$ in both cases), and also a significant difference between MMP and MMT but with a higher p-value ($p = 0.0027$). No significant difference was found between the two AM methods. Bonferroni correction ($\alpha = 0.005$) did not modify the results.

Concerning the relevancy dimension, based on the interviews conducted, it can be concluded that most of the teachers (9) consider PR graphs tied to the text as a potentially very useful feature for educational purposes. Some teachers (3) argue that usefulness depends on several factors. Additionally, 77% claim that it can support learners, 85% believes that it can be useful for teachers to organize the contents of lectures.

Task completion time. We computed the average time used to annotate each PR dataset generated using PREAP, MMP and MMT. Results show that the average time is lower for PREAP than for the other methods, indicating 43% and 29% less time for PREAP than for MMP and MMT, respectively (Table 2, right side). We used one-way analysis of variance (ANOVA) to check if the difference between the averages of the three groups was significant. Results revealed statistically significant differences among the three groups, each of equal size (annotators*domains) for each method ($F(2,27)=12.57$, $n^2=0.48$, $p < .001$). After the one-way ANOVA Test, Tukey Test was used as a complementary Post-Hoc analysis for pairwise comparisons. The difference resulted to be significant ($p < 0.05$) for the pairs PREAP-MMP ($p < .001$) and PREAP-MMT ($p = 0.027$), while it was not found to be significant for the MMP-MMT pair ($p = 0.08003$).

4.3 Discussion

The evaluation provided a rich source of quantitative and qualitative data. Limiting the analysis to what concerns the research questions of this study, we highlight the following results.

RQ1 was aimed at evaluating the reliability of the datasets annotated using PREAP (Goal1) by considering: the consistency of annotations in terms of inter-annotator agreement (RQ1.1) and the completeness and accuracy of the resulting datasets (RQ1.2). As for RQ1.1, results in the previous section show that the *inter-annotator agreement scores* obtained on the PREAP datasets are much higher than those obtained relying on the other

manual methods. Apart from one case (biology annotated using MMP), the agreement scores obtained using MMP and MMT are generally slight, while they raise to moderate and substantial with PREAP (Landis and Koch, 1977). Considering RQ1.2, we can observe that all the manual MM methods perform considerably better than the automatic AM methods, both for completeness and accuracy. Focusing on PREAP against MMs, it appears that the three methods are mostly comparable in terms of *completeness* of the datasets, while in terms of *accuracy* PREAP turns out to yield better results. The main reason for the lower accuracy of MMP and MMT is the incoherence of some resulting PR paths. This can be attributed to the requirement of these methods of annotating set of concept pairs, identifying if a PR relation exists or not among the two. This seems to induce annotators to find more relations than necessary. For instance, teachers evaluated as wrong or borderline acceptable, but not good, PR relations between 'product' and 'enzyme', 'product' and 'activation energy', that were included in the biochemistry dataset annotated using the MMP method.

RQ2 was aimed to evaluate the textbook-driven annotation approach (Goal2). The results reported in Fig. 5c. The statistical analysis shows that PREAP-annotated datasets perform better in terms of *correspondence* between the annotation and the content of the source text. The interview clarified also the errors attributed to each method. In the case of PREAP, the main error reported was false prerequisite concepts mentioned in the explanation of another concept, whereas they were rather supplementary explanations or primary notions. For example, three teachers noted that the sentence 'Elementary algebra differs from arithmetic in the use of abstractions, such as using letters to stand for numbers' means that letters and numbers are prerequisites for abstraction but elementary algebra is not a prerequisite for them, as it resulted in the PREAP-annotated dataset. Another example is the sentence 'A horizontal layer interface will be recorded on a plan which shows the boundary contours of the deposit and, therefore, the limits of the interface', which made teachers raise concerns about the correctness of boundary contour as the prerequisite of deposit. *Relevancy* of the annotated datasets was discussed with the teachers and the pedagogist. The aim was to get hints about the value of such PR graphs for educational uses. As seen, almost all considered them a useful support for teachers and most of them for learners. The pedagogist pointed out concerns about its practical direct use with large graphs, suggesting splitting into sub-graphs. It was also observed that graph accuracy is essential for its usefulness (*relevancy*), and that PREAP is the method that most accurately helps to highlight the lesson structure underlying the educational text (*exact correspondence*), also thanks to its higher readability.

Finally, if we look at such results in light of the average *completion time* required for completing the annotations, we observe that not only PREAP improves annotation consistency, accuracy, and match with the text, but it is also faster than the baseline methods. No specific and recurrent differences have been found across domains for any of the metrics.

5 Annotation Project Case Study

This section presents the annotation project we carried out in the last cycle of development, following PREAP procedures described in Sect. 3.2, Fig. 3.

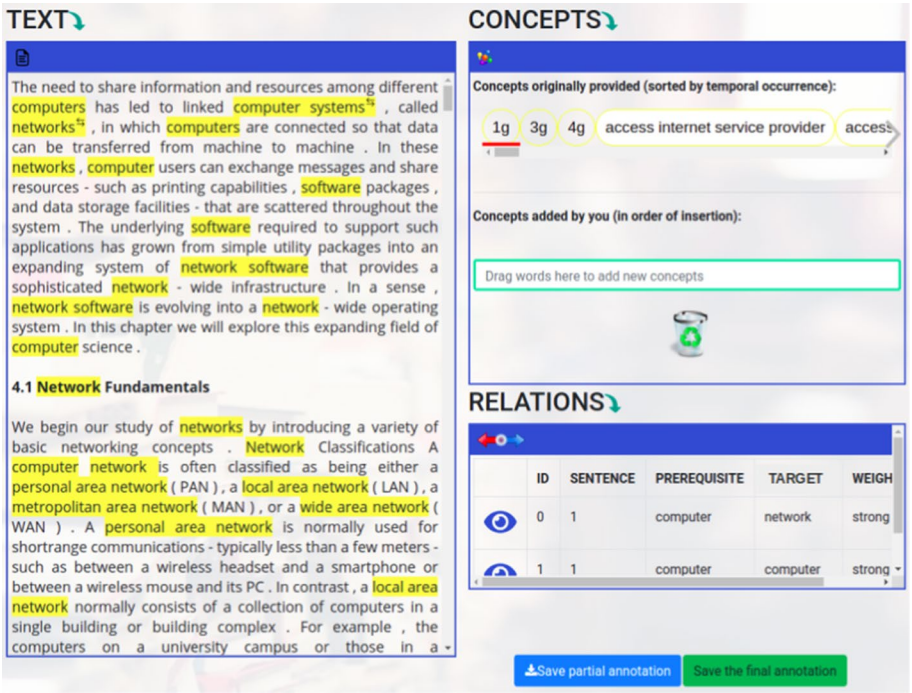


Fig. 6 PRAT tool annotation interface

5.1 Project Management

(i) Annotation goals and features: obtaining a gold dataset to be used for linguistic analysis of PR instances and for testing an automatic PR learning system based on linguistic features. While the latter use is presented in Sect. 6 of this paper, the linguistic analysis is left out for space limits.

(ii) Corpus selection and annotation tool: the annotation project relies on the fourth chapter of the computer science textbook (Brookshear & Brylow, 2015), 'Networking and the Internet' (20,964 tokens distributed along 780 sentences). The chosen tool for supporting PR annotation is the PRAT tool that we developed for PR annotation and analysis.

(iii) Annotators recruitment and training: the project manager recruited four master's students in Computer Science. Although they were domain experts with regard to the book content, none of them was familiar with annotation procedures or the annotation protocol. Hence, a preliminary training phase was conducted before starting the annotation task. The guidelines in the *annotation manual* were first explained by the project manager and then tried individually by each annotator in a trial annotation task. Then, annotators compared and commented on their individual annotations in a group discussion to address doubts.

(iv) Coordination of the other tasks: described below.

Table 3 For each Annotator [A1-A4] the Annotation block reports the annotated PRs and the distribution of Strong and Weak weights; the Revision block reports the number of PRs checked in the revision phase by each annotator (and the percentage out of their total PRs), the number of deleted and modified PRs

	Annotation			Revision		
	PRs	Strong (%)	Weak (%)	Checked	Deleted	Modified
A1	141	96.45	3.54	39 (27.66%)	11	4
A2	257	84.82	15.17	85 (33.07%)	21	25
A3	199	89.45	10.55	50 (25.12%)	15	10
A4	163	90.18	9.82	46 (28.22%)	20	9

5.2 Annotation Task

After the training phase, each annotator performed text annotation individually without consulting the other annotators.

(i) Concept annotation. Concept annotation, supervised by the project manager, was performed as an *autonomous* step with respect to PR annotation, adopting a semi-automatic approach. Specifically, the text underwent linguistic analysis[5] and semi-automatic terminology extraction through the Text-To-Knowledge[2] platform (Dell'Orletta et al., 2014). The platform returned a list of 185 candidate terms, then manually revised according to PREAP guidelines in order to remove non-concepts (e.g., busy channel, own network, term gateway, same machine) and add missing ones (e.g., router). The ultimate result was a terminology T of 140 concepts. The lists of automatically extracted and revised concepts are available among the shared resources. Note that Sect. 6 discusses different concept annotation options.

(ii) Prerequisite annotation. The PR annotation was carried out on PRAT tool. As shown in Fig. 6, the "Text" area displays the text and highlights the concepts of T (also listed in the upper part of the "Concepts" area). To create a PR pair, the expert selected the occurrence of the target concept in the text and entered its prerequisite concept, along with the weight of the relation (weak or strong) as specified in the annotation manual. The newly created PR is shown in the "Relations" area as a tuple encoding the following information: the pair ID, i.e. the id of the sentence where the target concept occurs and where the relation was entered, the prerequisite and target concepts, and the relation weight. The statistics about the annotations of each annotator [A1-A4] are reported in the 'Annotation' block of Table 3. As can be noted, although each expert produced different amounts of pairs, the distribution of weight labels is consistent. This is encouraging with regard to the effectiveness of KEQ in making annotators understand how weights should be assigned. Future analyses will investigate this in more depth.

5.3 Annotation Revision

After completing the annotation, experts performed the in-context revision of the PR annotations, checking the correctness of their own created pairs. As recommended by PREAP,

[5] Performed at the morpho-syntactic level by UDPipe pipeline (Straka et al., 2016).

Table 4 Gold-PR datasets created in three annotation projects where different options for concept annotation were employed

PREAP option	Dataset	Concepts	PRs	Agreement
Autonomous automatic	dataset v1	185	2,252	0.40
Autonomous semi-automatic	dataset v2	132	1,974	0.62
Simultaneous manual	dataset v3	353	6,768	0.25

each expert checked only the subset of PRs identified solely by her/himself and decided on confirming, deleting or modifying the weight of the pair.

Table 3, 'Revision' block, summarizes the statistics of the revision task. With respect to the overall number of PR pairs ('PRs' column), the revision involved a comparable number of pairs among annotators (between 25% and 33%). Considering the modified and deleted pairs, we obtain the following distributions: 38,46%, 54,12%, 50,00%, 63,04% for A1–A4 respectively. This means that an average of more than 50% of the checked PRs have been corrected in the revision phase, which shows the importance of this process in order to have reliable datasets.

5.4 Agreement and gold dataset

(i) Agreement evaluation. Annotations' consistency was computed pre- and post-revision using the inter-annotator agreement metrics adapted for PR introduced in Sect. 3.2.4. We computed both pairwise Cohen's (Cohen, 1960) and Fleiss' (Fleiss, 1971) k for all annotators. According to the common interpretation of k (Landis and Koch, 1977), we observe an average *moderate* agreement (0.60) among the original annotations when considering pairwise agreement (Cohen's k), which improves to 0.62 on the revised annotations. In fact, a small but consistent improvement is reported for all pairs of experts, confirming that the revision allowed obtaining more coherent and consistent annotations. Fleiss' k value rises from 0.43 to 0.45 when considering revised annotations. Confirming the results of the protocol evaluation (Sect. 4.2), PREAP seems to mitigate the disagreement attested when adopting different PR annotation strategies: Chaplot et al. (2016) and Fabbri et al. (2018), e.g., report an average pairwise agreement of around 0.30.

(ii) Annotations combination. The gold dataset was built by merging the four revised annotations (*Union* option): the 385 PR pairs annotated as PR by at least one expert appear in the gold dataset as *positive PRs*, i.e. showing a prerequisite relation[6]. The Union option aligns well with the project goal of creating a gold-PR dataset suitable for linguistic analysis of PR relations and for training a PR learning system using linguistic features. The conditions for the applicability of the Union option are also satisfied (ref. Sect. 3.2.4). These include the expertise level of the annotators, which ensures the understanding of the textbook content. Additionally, the average agreement among annotators provides assurance regarding the comprehension of the annotation guidelines. Moreover, the process of annotation revision resulted in not only a slight improvement in agreement (thus consistency) but above all augmented in correctness and, subsequently, reliability.

(iii) Gold dataset revision. To address potential inconsistencies and loops that may arise from the combination of annotations, we relied on the visualisation aids included in the

[6] Given that the annotation of PR weights remains a proposal, we did not take into account relation weights in this project.

Table 5 Performance of the PR learning model trained with each of the three gold datasets

Dataset	Precision (%)	Recall (%)	F1	Accuracy (%)
v1	83,75	87.85	85.71	85.34
v2	**87.87**	**89.67**	**88.73**	**88.60**
v3	65.08	87.94	74.79	70.36

The highest scores are marked using bold

PRAT tool. These allow to navigate the PR graph resulting from the combination of annotations and identify issues such as loops and lengthy paths, that stemmed from the annotations combination. Such issues were addressed through discussion among annotators, led by the annotation manager, similarly to Wang et al. (2021).

(iv) Gold dataset release. The gold-PR dataset is made available with the related documentation. It was also annotated with metadata according to schema.org vocabulary Dataset class (schema.org/Dataset) based on the W3C Data Catalog Vocabulary, encoded in JSON-LD format.

6 PREAP Options and Machine Learning Tests

This section discusses the use of PREAP options for concept annotation proposing three application scenarios that complete the case study presented in Sect. 5. In that case, concept annotation was performed using a *semi-automatic* approach as an *autonomous* step of the annotation. Here we present the results of two further annotation projects that differ only in the way concepts are annotated according to the other PREAP options: *autonomous automatic* and *simultaneous manual* annotation. The *autonomous manual* option can be assimilated to the case of autonomous semi-automatic annotation since candidate concepts were manually revised, as described. The example is intended to provide suggestions about the use of PREAP options for different purposes.

First, we describe the annotation projects using the three options, the resulting gold datasets, and the effect on the inter-annotator agreement. Then, we present the use of the datasets to train a machine learning algorithm for PR learning and discuss the effects on algorithm performance.

6.1 Annotation Projects Employing Different Options for Concept Annotation

Table 4 provides information on the three projects, including the details of the resulting datasets. The term 'Autonomous semi-automatic' corresponds to the case study discussed in Sect. 5. All projects rely on the same corpus and combination method for gold dataset creation described in Sect. 5. T2K^2 was employed for concept extraction in both projects relying on the autonomous option. Each project involved four different annotators, each with comparable levels of expertise.

As reported in Table 4, the number of concepts in each dataset version reflects the option of the protocol employed: v1 includes only the automatically extracted terms, in v2 the project manager post-processed the automatically extracted terms, as explained in Sect. 5.2, mostly removing non–concepts. Dataset v3 included also concepts manually added by annotators (agreement on concept annotations=0.71). This explains its larger size compared to

v1 and v2, and also the huge increase in PR relations identified by annotators. The inter-annotator agreement shows lower average Cohen agreement on v3 (k=0.25) compared to v2 (k=0.62), and also to v1 (k=0.40). This suggests that, while adding new concepts during annotation produced a richer set of concepts, it also created a less coherent dataset.

6.2 Training a Machine Learning Model and Performance Comparison

To show the use of the PR datasets to train a ML model for PR learning and to investigate the effect of the three options on the performance of the algorithm, we employed the deep learning classification model and the experimental setting of (Alzetta et al., 2019). This model acquires lexical (i.e. word embeddings) and global features (i.e. number of occurrences and measures of lexical similarity) for each pair of PRs from the raw textual corpus without relying on external knowledge bases, as in (Liang et al., 2015; Gasparetti et al., 2018; Talukdar and Cohen, 2012), which reflects PREAP annotation principles. The performance of the classifier trained with the three datasets is evaluated using precision, recall, F1, and accuracy computed in a 5-fold cross-validation setting, and compared against a Zero Rule baseline (accuracy=50%, F1=66.66%). As reported in Table 5, the results obtained by the three gold datasets exceed the baseline. The best performance is observed when the model is trained with v2 dataset and the worst with v3.

6.3 Discussion and Annotation Suggestions

Space constraints do not allow us to report the result of the analysis and to discuss them in detail. We just note that the results in terms of agreement and automatic extraction suggest a positive effect of annotating concepts as an autonomous step, as in v1 and v2. This is coherent with our recommendation to avoid simultaneous annotation of concepts and prerequisites (as in v3) unless specific requirements are given, e.g., in terms of dataset richness.

If we now focus the analysis on comparing v1 and v2 datasets, we observe that v2 results in higher agreement and better PR extraction performance. However, the recommendation for semi-automatic vs automatic concept annotation is not straightforward, and annotation managers should consider at least two factors. The first one is the time required for post-processing the automatically extracted concepts. Even though we found that it is, on average, lower than performing manual annotation, post-processing takes time to read the text and revise the list, as explained in the case study. The percentage increase in performance of v2 compared to v1 is too low to warrant the effort. However, the choice depends on the annotation project goals and the expected quality of annotated data. In particular, if the project aims to produce a dataset for ML training, using an automatic approach for concept extraction can be reasonable. In such cases, the subsequent manual PR annotation step can help mitigate the errors in automatic extraction since annotators are expected not to add PR relations between terms that do not represent domain concepts. This likely accounts for the slight performance decrease observed in v1. Conversely, if the knowledge graph has to be used per se, e.g. for intelligent textbook applications or as ground truth for evaluation tasks, higher correctness and coherence should guide the choice. The second factor to take into account in the decision is that PR extraction results are much affected by the algorithms employed for concept and PR learning, thus better performance might be achieved by other models than those used in this case study.

7 Conclusion and Limits

In this paper, we presented the PREAP protocol for textbook annotation, a systematic procedure for the creation of gold datasets annotated with prerequisite relations.

As a first goal, the protocol is intended to cover a gap in the current literature on prerequisite annotation which lacks systematic procedures for the manual annotation of prerequisites. The aim is to produce reliable datasets built using reproducible methods, adequate for reuse and comparison in learning tasks. The mixed quantitative-qualitative evaluation of the protocol against baseline methods for manual annotation in five domains shows that PREAP succeeds in obtaining datasets that present higher consistency of annotations and accuracy. While dataset completeness is generally comparable across methods, the annotation process using PREAP significantly reduces the required time compared to the other methods. Additionally, a comparison between PREAP and automated methods for PR annotation reveals that automated approaches are not yet able to match the annotation quality achieved through manual methods.

The second goal of PREAP was to design an annotation method aimed at capturing the prerequisite relations as expressed in the text: we refer to it as textbook-driven annotation approach, a method that is very common in concept annotation but still not widely addressed for prerequisite annotation. The annotation approach defined by PREAP also proposes to weight PRs differently based on the concepts' description in the textbook. This use of PR weights is still a proposal and will be further investigated in future studies. However, we did discuss this PR feature with annotators during the protocol design phases and they expressed a preference for being able to indicate the degree of importance of a prerequisite for a specific target concept. For the evaluation, we used the metrics of annotation correspondence with the source text content and relevancy, defined as the extent to which the annotated data are applicable and helpful for learning support. Also for this evaluation, we compared PREAP against manual and automated methods. Results confirm the validity of PREAP for the two metrics and highlight the expected value of such datasets for applications in education, including learning support for students, support to teachers for instructional design and for textbook comparison.

The paper reports also an annotation project case study that provides a detailed example of protocol application and discusses some of its options and uses for prerequisite extraction in a ML task. The datasets and all the text sources are publicly available with documentation and semantic meta-annotation based on the W3C Web Annotation Vocabulary (see Sect. 8).

The protocol has been applied to several texts belonging to different domains. Although we did not find specific and recurrent differences across domains, we cannot claim that the protocol fits all domains and needs, and further evaluations are necessary in this respect. However, we believe that PREAP contributes to the literature by introducing a method that addresses the aforementioned gaps and achieves the goals it was designed for, recalled above. We hope the results presented in this contribution can represent the starting point for the creation of novel resources for analysing PRs in new domains and scenarios, given the relevance of prerequisite relations for enhanced educational systems.

As a limit of the approach and a direction for future work, we observe that having annotations produced based on the content of multiple textbooks would be highly useful for comparing the content reported in different educational resources dealing with the same topic. Also, producing a unique dataset starting from multiple resources could be useful for educational

purposes and for ITSs (for improving the dataset coverage, for instance). However, this would require careful and accurate combination strategies to avoid inconsistencies and conflicts in the annotation. We are currently experimenting whether PR weights could be effectively exploited in this scenario, but this research goes beyond the goals of this manuscript. Moreover, future research could investigate other approaches for automatic concept extraction and reconciling annotators' revisions of concepts extracted from corpora through automatic methods. The high inter-annotator agreement reported in Sect. 6 and in previous analyses (Alzetta et al., 2020a) suggests a shared understanding of PREAP guidelines about the notion of concept used in the protocol. However, further experiments could be carried out in order to confirm this result, since the proper identification of concepts is a requirement for the reliability of PR annotation. Tests could also be conducted to investigate the balance between reducing the costs associated with concept annotation by involving non-experts and maintaining annotation quality. In this regard, the work of Lee et al. (2022) can provide inspiration for future research towards this direction.

8 Datasets and Resources

The materials and data presented in this paper are publicly available and have been archived in a public online repository, which can be accessed via the following link: https://github.com/IntAIEdu/PRAT/. Below is a list of the available datasets, documents and software that can be found in the repository. Researchers and interested parties are encouraged to visit the repository to access and utilize these materials for further exploration and analysis.[7]

1. PREAP annotation protocol:
   - PREAP Annotation Manual for Annotators
   - PREAP Specifications for Project Management
2. Datasets used for PREAP evaluation and case study
   - Evaluation: Datasets and Source Texts used in PREAP Evaluation with education experts
   - Case Study: Annotation project example
     - List of concepts, annotated PR-dataset, row text
     - JSON-LD and visual RDF graph encoding metadata information about the dataset and the related annotation process.
   - Case Study: Datasets and related data used in the ML Experimental tests
3. PRAT tool for PR annotation and analysis

Appendix

See Table 6.

[7] https://github.com/IntAIEdu/PRAT/.

Table 6 Results of the evaluation for *Completeness*, *Accuracy*, *Match with text* metrics

n evaluation	Teacher	Domain	PREAP	MMP	MMT	RefD	Burst
Completeness							
n1	t1	Algebra	3.0	3.5	3.5	0.7	0.6
n2	t2	Algebra	3.1	3.4	3.4	0.6	0.5
n3	t3	Algebra	3.5	3.1	3.3	0.7	0.7
n4	t3	Statistics	3.3	3.3	3.3	0.8	0.2
n5	t4	Statistics	3.2	3.4	3.3	0.7	0.2
n6	t5	Statistics	3.6	3.5	3.5	0.7	0.3
n7	t6	Biology	3.2	3.3	3.3	0.8	0.5
n8	t7	Biology	3.3	3.3	3.3	0.6	0.6
n9	t8	Biology	3.3	3.2	3.2	0.6	0.5
n10	t6	Biochemistry	2.7	3.0	3.1	0.9	1.0
n11	t8	Biochemistry	3.1	3.1	3.1	0.8	0.9
n12	t9	Biochemistry	2.8	2.9	3.0	0.6	0.7
n13	t10	Archaeology	3.0	2.9	2.9	0.0	0.9
n14	t11	Archaeology	3.0	3.1	3.0	0.0	1.0
n15	t12	Archaeology	3.2	3.1	3.1	0.0	0.9
		AVG	3.2	3.2	3.2	0.6	0.6
		SD	0.2	0.2	0.2	0.3	0.3
Accuracy							
n1	t1	Algebra	3.8	2.2	2.8	0.7	0.5
n2	t2	Algebra	4.0	2.3	3.1	0.6	0.8
n3	t3	Algebra	3.3	1.9	2.9	0.5	0.5
n4	t3	Statistics	3.1	2.5	2.7	0.2	0.0
n5	t4	Statistics	3.3	2.7	3.0	0.3	0.0
n6	t5	Statistics	3.3	1.7	2.7	0.2	0.0
n7	t6	Biology	3.9	1.9	2.5	0.0	0.3
n8	t7	Biology	3.8	2.6	2.5	0.2	0.2
n9	t8	Biology	4.0	3.0	2.4	0.0	0.0
n10	t6	Biochemistry	3.4	1.9	2.0	0.6	0.9
n11	t8	Biochemistry	3.3	1.7	2.2	0.7	0.9
n12	t9	Biochemistry	3.4	2.8	3.4	0.7	0.7
n13	t10	Archaeology	3.2	1.9	2.2	0.0	0.5
n14	t11	Archaeology	3.0	1.8	2.1	0.0	0.2
n15	t12	Archaeology	3.5	1.8	2.5	0.0	0.7
		AVG	3.5	2.2	2.6	0.3	0.4
		SD	0.3	0.4	0.4	0.3	0.3
n evaluation	Teacher	Domain	PREAP	MMP	MMT	RefD	Burst
Match with text							
n1	t1	Algebra	3.2	2.4	2.7	0.7	0.6
n2	t2	Algebra	3.3	2.3	2.7	0.6	0.5
n3	t3	Algebra	3.7	2.7	2.9	0.7	0.7
n4	t3	Statistics	3.6	2.4	2.9	0.8	0.2
n5	t4	Statistics	3.3	2.4	3.0	0.7	0.2
n6	t5	Statistics	3.6	2.9	3.1	0.7	0.3

Table 6 (continued)

n evaluation	Teacher	Domain	PREAP	MMP	MMT	RefD	Burst
n7	t6	Biology	3.3	3.1	3.1	0.8	0.5
n8	t7	Biology	3.5	2.8	2.8	0.6	0.6
n9	t8	Biology	3.8	3.0	3.1	0.6	0.5
n10	t6	Biochemistry	3.4	2.7	3.1	0.9	1.0
n11	t8	Biochemistry	3.6	3.1	3.1	0.8	0.9
n12	t9	Biochemistry	3.3	2.4	3.0	0.6	0.7
n13	t10	Archaeology	3.6	2.5	2.9	0.0	0.9
n14	t11	Archaeology	3.3	2.4	3.0	0.0	1.0
n15	t12	Archaeology	3.4	2.7	3.1	0.0	0.9
		AVG	3.5	2.7	3.0	0.6	0.6
		SD	0.2	0.3	0.1	0.3	0.3

Acknowledgements We thank the colleagues from the Institute for Computational Linguistics ``A. Zampolli'' (CNR-ILC, IT), from the PhD Committee of Digital Humanities (University of Genoa), and the PAWS Lab (School of Computing and Information, Pittsburgh University), who respectively took part in the definition and evaluation phases, and provided fruitful insights and discussion.

Funding Open access funding provided by Università degli Studi di Genova within the CRUI-CARE Agreement. No funding was received for conducting this study.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Adorni, G., & Koceva, F. (2016). Educational concept maps for personalized learning path generation. In *AI* IA 2016 Advances in Artificial Intelligence: XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, November 29–December 1, 2016, Proceedings XV* (pp. 135–148). Springer International Publishing.

Adorni, G., Alzetta, C., Koceva, F., Passalacqua, S., & Torre, I. (2019). Towards the identification of propaedeutic relations in textbooks. *International conference on Artificial Intelligence in Education*: Springer.

Alzetta, C., Koceva, F., Passalacqua, S., Torre, I., & Adorni, G. (2018). PRET: Prerequisite-Enriched Terminology. A Case Study on Educational Texts. In *Proceedings of the Italian Conference on Computational Linguistics (CLiC-it 2018).*

Alzetta, C., Miaschi, A., Adorni, G., Dell'Orletta, F., Koceva, F., Passalacqua, S., & Torre, I. (2019). Prerequisite or not prerequisite? That's the problem! an NLP-based approach for concept prerequisites learning. In: 6th Italian Conference on Computational Linguistics, CLiC-it 2019, CEUR-WS, vol 2481.

Alzetta, C., Galluccio, I., Koceva, F., Passalacqua, S., & Torre, I. (2020a). Digging Into Prerequisite Annotation. In *iTextbooks@ AIED* (pp. 29–34).

Alzetta, C., Miaschi, A., Dell'Orletta, F., Koceva, F., & Torre, I. (2020b). PRELEARN@EVALITA 2020: Overview of the prerequisite relation learning task for Italian. In *Proceedings of 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Angel, J., Aroyehun, S.T., Gelbukh, A. (2020). NLP-CIC@ PRELEARN: Mastering prerequisites relations, from handcrafted features to embeddings. In: Proceedings of 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020).

Artstein, R. (2017). Inter-annotator agreement. *Handbook of linguistic annotation* (pp. 297–313). Springer.

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

Atkinson, J. (2020). *Technology, change and the academic library: Case studies. Trends and reflections*. Chandos Publishing.

Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A. (2017). Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), (pp. 546–555).

Ausubel, D. P., Novak, J. D., Hanesian, H., et al. (1968). *Educational psychology: A cognitive view*. Rinehart and Winston.

Bai, Y., Zhang, Y., Xiao, K., Lou, Y., & Sun, K. (2021). A BERT-based approach for extracting prerequisite relations among Wikipedia concepts. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2021/3510402

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics, 37*(4), 699–725. https://doi.org/10.1162/COLI_a_00074

Brookshear, G., & Brylow, D. (2015). *Computer science: An overview* (Global Edition). Pearson Education Limited.

Brusilovsky, P., & Vassileva, J. (2003). Course sequencing techniques for large-scale web-based education. *International Journal of Continuing Engineering Education and Life Long Learning, 13*(1–2), 75–94.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*(5), 423–429.

Cabré, M. T. (1999). *Terminology: Theory, methods, and applications* (Vol. 1). John Benjamins Publishing.

Carey, S. (1999). Knowledge acquisition: Enrichment or conceptual change. *Concepts: Core readings* (pp. 459–487). MIT Press.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Carvalho, P.F., Gao, M., Motz, B.A., Koedinger, K.R. (2018). Analyzing the relative learning benefits of completing required activities and optional readings in online courses. International Educational Data Mining Society.

Changuel, S., Labroche, N., & Bouchon-Meunier, B. (2015). Resources sequencing using automatic prerequisite-outcome annotation. *ACM Transactions on Intelligent Systems and Technology (TIST), 6*(1), 1–30.

Chaplot, D.S., Yang, Y., Carbonell, J.G., Koedinger, K.R. (2016). Data-driven automated induction of prerequisite structure graphs. In: Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016, (pp. 318–323).

Chau, H., Balaneshin, S., Liu, K., Linda, O. (2020a). Understanding the tradeoff between cost and quality of expert annotations for keyphrase extraction. In: Proceedings of of the 14th Linguistic Annotation Workshop, (pp. 74–86).

Chau, H., Labutov, I., Thaker, K., He, D., & Brusilovsky, P. (2020). Automatic concept extraction for domain and student modeling in adaptive textbooks. *International Journal of Artificial Intelligence in Education, 31*, 820–846.

Cimiano, P., Völker, J. (2005). Text2Onto. Natural language processing and information systems. In: International Conference on Applications of Natural Language to Information Systems (NLDB), (pp 15–17).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Council, N. R., et al. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press.

Dell'Orletta, F., Venturi, G., Cimino, A., Montemagni, S. (2014). T2K^2: A system for automatically extracting and organizing knowledge from texts. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014).

Di Eugenio, B., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics, 30*(1), 95–101.

Dickinson, M. (2015). Detection of annotation errors in corpora. *Language and Linguistics Compass, 9*(3), 119–138.

Eighmy-Brown, M., McCready, K., & Riha, E. (2017). Textbook access and affordability through academic library services: A department develops strategies to meet the needs of students. *Journal of Access Services, 14*(3), 93–113.

Eskin, E. (2000). Detecting errors within a corpus using anomaly detection. In: 1st Meeting of the North American Chapter of the Association for Computational Linguistics.

Fabbri, A.R., Li, I., Trairatvorakul, P., He Y., Ting, W., Tung, R., Westerfield, C., Radev, D. (2018). Tutorialbank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, (pp. 611–620).

Faure, D., & Nedellec, C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium. International Confernce on Knowledge Engineering and Knowledge Management, (pp. 329–334) Springer.

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

Fort, K., Nazarenko, A., Claire, R. (2011). Corpus linguistics for the annotation manager. In: Corpus Linguistics.

Gagne, R. M. (1962). The acquisition of knowledge. *Psychological Review, 69*(4), 355.

Gasparetti, F. (2022). Discovering prerequisite relations from educational documents through word embeddings. *Future Generation Computer Systems, 127*, 31–41.

Gasparetti, F., De Medio, C., Limongelli, C., Sciarrone, F., & Temperini, M. (2018). Prerequisites between learning objects: Automatic extraction based on a machine learning approach. *Telematics and Informatics, 35*(3), 595–610.

Golik, W., Bossy, R., Ratkovic, Z., & Nédellec, C. (2013). Improving term extraction with linguistic analysis in the biomedical domain. *Research in Computing Science, 70*, 157–172.

Gordon, J., Zhu, L., Galstyan, A., Natarajan, P., Burns, G. (2016). Modeling concept dependencies in a scientific corpus. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol 1, (pp. 866–875).

Gordon, J., Aguilar, S., Sheng, E., Burns, G. (2017). Structured generation of technical reading lists. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, (pp. 261–270).

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199–220.

Hippisley, A., Cheng, D., & Ahmad, K. (2005). The head-modifier principle and multilingual term extraction. *Natural Language Engineering, 11*(2), 129–157.

Hovy, E., & Lavid, J. (2010). Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation, 22*(1), 13–36.

Hripcsak, G., & Wilcox, A. (2002). Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. *Journal of the American Medical Informatics Association, 9*(1), 1–15.

Hu, X., He, Y., Sun, G. (2021). Active learning for concept prerequisite learning in Wikipedia. In: 13th International Conference on Machine Learning and Computing, (pp. 582–587).

Huang, C., Li, Q., Chen, Y., Zhan, D. (2021). An effective method for constructing knowledge graph of online course. In: 4th International Conference on Big Data and Education, (pp. 12–18).

Hübscher, R. (2001) What's in a prerequisite. In: International Confernace on Advanced Learning Technology (ICALT), Citeseer.

Ide, N., & Pustejovsky, J. (2017). *Handbook of linguistic annotation*. Springer.

Labutov, I., Huang, Y., Brusilovsky P, He, D. (2017). Semi-supervised techniques for mining learning outcomes and prerequisites. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, (pp. 907–915).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lee, J. U., Klie, J. C., & Gurevych, I. (2022). Annotation curricula to implicitly train non-expert annotators. *Computational Linguistics, 48*(2), 343–373.

Li, B., et al. (2021). Prerequisite learning with pre-trained language and graph embedding. In: International Conference on NLP and Chinese Computing, (pp. 98–108)

Li, I., Fabbri, A.R., Tung, R.R., Radev, D.R. (2019). What should i learn first: Introducing lecturebank for NLP education and prerequisite chain learning. In: Proceedings of AAAI 2019.

Liang, C., Wu, Z., Huang, W., Giles, C.L. (2015). Measuring prerequisite relations among concepts. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, (pp. 1668–1674).

Liang, C., Ye, J., Wu, Z., Pursel, B., Giles, C.L. (2017). Recovering concept prerequisite relations from university course dependencies. In: AAAI, (pp 4786–4791).
Liang, C., Ye, J., Wang, S., Pursel, B., Giles, C.L. (2018). Investigating active learning for concept prerequisite learning. In: Proceedings of EAAI.
Liang, C., Ye, J., Zhao, H., Pursel, B., Giles, C.L. (2019). Active learning of strict partial orders: A case study on concept prerequisite relations. In: 12th International Conference on Educational Data Mining, EDM 2019, (pp. 348–353).
Limongelli, C., Gasparetti, F., Sciarrone, F. (2015). Wiki course builder: A system for retrieving and sequencing didactic materials from Wikipedia. In: 2015 International Conference on Information Technology Based Higher Education and Training (ITHET), IEEE, (pp. 1–6).
Lu, W., Zhou, Y., Yu, J., Jia, C. (2019). Concept extraction and prerequisite relation learning from educational data. In: Proceedings of the Conference on Artificial Intelligence, vol 33, (pp. 9678–9685).
Manrique, R., Sosa, J., Marino, O., Nunes, B. P., Cardozo, N. (2018). Investigating learning resources precedence relations via concept prerequisite learning. 2018 IEEE/WIC/ACM Int IEEE: Conference on Web Intelligence (pp. 198–205).
Martinez-Rodriguez, J. L., Hogan, A., & Lopez-Arevalo, I. (2020). Information extraction meets the semantic web: A survey. *Semantic Web, 11*(2), 255–335.
Mendes, P.N., Mühleisen, H., Bizer, C. (2012). Sieve: Linked data quality assessment and fusion. In: Proceedings of the 2012 joint EDBT/ICDT workshops, (pp. 116–123).
Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development, 50*(3), 43–59.
Miaschi, A., Alzetta, C., Cardillo, F.A., Dell'Orletta, F. (2019). Linguistically-driven strategy for concept prerequisites learning on Italian. In: Proceedings of 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2019).
Novak, J. D. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching, 27*(10), 937–949.
Novak, J. D., & Cañas, A. J. (2008). *The theory underlying concept maps and how to construct and use them*. Institute for Human and Machine Cognition.
Pan, L., Li, C., Li, J., Tang, J. (2017a). Prerequisite relation learning for concepts in MOOCS. In: Proceedings of the 55th Meeting of the Association for Computational Linguistics, (pp. 1447–1456).
Pan, L., Li, C., Li, J., Tang, J. (2017b). Prerequisite relation learning for concepts in MOOCS. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol 1, (pp. 1447–1456).
Pan, L., Wang, X., Li, C., Li, J., Tang, J. (2017c). Course concept extraction in moocs via embedding-based graph propagation. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, (pp. 875–884)
Plank, B., Hovy, D., Søgaard, A. (2014). Linguistically debatable or just plain wrong? In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, (pp. 507–511).
Pustejovsky, J. (2006). Unifying linguistic annotations: A time ML case study. In: Proceedings of Text, Speech, and Dialogue Conference.
Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning: A guide to corpus-building for applications*. O Reilly Media.
Rani, M., Dhar, A. K., & Vyas, O. (2017). Semi-automatic terminology ontology learning based on topic modeling. *Engineering Applications of Artificial Intelligence, 63*, 108–125.
Roy, S., Madhyastha, M., Lawrence, S., Rajan, V. (2019). Inferring concept prerequisite relations from online educational resources. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, (pp. 9589–9594).
Ruiz-Primo, M. A. (2000). On the use of concept maps as an assessment tool in science: What we have learned so far. *REDIE Revista Electrónica de Investigación Educativa, 2*(1), 29–53.
Sager, J. C. (1990). *Practical course in terminology processing*. John Benjamins Publishing.
Sayyadiharikandeh, M., Gordon, J., Ambite, J.L., Lerman, K. (2019), Finding prerequisite relations using the Wikipedia clickstream. In: Companion Proceedings of the WWW Conference, (pp. 1240–1247).
Shen, W., Wang, J., & Han, J. (2014). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering, 27*(2), 443–460.
Straka, M., Hajic, J., Straková, J. (2016), UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In: Proceedings of the International Conference on Language Resources and Evaluation, (pp. 4290–4297).
Suresu, S., Elamparithi, M. (2016). Probabilistic relational concept extraction in ontology learning. International Journal of Information Technology, 2(6)

Talukdar, P.P., Cohen, WW. (2012). Crowdsourced comprehension: predicting prerequisite structure in Wikipedia. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics, (pp 307–315).
Velardi, P., Faralli, S., & Navigli, R. (2013). Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics, 39*(3), 665–707.
Wang, M., Chau, H., Thaker, K., Brusilovsky, P., & He, D. (2021). Knowledge annotation for intelligent textbooks. *Technology, Knowledge and Learning, 28*, 1–22.
Wang, S., Liang, C., Wu, Z., Williams, K., Pursel, B., Brautigam, B., Saul, S., Williams, H., Bowen, K., Giles, C.L. (2015), Concept hierarchy extraction from textbooks. In: Proceedings of the 2015 ACM Symposium on Document Engineering, (pp. 147–156).
Wang, S., Ororbia, A., Wu, Z., Williams, K., Liang, C., Pursel, B., Giles, C.L. (2016), Using prerequisites to extract concept maps from textbooks. In: Proceedings of the 25th ACM International on Confernce on information and knowledge management, ACM, (pp. 317–326).
Wen, H., Zhu, X., Zhang, M., Zhang, C., & Yin, C. (2021). Combining Wikipedia to identify prerequisite relations of concepts in MOOCS. In: International Conference on Neural Information Processing, (pp. 739–747) Springer.
Xiao, K., Bai, Y., & Wang, S. (2021). Mining precedence relations among lecture videos in MOOCS via concept prerequisite learning. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2021/7655462
Xiao, K., Bai, Y., & Wang, Z. (2022). Extracting prerequisite relations among concepts from the course descriptions. *International Journal of Software Engineering and Knowledge Engineering, 32*(04), 503–523.
Yang, Y., Liu, H., Carbonell, J., Ma, W. (2015). Concept graph learning from educational data. In: Proceedings of the Eighth ACM International Confernce on Web Search and Data Mining, ACM, (pp. 159–168).
Yu, X., Stahr, M., Chen, H., Yan, R. (2021). Design and implementation of curriculum system based on knowledge graph. In: IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), IEEE, (pp 767–770).
Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., & Hitzler, P. (2013). Quality assessment methodologies for linked open data. *Submitted to Semantic Web Journal, 1*(1), 1–5.
Zhao, G., Zhang, X. (2018), Domain-specific ontology concept extraction and hierarchy extension. In: Proceedings of the 2nd International Confernce on Natural Language Processing and Information Retrieval, ACM, (pp. 60–64).
Zhao, Z., Yang, Y., Li, C., & Nie, L. (2020). Guessuneed: Recommending courses via neural attention network and course prerequisite relation embeddings. *ACM Transactions on Multimedia Computing, Communications, and Applications, 16*(4), 1–17.
Zhou, Y., Xiao, K. (2019). Extracting prerequisite relations among concepts in Wikipedia. 2019 International IEEE: Joint Conference on Neural Networks. (pp. 1–8).

Authors and Affiliations

Chiara Alzetta[1] · Ilaria Torre[2] · Frosina Koceva[3]

✉ Ilaria Torre
ilaria.torre@unige.it

Chiara Alzetta
chiara.alzetta@ilc.cnr.it

[1] Institute of Computational Linguistics "A. Zampolli", ItaliaNLP Lab, CNR-ILC, Pisa, Italy

[2] Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genoa, Italy

[3] E-Learning and Knowledge Management Lab, DIBRIS, University of Genoa, Genoa, Italy