ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 10, Number 1
june 2024

aA Accademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 9791255000983

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_10_1

aAccademia university press

Accademia University Press è un marchio registrato di proprietà
di LEXIS Compagnia Editoriale in Torino srl

IJCoL

Volume 10, Number 1
june 2024

CONTENTS

Investigating the Interplay between Text Difficulty and Prerequisite Relation Identification in Educational Texts

Chiara Alzetta*
CNR, Istituto di Linguistica
Computazionale 'A.Zampolli'

Prerequisite relations (PR) are fundamental in knowledge acquisition and the applications of Artificial Intelligence to distance learning, particularly with regard to personalized learning plans. The role of these relations is to specify the sequence of information acquisition necessary for understanding a target concept. Despite their significance, identifying PRs in educational texts is challenging, mainly due to the lack of systematic procedures for their identification on educational texts. This paper contributes to the ongoing research on PR identification by exploring the relationship between text difficulty, assessed across various linguistic properties and target audiences, and prerequisite relations. We conducted a crowd-based study on the novel task of prerequisite concept ordering. The study yielded preliminary yet valuable insights into the impact of text difficulty on the task. Such evidence sheds light on the need to account for the linguistic properties of texts when identifying PRs, thus advancing the field's comprehension of PRs within the educational landscape. Ultimately, we hope that this work could foster novel linguistically-aware research on PR.

1. Introduction and Motivation

Educational materials, such as textbooks and lecture notes, serve as vital resources for providing students with knowledge about topics and subject matters. Domain experts and educational publishers design such materials with the goal of supporting content comprehension for their target users, i.e., learners. The overarching aim is to prevent frustration, misunderstanding, and disorientation among learners during the learning process (Gagne 1962). Achieving this goal requires a well-thought presentation of concepts within the educational text, which should follow the concepts' propaedeutic order. This order is effectively conveyed through prerequisite relations (PR).

In education, PRs hold great relevance as they encode the sequence in which concepts should be acquired. For instance, in arithmetic classes, typically $addition \prec multiplication$ (read as "addition is a prerequisite for multiplication"), since the former concept is usually introduced before the latter in the learning process. Formally, a PR represents a binary dependency relationship connecting a 'prerequisite' concept with a 'target' concept, where the former must be comprehended before the latter (Johnson-Laird 1980; Liang et al. 2019). This definition is widely adopted in the PR literature, a field of research dealing with the creation of datasets annotated with explicit PRs

* CNR, Istituto di Linguistica Computazionale 'A.Zampolli', ItaliaNLP Lab - via G.Moruzzi, 1, Pisa, Italy
E-mail: chiara.alzetta@ilc.cnr.it

between educational concepts (Chaplot et al. 2016; Fabbri et al. 2018; Talukdar and Cohen 2012; Wang et al. 2016; Alzetta, Torre, and Koceva 2023) and the development of systems for their automated identification within educational materials (Adorni et al. 2019; De Medio et al. 2016; Miaschi et al. 2019; Sabnis et al. 2021; Sayyadiharikandeh et al. 2019; Zhu and Zamani 2022). These efforts are driven by the goal of enriching educational applications with knowledge structures representing PRs, such as automatically-generated study plans (Agrawal, Golshan, and Papalexakis 2016; Gasparetti, Limongelli, and Sciarrone 2015; Zhao et al. 2021) and educational contents (Liang et al. 2016; Lu et al. 2019). This literature frequently leverages a quite operational definition of 'concept', intended as a piece of domain knowledge represented in an educational text by means of domain terms (single or multi-word noun phrases, such as *multiplication* or *natural number*) (Chau et al. 2021; Pan et al. 2017; Talukdar and Cohen 2012). We too adopt such a perspective in this work.

Anchoring concepts to text portions, while simplifying the process of identifying the concepts referred to in a textual document, has also opened lines of research aiming to identify PRs solely from the content of educational texts (Adorni et al. 2019; Lu et al. 2019; Wang et al. 2016; Alzetta et al. 2020; Zhu and Zamani 2022; Stamper et al. 2023). These are opposed to methodologies that rely on ontologies and structured knowledge bases to identify concepts and their relations. While the latter methods are generally prevalent (see. e.g., (Gordon et al. 2016; Roy et al. 2019; Talukdar and Cohen 2012; Zhou and Xiao 2019; Bai et al. 2021; Ma et al. 2022), they may face limitations when applied to domains lacking comprehensive coverage in external knowledge sources. On the other hand, the novel scenario that exploits solely the textual content of the source texts opens new avenues in PR research, with a stronger focus on using natural language processing (NLP) approaches to examine whether linguistic and semantic properties of the text influence PR identification. This work aims to contribute to such a line of research by investigating a relatively unexplored aspect: the interplay between the difficulty of texts designed for educational purposes and the ability of human annotators to identify PRs between the concepts therein.

Text difficulty refers to the accessibility of text to the reader, encompassing a broad range of factors that influence how challenging a text may be perceived by its intended audience (Fulcher 1997). In the literature, text difficulty has traditionally been assessed using readability metrics, which typically consider factors like text length and word frequency, often overlooking other, more fine-grained, linguistic properties known to affect text comprehension and perceived difficulty (Brunato et al. 2018; McNamara, Graesser, and LouwerseMax 2012). Notably, there is a nuanced but substantial difference between the meaning of text difficulty and text complexity (Cunningham and Anne Mesmer 2014; Pelánek, Effenberger, and Čechák 2022; Beckmann, Birney, and Goode 2017; Mesmer, Cunningham, and Hiebert 2012). Text complexity is used when accounting for specific linguistic or textual properties of a text that can be manipulated and that, collectively, influence its overall complexity. These properties might include elements of words, syntax, or discourse. Text difficulty, as mentioned above, extends beyond these elements to encompass the interaction between linguistic and textual features and reader characteristics. Thus, when using the term 'complexity', we refer, globally or individually, to the linguistic features of a text. Conversely, 'difficulty' is used with a broader meaning that subtends readers' perception. This dual perspective allows us to holistically measure the overall difficulty of texts, aligning with established definitions in the literature and providing a comprehensive understanding of text characteristics in our analyses.

Educational research has explored the role of text difficulty in the learning process (Frantz, Starr, and Bailey 2015; Chall, Conard, and Harris-Sharples 1991; Pelánek, Effenberger, and Čechák 2022), primarily driven by the understanding that language is a pivotal medium for learning (Halliday 1993). The evidence of these studies suggests that the complexity of educational texts negatively impacts reading comprehension (Spencer et al. 2019; Benjamin 2012) and is associated with higher levels of mind wandering among students (Feng, D'Mello, and Graesser 2013). Easier-to-read texts appear to be more conducive to learning, as they facilitate the recognition of connections between concepts. Such intuition was explored by (Manrique et al. 2018) and (Angel, Aroyehun, and Gelbukh 2020), who incorporated readability and complexity-based features for training automatic prerequisite learning models. However, neither tested the impact and significance of those features. We aim to build upon their findings and address the still unanswered research question:

- **RQ**: Does the textual difficulty of educational materials affect learners' ability to recognise the sequence of concept acquisition?

To investigate this question, we conducted a pilot crowd-based study on concept ordering. Sixty participants were tasked with ordering triples of concepts based on their prerequisite order after reading short concept descriptions at different difficulty levels. Given that readers can acquire information about a concept solely from its description, we used this study to explore the following complementary hypotheses:

- **HP1**: Easier-to-read texts convey PRs between concepts clearly and unambiguously.
- **HP2**: Difficult-to-read texts pose challenges in abstracting relations, making the identification of PRs more arduous.

Drawing upon HP1 and HP2, we expect that judgements regarding prerequisite ordering based on easier-to-read concept descriptions would exhibit consistency among participants owing to the less ambiguous expression of relations. Conversely, judgements of prerequisite ordering stemming from difficult-to-read concept descriptions might display lower consistency, indicating greater difficulty in abstracting relations from the text.

Among our main contributions is the resource we have newly created for this study, which we named the *Concept Description Variations* corpus. The novel resource comprises parallel descriptions of thirty concepts at three different difficulty degrees, for a total of 90 concept descriptions. Details about the resource and its construction process are presented in Section 2.2. For transparency and reproducibility, all data and materials from the study are made available through an online repository and freely accessible for research purposes.

The remainder of the paper is organised as follows. First, we present the methodological details of the study. Specifically, the task presentation and the data are reported in Section 2, while the experimental design of the crowd-based study is discussed in Section 3. Then, in Section 4 we report the results of the study on concept prerequisite ordering. Section 5 discusses the obtained results in light of the presented hypothesis and existing literature. We conclude the paper in Section 6.

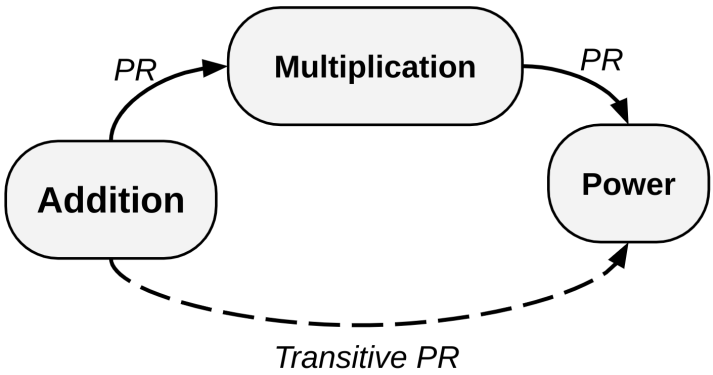

Figure 1
Example of PR relations between the concepts of 'Addition', 'Multiplication' and 'Power'. The dashed line represents the transitive PR.

2. Study Goals and Data

This section outlines the task defined for the crowd-based study, namely the *prerequisite concept ordering* task. The proposed task draws inspiration from the task of knowledge sequencing, a fundamental principle in instructional design and curriculum development across various educational settings (Brusilovsky 1992). Knowledge sequencing is typically performed by teachers, who structure educational content to guide learners from foundational to advanced topics. Such a structured progression is aimed at fostering effective learning, leveraging an approach that builds upon prior knowledge, thereby facilitating comprehension, retention, and skill development. By formalizing the learning order between two pieces of educational content, PRs enable the construction of concept sequences reflecting the optimal order for learning. As we delve into in Section 2.1, the construction of such sequences is informed by the presentation of concepts within educational texts.

Following the description of the prerequisite concept ordering task, in Section 2.2 we will offer an in-depth overview of the Concept Description Variation corpus employed in the study, discussing its construction process and detailing its composition.

2.1 Prerequisite Concept Ordering Task

Prerequisite concept ordering is a novel task which consists of manually ordering concepts according to the ideal sequence in which they should be presented in educational materials. This sequencing determines the order in which concepts should be introduced to prevent learners' disorientation and optimise comprehension. The sequence is formally represented through prerequisite relations, binary and directed relations that indicate which, between two concepts, should be acquired first by a learner (Johnson-Laird 1980). It should be noted that PRs, in addition to being binary and directed, hold the following properties: they are irreflexive, meaning that related concepts must be distinct, and transitive, implying that if concept x is a prerequisite of concept y, and y is a prerequisite of concept z, then x is also a prerequisite of z. These properties are exemplified in Figure 1 between the concepts of 'Addition', 'Multiplication' and 'Power'.

Following from the above, the task of prerequisite concept ordering is formally defined as follows: given three concepts A, B, and C, each accompanied by a concise descriptive text t_A, t_B and t_C, create a triple of prerequisites $T = (t_A \prec t_B \prec t_C)$ to indicate that $A \prec B$ and $B \prec C$.[1]

Hence, to align with the definition of PR, according to which a PR is a relation that subtends the requirement of mastering the most basic concept before acquiring the most advanced one, each triple T conveys the following information:

1. t_A introduces the foundational knowledge necessary to understand both t_B and t_C;
2. t_B can be understood only if t_A is known;
3. t_C requires familiarity with both t_A and t_B for complete understanding.

These three points underscore the importance of accounting for the content of the three concept descriptions when creating the sequence. In fact, rather than creating absolute resource-independent PRs, the proposed task requires building PR sequences that emerge from reading the concept descriptions. Consequently, the annotation process should assess how the content of each text informs the prerequisite relationships within the triple, reflecting the progressive nature of concept acquisition.

2.2 The Concept Description Variation Corpus

The novel 'concept description variation' corpus serves as a parallel resource, offering concise concept descriptions at three distinct levels of difficulty. This corpus stands as an original contribution of this work and is made publicly available online to ensure transparency and reproducibility.[2]

To address the research question posed in this study and align with the objectives of the prerequisite concept ordering task, the corpus features triples of concepts that, like the concepts reported in the example representation of Figure 1, are related by prerequisite relations. Specifically, the corpus comprises the ten concept triples, for a total of thirty distinct concepts, reported in Table 1. Despite its relatively limited size, the corpus covers a varied range of educational content.

The concept triples were sourced from AL-CPL (Liang et al. 2019), a dataset of concept pairs manually annotated with prerequisite relations by domain experts across four domains: geometry, precalculus, physics, and data mining. Notably, AL-CPL concept pairs have been validated as PR by three domain experts based on their domain-specific background knowledge. Only those pairs consistently labelled as PRs by the majority of annotators are included in our corpus, ensuring their accuracy as prerequisite relations. We included in the *Concept description variation* corpus only concept triples that appear in AL-CPL as $A \prec B$, $B \prec C$ and $A \prec C$.

Having defined the ten triples, we collected three descriptions for each concept. These concept descriptions consist of concise English texts, typically spanning 3 to 5 sentences and averaging around 100 tokens. These descriptions were gathered between March and April 2020 from the following three distinct sources:

1 We recall that $\prec$ should be read as *'is prerequisite of'*.
2 `https://github.com/chiaralz1/PR_difficulty`

Table 1
Concept triples of the Concept Description Variation corpus and gold prerequisite orderings based on AL-CPL Dataset.

Domain	#	Concept Triples: Concept of text A	Concept of text B	Concept of text C	Gold Ordering: Concepts	Sequences
Geometry	1	Geometry	Cone	Circle	Geometry - Circle - Cone	ACB
	2	Line	Angle	Point	Point - Line - Angle	CAB
	3	Addition	Summation	Arithmetic	Arithmetic - Addition - Summation	CAB
Physics	4	Gravity	Gravitational field	Physics	Physics - Gravity - Gravitational Field	CAB
	5	Skew Lines	Line	Parallel	Line - Parallel - Skew Lines	BCA
	6	Acceleration	Speed	Motion	Motion - Speed - Acceleration	CBA
Precalculus	7	Deformation	Hooke's Law	Elasticity	Deformation - Elasticity - Hooke's Law	ACB
	8	Polynomial	Number	Integer	Number - Integer - Polynomial	BCA
	9	Function	Mathematics	Limit of a function	Mathematics - Function - Limit of a function	BAC
Data Mining	10	Sample	Statistical significance	Confidence interval	Sample - Confidence interval - Statistical significance	ACB

- **Simple English Wikipedia**[3]: An online free encyclopedia written at a basic level of English for learners with cognitive impairments or early learners of English as a second language (Jatowt and Tanaka 2012; Vajjala and Meurers 2014).
- **English Wikipedia**[4]: An online encyclopedia created and maintained by volunteer contributors with the goal of disseminating knowledge to a broad audience.
- **Specialised Encyclopedias**: We relied on the Encyclopaedia of Mathematics[5] for precalculus, data mining, and geometry concepts, and on the Encyclopaedia of Physics[6] for physics concepts. Unlike the previous two, these encyclopedias focus on single domains and are tailored for domain experts who already mastered fundamental knowledge of the discipline.

All three sources are works of encyclopedic scope, organised in entries (articles) and providing factual information about the concept covered. Notably, these sources target distinct audiences, resulting in varying levels of text difficulty (Dale and Chall 1949; Fulcher 1997). Simple English Wikipedia employs basic language suitable for learners with cognitive impairments or non-native English speakers. English Wikipedia is designed to serve a general audience, whereas specialised encyclopedias target domain

3 https://simple.wikipedia.org
4 https://en.wikipedia.org
5 https://encyclopediaofmath.org
6 Besancon, R. (2013). *The Encyclopedia of Physics.* Springer Science and Business Media.

experts. The diversity in audience and purpose contributes to differences in text complexity. We also empirically evaluated the text complexity of such sources accounting for a wide range of linguistic phenomena (cf. Sec. 4.1).

To collect the concept descriptions for the thirty concepts of the corpus, we initially identified the articles corresponding to each concept in the three selected sources. Then, we extracted the first lines of each page, roughly encompassing the initial 3 to 5 sentences. This approach was guided by the intuition that the opening lines of an encyclopedic article typically provide a concise definition of the concept. Through this process, we obtained a total of 90 concept descriptions, with 30 descriptions sourced from each of the three selected sources. Accordingly, t_xS, t_xW and t_xE refer to the descriptions of the concept acquired, respectively, from Simple Wikipedia (S), Wikipedia (W) and encyclopedias (E). Eventually, we manually reviewed all texts to ensure that they indeed contained the information necessary to establish the gold PRs orderings between concepts of the Concept Description Variation corpus.

3. Crowd-based Study on Prerequisite Concept Ordering

In this section, we describe how the concept triples of the Concept Description Variation corpus were employed in a crowd-based study to investigate our primary research question: Is there an impact of text difficulty on the manual identification of concept sequences? This research question is explored through the Prerequisite Concept Ordering task presented to the study participants.

In what follows, we will first outline the design of our crowd-sourced study and introduce the recruited participants in Section 3.1. Then, in Section 3.2, we will describe the tools and metrics employed to analyse the data and evaluate the responses provided by the participants.

3.1 Crowd-sourcing Design

We defined three tasks on prerequisite concepts ordering administered through Prolific[7], a crowd-sourcing platform that allows to recruit and pay participants. Each task is presented to participants through a distinct questionnaire Q_y, where $y = \langle S, W, E \rangle$ depending on the source of the concept descriptions. In these questionnaires, participants are tasked with establishing the prerequisite sequence for 10 randomly arranged concept triples (T). Notably, the order of triples and concepts is consistent across questionnaires, thus the source of concept descriptions is the only variable at play. For instance, the concept triple #1 of Table 1, $T = (Geometry \prec Circle \prec Cone)$, is represented in Q_S through the descriptions of the concepts *geometry*, *circle* and *cone* acquired from Simple Wikipedia. In contrast, the same triple is introduced in Q_W with descriptions sourced from English Wikipedia and in Q_E with descriptions obtained from the Encyclopaedia of Mathematics. All questionnaires include two control questions to identify participants who provided unreliable responses.

Before taking the questionnaire, participants are provided with instructions to guide them through the task. The instructions are complemented by a solved example question, displayed in Figure 2. This example serves to familiarise participants with the questionnaire's structure and the ordering task.

7 https://www.prolific.co

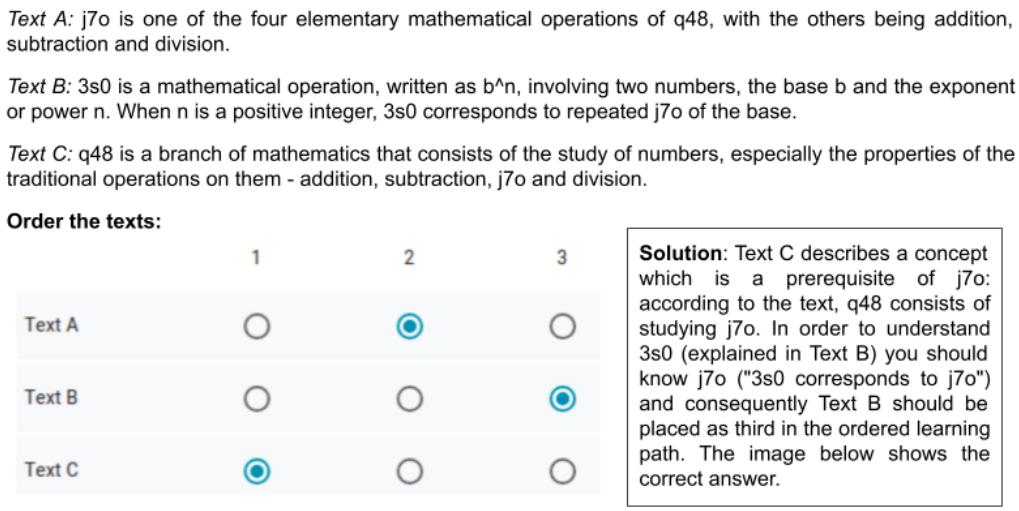

Figure 2
Test question presented to participants and solution explanation.

Notably, the concepts mentioned in the text are anonymised using alphanumeric codes. This was done to prevent any potential facilitation effect arising from background knowledge, aligning with established best practices in cognitive science research (Anderson and Pearson 2016; Weber 1991).

Upon completing the questionnaire, participants are asked to provide feedback by rating the task's difficulty on a 5-point Likert scale where 1 equals "very difficult" and 5 signifies "very easy".

3.1.1 Participants

We recruited a total of 60 participants. Specifically, we allocated 20 participants to each of the three questionnaires Q_S, Q_W and Q_E. Since the study involved English texts, being English native speakers was an essential requirement for participation. Another factor that we deemed significant was the education level, thus we required participants to have a minimum of secondary education. These criteria were established to ensure a homogeneous group with a shared language background and familiarity with educational textual content. Notably, gender was not controlled for in this study since we align with research indicating no substantial gender-related differences in individuals' ability to learn and comprehend textual content (Asgarabadi, Rouhi, and Jafarigohar 2015; Fahim, Barjesteh, and Vaseghi 2012).

Before starting the study, all subjects were informed about the research objectives and the nature of data collection. In accordance with ethical guidelines, informed consent was obtained from each participant, who also agreed to the terms and conditions of the study. To ensure the anonymity of any individual contributor, any data collected that could potentially identify an individual, such as age and gender, will be presented in aggregate form. Additionally, all data were stored securely and only accessed by authorised researchers involved in the study.

To ensure the quality and reliability of the data in the analyses, we excluded from the study participants who completed the questionnaire in less than five minutes (experimentally defined as the minimum time possible for completing the questions) and

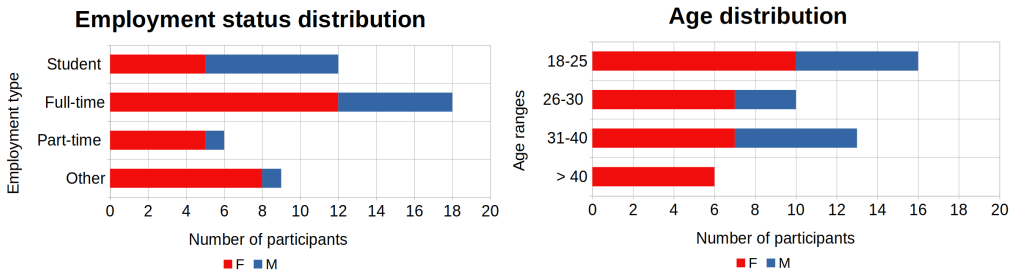

Figure 3
Employment status and range of age in the subgroups of females (F) and males (M) participants.

those who failed the control questions. Consequently, our analysis was based on the responses of 45 participants[8] (30 females; average age = 30.91; SD = ±11.1), namely 15 individuals for each of the three questionnaires. Details about the demographics of the full set of participants are reported in Figure 3. As we can see from the plot on the left, most participants declared to have a full-time job (12 females, 6 males, 18 in total). The plot on the right shows that 16 participants (10 females, 6 males) out of 45 are between 18 and 25 years of age. Note that the proportions of these distributions hold consistently within the subgroups who took the three questionnaires.

3.2 Analysis Tools and Metrics

Linguistic Profiling. To assess the linguistic complexity of concept descriptions, we employed Profiling-UD (Brunato et al. 2020), a web-based application designed to capture a comprehensive range of linguistic characteristics that contribute to characterising language variation within and across texts (Brunato et al. 2020; Deutsch, Jasbi, and Shieber 2020; van Halteren 2000). These properties encompass various aspects, including raw text features, lexical diversity, morpho-syntactic information, verbal predicate structure, global and local parse tree structures, syntactic relations, and the use of subordination. This information is extracted from linguistically analysed texts and automatically parsed following the Universal Dependencies annotation schema (de Marneffe et al. 2021). The complete set of features is detailed in Appendix A.

Difficulty Variations. To evaluate the linguistic variation among concept descriptions acquired from different sources, we employed the non-parametric Kruskal-Wallis (KW) test, which provides a means of determining whether statistically significant differences exist between the values of variables between three or more independent groups. Additionally, we employed Principal Component Analysis (PCA) to visually inspect the data. PCA is a classic data analysis method that reduces data dimensionality while preserving most of the variation by identifying principal components that capture maximal data variance (Jolliffe and Cadima 2016).

Questionnaire Analysis. The questionnaires underwent an in-depth analysis focusing on two key factors: *completion time* and *question accuracy*. The former measures the time

8 Participants providing valid responses were compensated at 6.27£ per hour, a payment rate certified as 'Fair' by the Prolific platform.

taken by a subject to complete a questionnaire, and the latter accounts for the percentage of sequences matching the order in the AL-CPL dataset. We employed classical statistical metrics to assess the significance of variations between the answers of each participant. Specifically, we relied on the T-test and the Pearson Correlation Coefficient (PCC) metrics. The former determines whether there is a statistically significant difference between the means of independent groups, while PCC is a measure of the linear relationship between two continuous variables. These are complementary measures since the T-test helps verify whether a variation in the questionnaire setting resulted in statistically different results, while PPC assesses the strength and direction of the relationship between the prerequisite sequences produced in different settings.

4. Results

4.1 Linguistic Profile of Concept Description Sources

The first step of the analysis concerns the assessment of the complexity level of concept descriptions. To this aim, we relied on the set of features acquired using Profiling-UD, which reveals the distribution of various linguistic phenomena within the short texts.[9]

Using KW, we identified the linguistic properties that vary significantly across the text sources, namely, the Profiling-UD features that exhibited statistically significant variations of their values among the concept descriptions acquired from Simple Wikipedia, English Wikipedia and Specialised Encyclopedias. Table 2 reports the mean values and standard deviations of the features showing significant variance according to the KW test. For transparency, in Appendix B we report the KW χ score and means values for all the features acquired using Profiling-UD, including those whose values do not vary significantly across the three concept description sources.

The variations between concept descriptions of different sources encompass all the linguistic levels monitored by the tool. Notably, the most significant variations are observed in the structure of parsed trees (*'Tree Structure'* group), the properties of raw text, and the distribution of syntactic dependencies. Specifically, we observe that Simple Wikipedia texts consistently yield flatter and shorter syntactic trees than Wikipedia and Encyclopaedia texts. As evidence, consider the average values of the features *Depth avg max*, which measures the mean distance from the root to the furthest leaf node in the parsed syntactic trees, and *Link len. max*, i.e. the average length of the longest link in each tree. As shown in Table 2, while the values of these features are relatively similar between Wikipedia and Encyclopedia sub-corpora, the values computed on Simple Wikipedia texts are significantly lower. This is particularly evident in the length of the longest links: in Simple Wikipedia, they are about half as long (6.95) as the links in the other texts (12.81 for Wikipedia; 11.73 for Encyclopedia).

9 The raw data obtained from the Profiling-UD analyses are available in full at `https://github.com/chiaralz1/PR_difficulty`, 'Linguistic Analysis' folder.

Table 2: Features that vary significantly across sources according to the KW test. For all features, it is reported the average value and standard deviation in each source. Significance levels (p-values) are indicated as follows: * ($p < 0.05$, significant), ** ($p < 0.01$, highly significant), and *** ($p < 0.001$, extremely significant).

	Mean Values (standard deviation)		
Feature	**Simple Wiki**	**Wikipedia**	**Encyclopedia**
	Raw Text		
Sent len (tokens)***	15.21 ±8.3	26.71 ±13.2	24.29 ±11.9
Char. per token***	4.45 ±1.1	4.83 ±0.7	4.99 ±0.9
	UPOS Distribution		
Adjectives***	6.25 ±7	7.77 ±5.5	10.4 ±8.2
Adpositions***	8.42 ±6.9	12.13 ±6	13.28 ±5.8
Auxiliaries***	7.45 ±5.8	6.33 ±4.6	4.6 ±4
Coord. Conj.**	2.64 ±4.1	3.26 ±3.2	2.37 ±3.5
Particle***	1.33 ±3	1.58 ±2.6	0.43 ±1.3
Punctuation**	15.96 ±13	13.11 ±6.7	12.24 ±7.7
	Verb Inflection		
Verb past***	30.85 ±43.7	54.86 ±46	51.87 ±46.8
Verb pres*	35.22 ±45.4	23.95 ±38.1	20.93 ±36.6
Verb ger**	6.3 ±19.7	9.34 ±20.4	13.63 ±27.2
Verb inf**	7.98 ±20.2	12.36 ±22.7	5.89 ±16.7
Verb part**	26.09 ±39	37.92 ±37.2	39.16 ±41.5
Aux pres*	67.06 ±47	76.78 ±41.7	61.2 ±48.3
Aux ind*	69.64 ±46.1	80.51 ±39.8	64.8 ±48
Aux inf*	3.56 ±12.1	6.78 ±15.1	4 ±14.6
	Verb Predicate		
Verb heads per sent.***	1.84 ±1.2	2.74 ±2	2.06 ±1.5
Perc. verbal roots***	88.69 ±31.8	94.07 ±23.7	71.2 ±45.5
Verb edges 0**	3.82 ±16.6	7.75 ±18.9	3.73 ±13
Verb edges 1*	6.05 ±20.2	9.37 ±21.6	11.88 ±24.2
Verb edges 5***	2.43 ±14.1	4.07 ±11.2	10.2 ±26.4
	Tree Structure		
Depth avg max***	3.39 ±1.3	4.5 ±1.5	4.66 ±1.7
Tok per clause avg***	8.26 ±4.9	11.85 ±6.8	11.6 ±8.2
Link len. avg***	2.29 ±0.7	2.76 ±0.6	2.68 ±0.7
Link len. max***	6.95 ±5.1	12.81 ±8.3	11.73 ±7.9
Prep. chain len. avg***	0.61 ±0.7	0.99 ±0.7	1.13 ±0.7
n. prep. chains***	0.67 ±0.8	1.53 ±1.3	1.63 ±1.1
Prep dist 1***	43.25 ±48.6	66.1 ±45	69.22 ±39.7
Prep dist 2***	6.25 ±22.7	9.39 ±25.5	13.26 ±25.6
	Order		
Subject pre***	92.11 ±26.1	97.32 ±14.3	82.64 ±37.6
	Syntactic Dependencies Distribution		
Adjectival modifier***	4.58 ±5.7	6.87 ±5.3	8.99 ±7.3
Appositional mod.***	0.56 ±2.1	0.7 ±2.1	1.49 ±2.7
Auxiliary pass*	1.93 ±3.7	1.92 ±2.9	2.41 ±3.1
Case marker***	8.61 ±7.1	12.57 ±6.2	13.41 ±5.8

Table 2 continued from previous page

	Mean Values (standard deviation)		
Feature	**Simple Wiki**	**Wikipedia**	**Encyclopedia**
Cordinating conj.**	2.6 ±4	3.27 ±3.2	2.41 ±3.5
Compound**	1.8 ±3.9	2.46 ±3.7	2.59 ±4.1
Conjunction**	3.44 ±5.5	4.64 ±4.9	3.52 ±4.7
Copula***	4.51 ±4.8	3.26 ±3.3	1.49 ±3
Goes with*	0 ±0	0 ±0	0.13 ±1
Marker*	2.16 ±3.8	2.05 ±3	1 ±2.1
Nominal modifier***	5.19 ±5.8	7.37 ±6.1	8.53 ±5.7
Nominal subject***	8.25 ±5.4	5.3 ±3.8	3.46 ±3.9
Oblique compl.***	2.84 ±4.2	4.47 ±3.9	4.32 ±4.2
Parataxis***	0.37 ±1.7	0.56 ±1.4	0.07 ±0.4
Punctuation*	14.84 ±9.4	13.19 ±6.8	12.35 ±7.9
Root***	9.31 ±10.7	4.82 ±2.7	6.01 ±6.1
Subordinate Structure			
Dist. of principal prop.***	61.91 ±35.4	56.68 ±33	43.87 ±37.4
Dist. of subord. prop.**	30.35 ±31.8	41.62 ±32.6	42.53 ±37.2
Subordinate post*	39.53 ±48	56.75 ±46.9	46.6 ±49.4
Subord chain len avg*	0.59 ±0.6	0.78 ±0.7	0.73 ±0.7
Subordinate dist 1*	43.25 ±49.6	57.84 ±47.2	50.53 ±49.8

Additionally, it is worth highlighting that Simple Wikipedia sentences tend to feature fewer embedded chains of nominal modifiers, as captured by the feature *n. prep. chains* of the *'Tree Structure'* group. This characteristic contributes to an overall simplification of the sentences in Simple Wikipedia texts, which obtain a feature value of 0.67, compared to the 1.53 and 1.63 for Wikipedia and Encyclopedia, respectively. To illustrate, consider the two embedded nominal chains acquired from Simple Wikipedia descriptions, *'distance of an object'* and *'amount of time'*, in contrast to the following chain acquired from an Encyclopaedia text: *'the rate of change of the speed of an object'*. In the former examples, the chain comprises only two nouns, whereas the Encyclopaedia chain contains four nouns. However, it is worth highlighting that the standard deviation of this feature in Simple Wikipedia is higher than the average value, indicating significant variability across the texts. Conversely, texts from other sources exhibit a smaller standard deviation compared to the mean value, suggesting more stable feature values in these sub-corpora.

Concerning traits that suggest higher linguistic complexity, specialised encyclopedia texts, that represent the difficult variety of the corpus, exhibit a richer subordinate structure. This is captured by the *Dist. subord prop.* feature, which measures the distribution of subordinate propositions. Interestingly, Wikipedia scores slightly lower than specialized encyclopedia texts (41.62 versus 42.53, respectively). This, along with other traits discussed, suggests that texts from these two sources are more similar in terms of syntactic structure than texts from Simple Wikipedia.

The higher complexity of Specialised Encyclopedia texts is also suggested by the higher number of dependency links, encompassing both arguments and modifiers, all centred around the same verbal head. This trait is evidenced by the features in the *'Verb Predicate'* group, and in particular by the *Verb edges avg* feature, which shows that the average amount of dependents of verbs increases with source complexity. Although

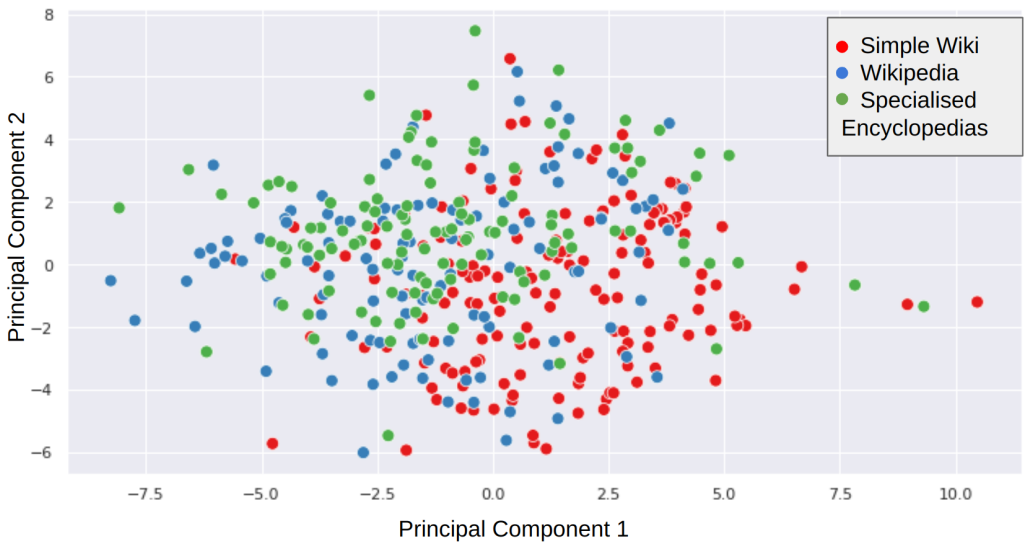

Figure 4
PCA visualization of the sentences in the concept descriptions.

this feature is considered non-significant (as shown in Appendix B), possibly due to the prevalence of verbs with 2 or 3 dependents, a more detailed look reveals a significantly higher frequency of verbs with five dependents in Encyclopedia sentences compared to other sources (see the *Verb edges 5* feature). With this regard, it is interesting to highlight that the frequency of nominal subjects, a syntactic relation depending on the verb, shows a higher frequency in Simple Wiki and Wikipedia. This aligns with the higher frequency in Specialised Encyclopedia texts of oblique, nominal and adverbial modifiers, as well as passive auxiliaries that are more typical in constructions found in scientific literature (e.g. *'the gravity force was first recognized by Sir Issac Newton'*).

The visual representation offered by the PCA further highlights differences due to the different linguistic properties of texts. Figure 4 displays how sentences acquired from the different description sources exhibit distinct spatial distributions. Notably, a predominant concentration of sentences is observed towards the centre of the plot. This may be due to the commonalities in terms of textual genre across the texts (i.e., articles in encyclopedias). Upon closer inspection, the PCA analysis reveals that sentences sourced from Simple Wikipedia (depicted by red dots) tend to aggregate towards the right side of the plot. Conversely, sentences from Specialized Encyclopedias (green dots) exhibit a prevalence towards the left side, indicating distinct linguistic attributes associated with this source. In contrast, sentences extracted from Wikipedia (blue dots) present a more scattered distribution across the plot, leaning slightly towards the left. This visual representation underscores subtle yet significant distinctions in the linguistic characteristics of sentences based on their source. As we will discuss in Section 5, the observed patterns align coherently with both the output of the linguistic profiling analyses and our expectations about the distinct linguistic structures that contribute to the overall complexity of the texts.

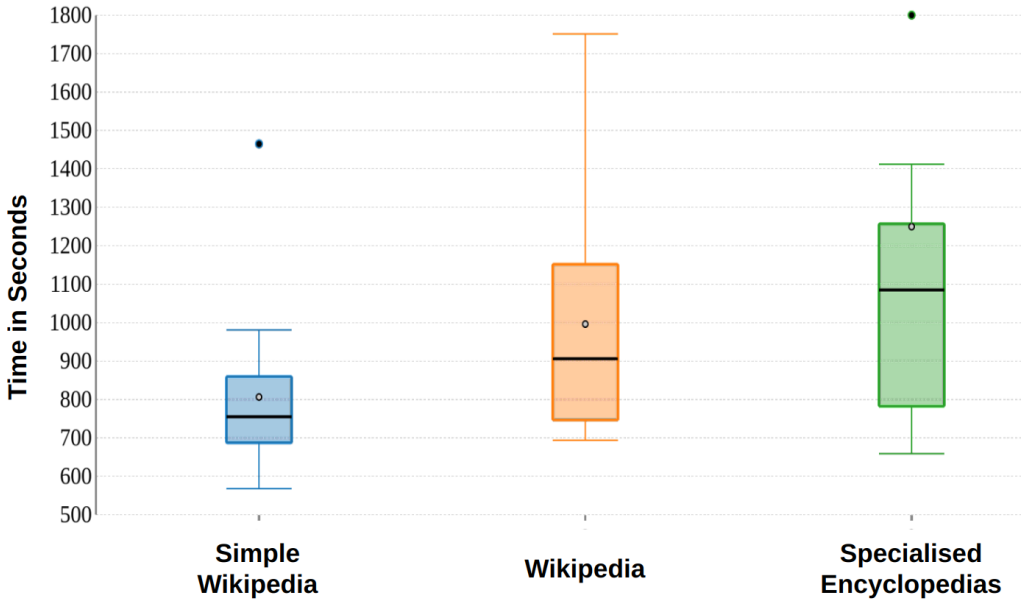

Figure 5
Completion times of questionnaires. White dots represent the mean value of the group; black bolded lines represent the median time; dots outside the plot present outliers.

4.2 Questionnaire Analyses

4.2.1 Completion Time

Figure 5 displays the completion times of each group of participants. Notably, the average time required to complete the questionnaires increases as the text difficulty level rises. For instance, the group presented with Q_S completed the study more quickly than the other groups. Specifically, the mean time for completing the questionnaire is 0:13:26[10] (SD = ±0:03:35) for the Simple Wikipedia group, whereas the Encyclopaedia group shows a mean time of 0:20:50 (SD = ±0:13:05). Q_W, on the other hand, required an average of 0:16:36 (SD = ±0:05:15) to complete, it between the completion times of Q_S and Q_E.

Figure 5 reveals that Q_S and Q_E are characterized by the presence of one outlier each, which is not attested in Q_W. The completion time of these outliers is 0:24:25 for Q_S and 1:01:16 for Q_E. While the former may correspond to a participant who took particular care in answering the questions, it is likely that the outlier in Q_E simply reflects a participant's lack of experience with the platform or who didn't actively notify the end of the test. If we exclude this outlier, the average completion time for Q_E is 0:17:56 (SD = ±0:07:03), more similar to the values of Q_W.

The Kruskal-Wallis test indicated that there is a significant difference between the time employed to complete the questionnaire ($p < 0.05$). The post-hoc Dunn's test using a Bonferroni correction indicated that the most significant difference is observed between the times of Q_S and Q_E.

10 Time is formatted in standard time format: 0 hours, 13 minutes, 26 seconds.

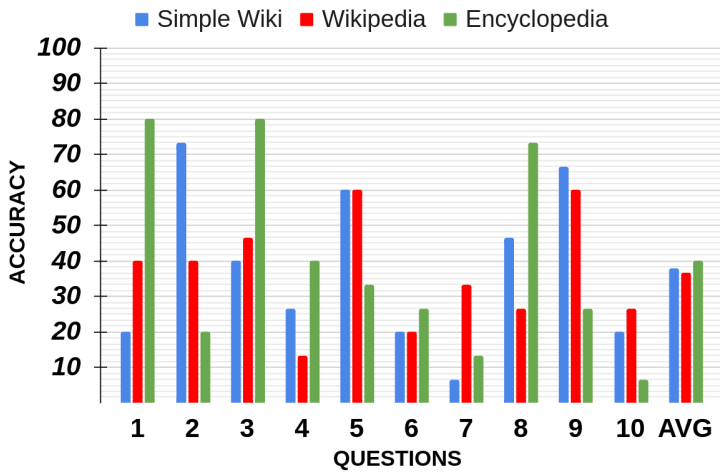

Figure 6
Ordering accuracy of questions in the three questionnaires. Note that question numbers correspond to the numbering of concept triples in Fig. 1.

4.2.2 Question Accuracy

Figure 6 shows the accuracies for each question and on average for the questionnaires (column AVG). Notably, the average accuracies of questionnaires are quite similar, with 38%, 36.7%, and 40% of accurately ordered triples in Q_S, Q_W, and Q_E respectively. The highest reported accuracy is 80% on questions 1 and 3 in Q_E, while the lowest accuracy of the study (<10%) is observed for questions 7 in Q_S and 10 in Q_E. Overall, only 26.66% (8) questions show an accuracy value higher than 50%.

Interestingly, the initial concepts of the triples tend to have higher accuracy rates than the final ones, regardless of the text source. In Q_S initial concepts are correctly identified in 68% of cases, while final concepts in only 45.34%. Similarly, in Q_E, initial concepts show 71.34% of accuracy, while final concepts are only 54.67%. In Q_W the gap is smaller, although still present, with initial and final concepts correctly identified in 58.67% and 51.34% of cases, respectively.

To further investigate the correlation between participants' answers and concept descriptions, we computed PCC on the questionnaires' answers. The highest correlation (PCC=0.54, $p < 0.001$) is observed when comparing the answers of Q_W with Q_S and Q_E. Simple Wikipedia and encyclopedia-based answers show a slightly weaker correlation (PCC=0.45, $p < 0.001$).

4.3 Post-Questionnaire Interview

While all three groups of participants found the task challenging, there is variation in the average difficulty scores among the groups. The group working on Q_S reported an average difficulty score of 2.4 (± 1.14) on the Likert scale, whereas the other groups reported average scores of 1.93 (± 0.96) and 2.0 (± 0.80) for Q_W and Q_E, respectively. The standard deviations suggest that the perceived complexity becomes more consistent within the group as the difficulty of the texts increases.

The KW test results indicate that there is no statistically significant difference between the scores assigned by participants in each group regarding the difficulty level

of the questionnaire ($p = 0.627$). This points to the fact that, despite the differences in average scores, the variations observed in perceived task difficulty across the groups are not statistically significant.

5. Discussion

The first analysis, conducted with the aid of Profiling-UD, revealed that the source of concept descriptions indeed has a notable impact on the distribution of their linguistic properties, and validates our choice of using Simple Wikipedia, English Wikipedia and Specialised Encyclopedias as representative of texts at different difficulty levels. As depicted in Figure 4, sentences obtained from the same source tend to cluster together, especially in the case of Simple Wikipedia and Encyclopaedia texts. This clustering indicates that the texts produced in the contexts of these sources are quite homogenous from a linguistic viewpoint as they exhibit a similar distribution of linguistic phenomena. The average values of linguistic features acquired from the Simple Wikipedia and Encyclopaedia texts using Profiling-UD (see Table 2) further indicate that concept descriptions obtained from these sources show different linguistic properties. These values suggest a spectrum of complexity within the Concept Description Variations corpus, aligned with the expected complexity levels of texts acquired from these sources indicated in the existing literature (Coster and Kauchak 2011; Den Besten and Dalle 2008; Napoles and Dredze 2010; Samoilenko et al. 2018; Snow 2010). This variability likely reflects differing levels of difficulty that readers may experience when engaging with these concept descriptions. In contrast, sentences from Wikipedia are more varied, as shown in the PCA plot (Figure 4). This aligns with prior research, which has highlighted the mixed and highly variable nature of the difficulty in Wikipedia texts (Jatowt and Tanaka 2012).

Focusing on the answers provided in the questionnaires, we notice that all text sources exhibit low question accuracy. This suggests that the task's complexity extends beyond the choice of texts used to describe the concepts. This observation emerges also from existing PR-annotated datasets, which often report low agreement between annotators and, consequently, variable performance of systems trained on such data (Chaplot et al. 2016; Fabbri et al. 2018; Gordon et al. 2016). Common causes of inconsistency in manual annotation include the lack of reproducible annotation procedures and poorly documented annotation guidelines (Ide and Pustejovsky 2017). To address these challenges, previous works on manual prerequisite relation identification have often involved domain experts assessing predetermined concept pairs based on their background knowledge. While this method is prevalent, limitations are frequently found when using these annotations in real-world scenarios: annotated pairs not anchored to a specific text can lead to sequences that may not align with the teaching approach of a given lecture or textbook. Therefore, in designing the prerequisite concept ordering task, we aimed to model the content of the text presented to annotators and implemented strategies to minimise the impact of background knowledge as much as possible. First of all, we required a minimum of secondary education to make sure that all annotators have familiarity with educational contents, but not necessarily specific knowledge about the four investigated domains. Then, we defined the task aimed at modelling the content of the text that annotators were reading rather than representing the abstract domain knowledge. We verified that the pairs from AL-CPL could be identified by reading the concept descriptions and designed task instructions to explicitly guide annotators to rely solely on the text they were reading to create concept sequences. Masking concept names was another measure we implemented to mitigate the impact of background

knowledge on the task, allowing participants to focus more on the text content. Thanks to this design, studies like the present one play a critical role in improving the current definition of PRs, which often tend to be fairly basic and naïve.

Concerning the impact of the difficulty levels of the texts on the task, a comparative analysis of questionnaire accuracies revealed interesting differences. First of all, we observe that the overall task accuracies are quite similar across the text sources. However, the instances where the complexity of the texts is higher seem to convey PRs more clearly. Wikipedia texts, for instance, showed low accuracy, which may be explained by the collaborative editing nature of Wikipedia, where contributors (mainly amateurs and enthusiasts of the domain) may have varying levels of expertise and training in writing educational texts (Dang and Ignat 2016; Shen, Qi, and Baldwin 2017). However, the most striking results were observed in the questionnaires based on Simple Wikipedia and Encyclopaedias. Simple Wikipedia, addressing young learners and readers with cognitive impairments, was expected to convey concept relationships clearly and unambiguously, as we formulated in HP1. Surprisingly, participants were generally more accurate at identifying the ordering of concepts when relying on the descriptions from more difficult texts. This could be attributed to the fact that, beyond the inherent difficulty of the text, encyclopedic concept descriptions, designed for domain experts, are often more precise and accurate, thus revealing concept relationships more clearly. On the other hand, the increased text complexity may encourage readers to engage more deeply with the descriptions, leading to better comprehension and more accurate ordering of concepts.

It should be noted that, although the average accuracy of answers produced when reading Simple Wikipedia and Encyclopaedia texts is quite similar, the answers do not converge on the same triples. As demonstrated by the lack of correlation between the orderings (cf. Section 4.2.2), the questions that showed high accuracy in Simple Wikipedia exhibited low accuracy in the Encyclopaedia questionnaire, and vice versa.

These results indicate that text difficulty significantly influences the outcomes of prerequisite concept ordering tasks, but the effect is contrary to the initial expectations: the more difficult the text is to read, the more accurate the ordering tends to be. However, it's essential to recognize that despite the higher quality of the orderings produced when reading Encyclopaedia texts, the actual and perceived difficulty of the task is greater. This is evident from the longer average reading time required for the questionnaire based on the Encyclopaedia, which can serve as a proxy for text processing effort (Wallot et al. 2014). The extended reading time indicates that participants are spending more time trying to understand the complex text. Furthermore, participants explicitly expressed this higher difficulty in the post-questionnaire interviews. They reported finding the Encyclopaedia texts more challenging to understand and requiring more effort to process. These subjective reports, coupled with the objective measure of longer reading times, highlight the dual nature of complex texts: while they may lead to better comprehension and task performance, they also demand significantly more cognitive resources and effort from the readers. This balance between improved accuracy and increased difficulty is a crucial consideration for designing educational materials and assessments that involve concept ordering tasks.

5.1 Limitations and Future Work

The work presented in this contribution serves as a preliminary study on the novel prerequisite concept ordering. While it has provided valuable insights into the impact

of text difficulty on the task, there are opportunities for more extensive investigations with larger participant groups and expanded sets of questions for each questionnaire.

One noteworthy observation is that lexical properties did not appear to play a significant role in determining text difficulty in this study. In particular, we refer to Profiling-UD features that measure the Type-Token Ratio (TTR) and the lexical diversity of the texts. TTR was computed on lemmas and base forms of tokens as the ratio between the types (i.e., the total number of different words) and the total number of tokens in a concept description text. Lexical diversity represents the ratio of content words (namely, nouns, verbs, adjectives and adverbs) to the total number of words in the concept description. Our analysis revealed that these lexical features did not exhibit significant variations among the Simple Wikipedia, Wikipedia, and Specialised Encyclopedia groups (see Appendix B), and possibly they did not contribute significantly to the observed differences in text difficulty. However, this result may be influenced by the limited size of the corpus and its composition, which consisted of parallel texts (i.e. acquired from different sources but discussing the same concepts). Further research is needed to thoroughly explore the impact of the lexicon on the PR identification task in different settings and on larger text collections.

In the future, this experimental setup could be employed to test the effectiveness of simplification patterns used by professionals to create parallel descriptions of educational content. This could involve asking participants to produce PR orderings based on original and manually simplified texts to investigate which linguistic constructions make PR recognition more challenging or accurate. On a similar note, one could employ textbooks for different grade levels to acquire concept descriptions. While this could offer interesting insights, the cohesive nature of textbooks makes isolating concise sentences defining concepts challenging. This difficulty is compounded in a crowd-sourcing setting, where pinpointing such sentences becomes impractical. The data collected in the current experiment do not enable us to investigate this aspect in such depth and it may require a modification in the experimental design, potentially switching to in-person annotation sessions for a more nuanced analysis.

Additionally, future work could delve into PRs at a more fine-grained level by investigating whether authors of educational materials tend to employ different writing styles when describing fundamental and advanced concepts. This idea was initially explored while examining the accuracy achieved with the initial and final concepts of the triples, where the former can be seen as representing the more fundamental concepts and the latter the more advanced ones. Preliminary findings in this direction suggest the presence of significant differences in linguistic features associated with the verbal predicate structure of texts introducing fundamental and advanced concepts. Specifically, the average values of such features are higher for fundamental concepts, indicating a higher level of complexity for texts describing these concepts compared to those describing advanced ones. In the case of specialised encyclopedia texts, higher accuracy is observed when identifying the first elements of the sequence rather than the last ones. Although these results are currently preliminary and somewhat limited in scope, they offer a glimpse of potential linguistic differences associated with the pedagogical roles of concepts, which should be explored further in future research.

Future research should also explore the use of recent large language models (LLMs) in addressing the challenge of prerequisite concept ordering. While this study primarily delved into the impact of text difficulty on human perceptions, we conducted preliminary experiments using a generative model like ChatGPT for annotation purposes. Notably, when presented with the same questionnaires administered to human participants, the model's responses exhibited a different trend from human annotators. It

produced more accurate concept orderings based on Wikipedia and Simple Wikipedia texts, whereas orderings derived from specialized encyclopedias only matched the gold standard for one question (#7). Notably, this is the question achieving the lowest human annotator accuracy. Moreover, in the cases of Wikipedia and Simple Wikipedia, the model achieved correct orderings for only four and three questions, respectively. These results underscore the necessity for further investigation into the effectiveness of LLMs in this context.

6. Conclusion

This paper presents the results of a study on a novel task, prerequisite concept ordering, which involves arranging triples of concepts in the correct sequence based on the prerequisite relationships among them. We conducted a crowd-based study where three questionnaires, varying with respect to the difficulty of their texts, were administered to multiple participant groups to assess their performance on this task. The crowd-sourcing task is carried out on the basis of the Concept Description Variations corpus, a novel resource of parallel concept descriptions. The corpus, freely available online along with the linguistic analysis of the concept descriptions and the results of the concept ordering task, represents one of the original contributions of this work.

From the results obtained, two key factors emerge. First of all, the results underscore the complexity of the prerequisite ordering task, which extends beyond the difficulty of the texts used to describe the concepts. This complexity is evident in the low question accuracy observed across all text sources, indicating the inherent difficulty of the task itself. Additionally, it is worth highlighting that, nonetheless, the study highlights the significant impact of text difficulty on the task. It suggests that more consistent and reliable prerequisite annotations can be obtained by carefully selecting texts for such studies. Interestingly, it appears that difficult-to-read texts may convey prerequisite relationships more clearly, in contrast to what was expected based on our original hypotheses. This surprising result sheds light on the necessity to carefully consider the texts used for PR annotation (and possibly any type of annotation performed on educational data): while easy-to-read texts, being intended for non-experts in the domain, might seem the most valuable choice to guarantee access to the content to annotators, they might produce sub-optimal results.

We hope that the results of this study foster further research in this field, possibly delving deeper into the pedagogical role of concepts in prerequisite identification. This work serves as a foundational step in understanding the dynamics between text difficulty and prerequisite concept ordering, paving the way for more comprehensive investigations in the future.

Acknowledgments

This work was partially founded by the Small Grant 2022 founded by the Italian association Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD).

Appendix A: Linguistic Features Acquired from Concept Descriptions

Description and label of linguistic features computed by Profiling-UD at the document (concept description) level, aggregated by group.

Linguistic Feature	Label
Raw Text Properties (*Raw Text*)	
Average document length	n. tokens,n. sentences
Average sentence length	Sent len (tokens)
Average word length	Char. per token
Vocabulary Richness (*Lexical Variety*)	
Type/Token Ratio for words and lemmas	TTR (form), TTR (lemma)
Morphosyntactic information (*UPOS Distribution*)	
Distribution of POS	[UD pos tag]
Lexical density	Lexical density
Inflectional morphology (*Verb Inflection*)	
Inflectional morphology of lexical verbs and auxiliaries	Verbs/Aux (tense/mood/num/pers/form)
Verbal Predicate Structure (*Verb Predicate*)	
Distribution of verbal heads per sentence	Verbal heads per sent
Percentage of sentence switch a verbal root	Perc. verbal roots
Verb arity and distribution of verbs by arity	Verb edges n./avg)
Global and Local Parsed Tree Structures (*Tree Structure*)	
Average depth of the whole syntactic tree	Depth (avg max)
Average and maximum dependency link lengths	Link length (avg/max)
Average number of prepositional chains per sentence	n. prep. chains
Average length of prepositional chains and distribution by depth	Prep. chain len. (avg), Prepositional distr. (n)
Average clause length	Tok per clause (avg)
Order of elements (*Order*)	
Relative order of subject and object with respect to the verb	Subject pre/post, Object pre/post
Syntactic Relations (*Syntactic Deps*)	
Distribution of dependency relations	[UD dependency tag]
Use of Subordination (*Subordinate Strucutre*)	
Distribution of principal and subordinate clauses	Dist. of principal/subord. prop.
Average length of subordination chains and distribution by depth	subord. chain len. (avg), subordinate dist. (n)
Relative order of subordinate clauses with respect to the principal proposition	Subordinate (pre/post)

Appendix B: Kruskal-Wallis Test on the Full Set of Profiling-UD Features

Results of the Kruskal-Wallis test (χ) on the full set of features acquired using Profiling-UD from concept descriptions on the Simple Wikipedia, English Wikipedia and Specialised Encyclopedia portions of the 'Concept Description Variation' corpus. Significance levels (p-values) are indicated as follows: * ($p < 0.05$, significant), ** ($p < 0.01$, highly significant), and *** ($p < 0.001$, extremely significant). For all features, the table also reports the average value and standard deviation in each source.

		Mean Values (standard deviation)		
Feature	χ	**Simple Wiki**	**Wikipedia**	**Encyclopedia**
Raw Text				
Sent len (tokens)	85.39***	15.21 ±8.3	26.71 ±13.2	24.29 ±11.9
Char. per token	26.55***	4.45 ±1.1	4.83 ±0.7	4.99 ±0.9
UPOS distribution				
Adjectives	24.08***	6.25 ±7	7.77 ±5.5	10.4 ±8.2
Adpositions	42.5***	8.42 ±6.9	12.13 ±6	13.28 ±5.8

Table 1 continued from previous page

		Mean Values (standard deviation)		
Feature	χ	**Simple Wiki**	**Wikipedia**	**Encyclopedia**
Adverbs	0.7583	3.54 ±5.2	2.57 ±3.5	2.5 ±3.7
Auxiliaries	21.4***	7.45 ±5.8	6.33 ±4.6	4.6 ±4
Coord. Conj.	9.776**	2.64 ±4.1	3.26 ±3.2	2.37 ±3.5
Determiners	3.73	12.48 ±8.2	12.07 ±7	13.68 ±6.8
Interjection	4.59	0 ±0	0 ±0	0. 07 ±0.6
Nouns	5025	25.47 ±9.2	27.56 ±7.5	26.4 ±7.3
Numerals	3998	1.75 ±4.3	1.66 ±4.3	1.97 ±3.5
Particle	17.97***	1.33 ±3	1.58 ±2.6	0.43 ±1.3
Pronouns	0.5129	3.45 ±5.3	2.24 ±3.2	2.36 ±3
Proper nouns	3648	1.62 ±4.8	0.96 ±2.3	1.51 ±3.2
Punctuation	12.22**	15.96 ±13	13.11 ±6.7	12.24 ±7.7
Subord. Conj.	1701	1.12 ±2.8	1 ±2.2	0.72 ±1.9
Symbols	0.8857	0.33 ±1.8	0.42 ±1.9	0.24 ±1.1
Verbs	3386	8.24 ±6.7	7.18 ±4.9	6.95 ±5
X	14.37	0 ±0	0.21 ±1.07	0.34 ±1.30
Lexical density	0.2735	0.56 ±0.2	0.56 ±0.2	0.56 ±0.3
Verb Inflection				
Verb past	23.02***	30.85 ±43.7	54.86 ±46	51.87 ±46.8
Verb pres	7.137*	35.22 ±45.4	23.95 ±38.1	20.93 ±36.6
Verb imp	4467	0.79 ±8.1	0 ±0	2.4 ±14
Verb ind	2015	40.28 ±49.1	36.44 ±48.3	32 ±46.4
Verb fin	2817	30.45 ±40.8	23.43 ±36.3	22.12 ±35
Verb ger	9.962**	6.3 ±19.7	9.34 ±20.4	13.63 ±27.2
Verb inf	9.644**	7.98 ±20.2	12.36 ±22.7	5.89 ±16.7
Verb part	13.07**	26.09 ±39	37.92 ±37.2	39.16 ±41.5
Verb sing,3	1.14	24.5 ±41.6	22.29 ±40.3	20 ±39.1
Aux	0.9443	2.58 ±15.5	3.73 ±17.3	3.6 ±17
Aux pres	6.844*	67.06 ±47	76.78 ±41.7	61.2 ±48.3
Aux ind	7.676*	69.64 ±46.1	80.51 ±39.8	64.8 ±48
Aux fin	4208	72.04 ±42.8	78.11 ±36	64.54 ±45.1
Aux ger	4308	0 ±0	0.28 ±3.1	1.06 ±6.9
Aux inf	8.144*	3.56 ±12.1	6.78 ±15.1	4 ±14.6
Aux part	2557	0 ±0	0.42 ±4.6	0.8 ±6.3
Aux sing,3	547	51.69 ±48.9	52.84 ±46.9	48.4 ±48
Verb Predicate				
Verb heads per sent.	17.16***	1.84 ±1.2	2.74 ±2	2.06 ±1.5
Perc. verbal roots	27.87***	88.69 ±31.8	94.07 ±23.7	71.2 ±45.5
Verb edges avg	3991	1.89 ±1.5	2.15 ±1.3	2.23 ±1.5
Verb edges 0	11.81**	3.82 ±16.6	7.75 ±18.9	3.73 ±13
Verb edges 1	9.014*	6.05 ±20.2	9.37 ±21.6	11.88 ±24.2
Verb edges 2	0.8075	23.36 ±34.5	19.18 ±29.8	19.96 ±32.9
Verb edges 3	2282	20.78 ±34.8	24.32 ±34.7	18.82 ±29.4
Verb edges 4	5133	12.3 ±28.6	17.71 ±30.8	14.2 ±28.4
Verb edges 5	14.92***	2.43 ±14.1	4.07 ±11.2	10.2 ±26.4
Verb edges 6	0.6456	2.08 ±12.8	0.64 ±5.1	2 ±13
Tree Structure				

Table 1 continued from previous page

		Mean Values (standard deviation)		
Feature	χ	**Simple Wiki**	**Wikipedia**	**Encyclopedia**
Depth avg max	57.38***	3.39 ±1.3	4.5 ±1.5	4.66 ±1.7
Tok per clause avg	36.76***	8.26 ±4.9	11.85 ±6.8	11.6 ±8.2
Link len avg	43.7***	2.29 ±0.7	2.76 ±0.6	2.68 ±0.7
Link len max	63.98***	6.95 ±5.1	12.81 ±8.3	11.73 ±7.9
Prep. chain len avg	49.56***	0.61 ±0.7	0.99 ±0.7	1.13 ±0.7
n. prep. chains	72.88***	0.67 ±0.8	1.53 ±1.3	1.63 ±1.1
Prep dist 1	23.83***	43.25 ±48.6	66.1 ±45	69.22 ±39.7
Prep dist 2	15.47***	6.25 ±22.7	9.39 ±25.5	13.26 ±25.6
Prep dist 3	2781	1.39 ±11.2	3.32 ±16.4	2.99 ±14.6
Prep dist 4	4048	0.3 ±3.9	0 ±0	1.33 ±8.6
Prep dist 5	1389	0 ±0	0.85 ±9.2	0.4 ±4.5
Order				
Object pre	0.07394	0.2 ±2.6	0.28 ±3.1	0.8 ±8.9
Object post	2064	37.9 ±48.5	46.33 ±49.9	42.4 ±49.6
Subject pre	13.85***	92.11 ±26.1	97.32 ±14.3	82.64 ±37.6
Subject post	1779	2.53 ±14.2	0.99 ±6.3	0.56 ±4.8
Syntactic Dependencies Distribution				
Adnominal clause	2398	1.12 ±2.8	1.04 ±2.1	1.11 ±2.1
Adverbial clause	1314	1.01 ±2.5	1.07 ±2.2	0.74 ±1.7
Adverbial mod.	0.2475	3.25 ±5	2.54 ±3.3	2.52 ±3.7
Adjectival mod.	34.87***	4.58 ±5.7	6.87 ±5.3	8.99 ±7.3
Appositional mod.	18.57***	0.56 ±2.1	0.7 ±2.1	1.49 ±2.7
Auxiliary	4799	1 ±2.6	1.1 ±2.1	0.71 ±1.8
Auxiliary pass	6.544*	1.93 ±3.7	1.92 ±2.9	2.41 ±3.1
Case marker	40.63***	8.61 ±7.1	12.57 ±6.2	13.41 ±5.8
Cordinating conj.	9.65**	2.6 ±4	3.27 ±3.2	2.41 ±3.5
Clausal compl.	1.42	0.72 ±2.2	0.29 ±1	0.29 ±1.1
Compound	10.92**	1.8 ±3.9	2.46 ±3.7	2.59 ±4.1
Conjunction	11.25**	3.44 ±5.5	4.64 ±4.9	3.52 ±4.7
Copula	37.59***	4.51 ±4.8	3.26 ±3.3	1.49 ±3
Clausal subject	0.2758	0.06 ±0.6	0.04 ±0.4	0.05 ±0.4
Determiners	4265	12.29 ±8.2	11.93 ±7	13.61 ±6.8
Direct object	0.3607	3.45 ±5.2	2.35 ±3.2	2.36 ±3.4
Discourse	756	0.1 ±1	0.12 ±0.8	0.07 ±0.6
Expletive	2777	0.18 ±1	0.08 ±0.6	0.02 ±0.2
Fixed	5868	0.16 ±1	0.28 ±1	0.18 ±0.9
Flat	1767	0.1 ±0.8	0.05 ±0.6	0.14 ±0.8
Goes with	6.898*	0 ±0	0 ±0	0.13 ±1
Marker	8.28*	2.16 ±3.8	2.05 ±3	1 ±2.1
Nominal modifier	24.72***	5.19 ±5.8	7.37 ±6.1	8.53 ±5.7
Nominal subject	68.04***	8.25 ±5.4	5.3 ±3.8	3.46 ±3.9
Nominal subj. pass	4272	1.93 ±3.7	1.86 ±2.9	2.19 ±3.1
Numeral modifier	1827	1.2 ±3.2	1.43 ±3.8	1.25 ±2.7
Oblique compl.	19.41***	2.84 ±4.2	4.47 ±3.9	4.32 ±4.2
Open cl. compl.	4851	0.82 ±2.6	0.89 ±2.3	0.83 ±2.3
Parataxis	15.13***	0.37 ±1.7	0.56 ±1.4	0.07 ±0.4

Table 1 continued from previous page

		Mean Values (standard deviation)		
Feature	χ	**Simple Wiki**	**Wikipedia**	**Encyclopedia**
Punctuation	9.116*	14.84 ±9.4	13.19 ±6.8	12.35 ±7.9
Relative clause	4375	0.74 ±2	0.84 ±1.6	1.04 ±2
Root	85.77***	9.31 ±10.7	4.82 ±2.7	6.01 ±6.1
Subordinate Structure				
Dist. of principal prop.	18.99***	61.91 ±35.4	56.68 ±33	43.87 ±37.4
Dist. of subord. prop.	11.98**	30.35 ±31.8	41.62 ±32.6	42.53 ±37.2
Subordinate post	8.269*	39.53 ±48	56.75 ±46.9	46.6 ±49.4
Subordinate pre	0.8942	11.06 ±29.8	9.35 ±24.1	15 ±34.8
Subordinate chain len avg	8.054*	0.59 ±0.6	0.78 ±0.7	0.73 ±0.7
Subordinate dist 1	6.078*	43.25 ±49.6	57.84 ±47.2	50.53 ±49.8
Subordinate dist 2	2239	6.75 ±24.9	6.07 ±19.1	10.67 ±30
Subordinate dist 3	4232	0.6 ±7.7	1.13 ±6.3	0.4 ±4.5
Subordinate dist 4	4978	0 ±0	1.06 ±9.5	0 ±0

References

Adorni, Giovanni, Chiara Alzetta, Frosina Koceva, Samuele Passalacqua, and Ilaria Torre. 2019. Towards the identification of propaedeutic relations in textbooks. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20*, pages 1–13. Springer.

Agrawal, Rakesh, Behzad Golshan, and Evangelos Papalexakis. 2016. Toward data-driven design of educational courses: A feasibility study. *Journal of Educational Data Mining*, 8(1):1–21.

Alzetta, Chiara, Alessio Miaschi, Felice Dell'Orletta, Frosina Koceva, and Ilaria Torre. 2020. PRELEARN @ EVALITA 2020: Overview of the Prerequisite Relation Learning Task for Italian. In *Proceedings of EVALITA Evaluation of NLP and Speech Tools for Italian, Online, December 17th, 2020*, volume 2765. Accademia University Press.

Alzetta, Chiara, Ilaria Torre, and Frosina Koceva. 2023. Annotation protocol for textbook enrichment with prerequisite knowledge graph. *Technology, Knowledge and Learning*, pages 1–32.

Anderson, Richard C. and P. David Pearson. 2016. A schema-theoretic view of basic processes in reading comprehension. In *Handbook of Reading Research*. Longman, Inc., pages 255–292.

Angel, Jason, Segun Taofeek Aroyehun, and Alexander Gelbukh. 2020. NLP-CIC @ PRELEARN: Mastering prerequisites relations, from handcrafted features to embeddings. In *Proceeedings of EVALITA Evaluation of NLP and Speech Tools for Italian, Online, December 17th, 2020*. CEUR-WS.

Asgarabadi, Yadollah Hosseini, Afsar Rouhi, and Manouchehr Jafarigohar. 2015. Learners' gender, reading comprehension, and reading strategies in descriptive and narrative macro-genres. *Theory and practice in Language Studies*, 5(12):2557.

Bai, Youheng, Yan Zhang, Kui Xiao, Yuanyuan Lou, and Kai Sun. 2021. A bert-based approach for extracting prerequisite relations among wikipedia concepts. *Mathematical Problems in Engineering*, 2021:1–8.

Beckmann, Jens F., Damian P. Birney, and Natassia Goode. 2017. Beyond psychometrics: the difference between difficult problem solving and complex problem solving. *Frontiers in psychology*, 8:1739.

Benjamin, Rebekah George. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:63–88.

Brunato, Dominique, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference, Marseille, France, May 11-16, 2020*, pages 7145–7151.

Brunato, Dominique, Lorenzo de Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? Do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, Brussels, Belgium, October 31 - November 4, 2018*, pages 2690–2699. Association for Computational Linguistics.

Brusilovsky, Peter L. 1992. A framework for intelligent knowledge sequencing and task sequencing. In *Intelligent Tutoring Systems: Second International Conference, ITS'92 Montréal, Canada, June 10–12 1992 Proceedings 2*, pages 499–506. Springer.

Chall, Jeanne S., Sue S. Conard, and Susan Harris-Sharples. 1991. *Should textbooks challenge students? The case for easier or harder textbooks.* Teachers College Press.

Chaplot, Devendra Singh, Yiming Yang, Jaime Carbonell, and Kenneth R. Koedinger. 2016. Data-Driven Automated Induction of Prerequisite Structure Graphs. In *International Educational Data Mining Society, EDM 2016, Raleigh, NC, USA, June 29 - July 2, 2016*.

Chau, Hung, Igor Labutov, Khushboo Thaker, Daqing He, and Peter Brusilovsky. 2021. Automatic Concept Extraction for Domain and Student Modeling in Adaptive Textbooks. *International Journal of Artificial Intelligence in Education*, 31(4):820–846.

Coster, William and David Kauchak. 2011. Simple English Wikipedia: A New Text Simplification Task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, 19-24 June, 2011*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Cunningham, James W. and Heidi Anne Mesmer. 2014. Quantitative measurement of text difficulty: What's the use? *The Elementary School Journal*, 115(2):255–269.

Dale, Edgar and Jeanne S Chall. 1949. Techniques for selecting and writing readable materials. *Elementary English*, 26(5):250–258.

Dang, Quang-Vinh and Claudia-Lavinia Ignat. 2016. Measuring quality of collaboratively edited documents: The case of Wikipedia. In *2016 IEEE 2nd international conference on collaboration and internet computing, CIC, Pittsburgh, PA, USA, November 1-3, 2016*, pages 266–275. IEEE.

de Marneffe, Marie Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

De Medio, Carlo, Fabio Gasparetti, Carla Limongelli, Filippo Sciarrone, and Marco Temperini. 2016. Mining prerequisite relationships among learning objects. *Communications in Computer and Information Science*, 618:221–225.

Den Besten, Matthijs and Jean-Michel Dalle. 2008. Keep it Simple: A Companion for Simple Wikipedia? Keep it Simple: A Companion for Simple Wikipedia? *Industry and Innovation*, 15(2):169–178.

Deutsch, Tovly, Masoud Jasbi, and Stuart M. Shieber. 2020. Linguistic Features for Readability Assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Seattle, WA, USA, July 10, 2020*, pages 1–17. Association for Computational Linguistics.

Fabbri, Alexander R., Irene Li, Prawat Trairatvorakul, Yijiao He, Wei Tai Ting, Robert Tung, Caitlin Westerfield, and Dragomir R. Radev. 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 (Volume 1: Long Papers), ACL 2018, Melbourne, Australia, July 15-20, 2018*, pages 611–620. Association for Computational Linguistics.

Fahim, Mansoor, Hamed Barjesteh, and Reza Vaseghi. 2012. Effects of critical thinking strategy training on male/female efl learners' reading comprehension. *English language teaching*, 5(1):140–145.

Feng, Shi, Sidney D'Mello, and Arthur C. Graesser. 2013. Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin and Review*, 20(3):586–592.

Frantz, Roger S., Laura E. Starr, and Alison L. Bailey. 2015. Syntactic Complexity as an Aspect of Text Complexity. *Educational Researcher*, 44(7):387–393.

Fulcher, Glenn. 1997. Text difficulty and accessibility: Reading formulae and expert judgement. *System*, 25(4):497–513.

Gagne, Robert M. 1962. The acquisition of knowledge. *Psychological review*, 69(4):355.

Gasparetti, Fabio, Carla Limongelli, and Filippo Sciarrone. 2015. Exploiting wikipedia for discovering prerequisite relationships among learning objects. In *2015 International Conference on Information Technology Based Higher Education and Training, ITHET 2015, Lisbon, Portugal, June 11-13, 2015*. Institute of Electrical and Electronics Engineers Inc.

Gordon, Jonathan, Linhong Zhu, Aram Galstyan, Prem Natarajan, and Gully Burns. 2016. Modeling Concept Dependencies in a Scientific Corpus. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 (Volume 2, Long Papers), Berlin, Germany, August 7-12, 2016*, pages 866–875. Association for Computational Linguistics.

Halliday, Michael A. K. 1993. Towards a Language-Based Theory of Learning. *Linguistics and Education*, 5:93–116.

Ide, Nancy and James Pustejovsky. 2017. *Handbook of Linguistic Annotation*. Springer Netherlands.

Jatowt, Adam and Katsumi Tanaka. 2012. Is Wikipedia too difficult? Comparative analysis of readability of Wikipedia, simple Wikipedia and Britannica. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12), Maui, HI, USA, October 29 - November 2, 2012*, pages 2607–2610.

Johnson-Laird, Philip N. 1980. Mental models in cognitive science. *Cognitive science*, 4(1):71–115.

Jolliffe, Ian T and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Liang, Chen, Shuting Wang, Zhaohui Wu, Kyle Williams, Bart Pursel, Benjamin Brautigam, Sherwyn Saul, Hannah Williams, Kyle Bowen, and C. Lee Gile. 2016. BBookX: Building online open books for personalized learning. In *30th AAAI Conference on Artificial Intelligence (AAAI 2016), Phoenix, AZ, USA, February 12-17, 2016*, volume 30 (1), pages 4369–4370.

Liang, Chen, Jianbo Ye, Han Zhao, Bart Pursel, and C Lee Giles. 2019. Active learning of strict partial orders: A case study on concept prerequisite relations. In *12th International Conference on Educational Data Mining (EDM 2019), Montréal, Canada, July 2–5, 2019*, pages 348–353. International Educational Data Mining Society.

Lu, Weiming, Pengkun Ma, Jiale Yu, Yangfan Zhou, and Baogang Wei. 2019. Metro maps for efficient knowledge learning by summarizing massive electronic textbooks. *International Journal on Document Analysis and Recognition*, 22:99–111.

Ma, Haiping, Jinwei Zhu, Shangshang Yang, Qi Liu, Haifeng Zhang, Xingyi Zhang, Yunbo Cao, and Xuemin Zhao. 2022. A Prerequisite Attention Model for Knowledge Proficiency Diagnosis of Students. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM22), Atlanta, GA, USA, October 17-21, 2022*, pages 4304–4308.

Manrique, Ruben, Juan Sosa, Olga Marino, Bernardo Pereira Nunes, and Nicolas Cardozo. 2018. Investigating learning resources precedence relations via concept prerequisite learning. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI18), Santiago, Chile, December 3–6, 2018*, pages 198–205. IEEE.

McNamara, Danielle S., Arthur C. Graesser, and LouwerseMax. 2012. Sources of text difficulty: Across genres and grades. In *Measuring Up: Advances in How We Assess Reading Ability*. Rowman & Littlefield Education, chapter 6, pages 89–116.

Mesmer, Heidi Anne, James W. Cunningham, and Elfrieda H. Hiebert. 2012. Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading research quarterly*, 47(3):235–258.

Miaschi, Alessio, Chiara Alzetta, Franco Alberto Cardillo, and Felice Dell'Orletta. 2019. Linguistically-driven strategy for concept prerequisites learning on Italian. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications, BEA 2019, Florence, Italy, August 2, 2019*, pages 285–295. Association for Computational Linguistics.

Napoles, Courtney and Mark Dredze. 2010. Learning Simple Wikipedia: A Cogitation in Ascertaining Abecedarian Language. In *Proceedings of HLT/NAACL Workshop on Computation Linguistics and Writing, Los Angeles, CA, USA, June 6, 2010*, pages 42–50.

Pan, Liangming, Chengjiang Li, Juanzi Li, and Jie Tang. 2017. Prerequisite Relation Learning for Concepts in MOOCs. In *55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Volume 1 - Long Papers), Vancouver, Canada, July 30-August 4, 2017*, pages 1447–1456. Association for Computational Linguistics.

Pelánek, Radek, Tomáš Effenberger, and Jaroslav Čechák. 2022. Complexity and difficulty of items in learning systems. *International Journal of Artificial Intelligence in Education*, 32(1):196–232.

Roy, Sudeshna, Meghana Madhyastha, Sheril Lawrence, and Vaibhav Rajan. 2019. Inferring Concept Prerequisite Relations from Online Educational Resources. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI19), Honolulu, HI, USA, January 27 – February 1, 2019*, volume 33, pages 9589–9594. AAAI Press.

Sabnis, Varun, Kumar Abhinav, Venkatesh Subramanian, Alpana Dubey, and Padmaraj Bhat. 2021. UPreG: An Unsupervised Approach for Building the Concept Prerequisite Graph. *International Educational Data Mining Society*.

Samoilenko, Anna, Florian Lemmerich, Maria Zens, Mohsen Jadidi, Mathieu Génois, and Markus Strohmaier. 2018. (Don't) mention the war: A comparison of Wikipedia and Britannica

articles on national histories. In *Proceedings of the 2018 world wide web conference, Lyon, France, April 23-27, 2018*, pages 843–852.

Sayyadiharikandeh, Mohsen, José Luis Ambite, Jonathan Gordon, and Kristina Lerman. 2019. Finding prerequisite relations using the Wikipedia clickstream. In *Companion Proceedings of The Web Conference 2019 - Companion of the World Wide Web Conference (WWW 2019), San Francisco, CA, USA, May 13-17, 2019*, pages 1240–1247. Association for Computing Machinery, Inc.

Shen, Aili, Jianzhong Qi, and Timothy Baldwin. 2017. A hybrid model for quality assessment of Wikipedia articles. In *Proceedings of the Australasian Language Technology Association Workshop 2017 (ALTA 17), Brisbane, Australia, December 6–8, 2017*, pages 43–52.

Snow, Catherine E. 2010. Academic language and the challenge of reading for learning about science. *Science*, 328(5977):450–452.

Spencer, Mercedes, Allison F Gilmour, Amanda C Miller, Angela M Emerson, Neena M Saha, and Laurie E Cutting. 2019. Understanding the influence of text complexity and question type on reading outcomes. *Reading and writing*, 32:603–637.

Stamper, John, Bharat Gaind, Karun Thankachan, Huy Nguyen, and Steven Moore. 2023. Hierarchical concept map generation from course data. In *AAAI 2023 Workshop on Artificial Intelligence in Education (AI4Edu), Washington DC, USA, February 13, 2023*.

Talukdar, Partha and William Cohen. 2012. Crowdsourced Comprehension: Predicting Prerequisite Structure in Wikipedia. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP (BEA 2012), Montreal, Canada, June 7, 2012*, pages 307–315.

Vajjala, Sowmya and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL14), Gothenburg, Sweden, April 26-30, 2014*, pages 288–297.

van Halteren, Hans. 2000. The detection of inconsistency in manually tagged text. In *Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora (LINC-2000), Luxembourg, August 6, 2000*, pages 48–55.

Wallot, Sebastian, Beth A. O'Brien, Anna Haussmann, Heidi Kloos, and Marlene S. Lyby. 2014. The role of reading time complexity and reading speed in text comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, 40(6):1745–1765.

Wang, Shuting, Alexander G. Ororbia, Zhaohui Wu, Kyle Williams, Chen Liang, Bart Pursel, and C. Lee Giles. 2016. Using prerequisites to extract concept maps from textbooks. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM 2016), Indianapolis, IN, USA, October 24-28, 2016*, pages 317–326. Association for Computing Machinery.

Weber, Rose-Marie. 1991. Linguistic diversity and reading in American society. In *Handbook of reading research, Volume 2*. Routledge Handbooks Online, mar, pages 97–119.

Zhao, Zhongying, Yonghao Yang, Chao Li, and Liqiang Nie. 2021. GuessUNeed: Recommending Courses via Neural Attention Network and Course Prerequisite Relation Embeddings. *ACM Transactions on Multimedia Computing, Communications and Applications*, 16(4).

Zhou, Yang and Kui Xiao. 2019. Extracting Prerequisite Relations among Concepts in Wikipedia. In *Proceedings of the International Joint Conference on Neural Networks, Budapest, Hungary, July 14-19, 2019*. Institute of Electrical and Electronics Engineers Inc.

Zhu, Yaxin and Hamed Zamani. 2022. Predicting prerequisite relations for unseen concepts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), Abu Dhabi, December 7–11, 2022*, pages 8542–8548.