# From Texts to Prerequisites

*Identifying and Annotating Propaedeutic Relations in Educational Textual Resources*

**PhD in Digital Humanities**

Cycle XXXIII

UNIVERSITÁ DEGLI STUDI DI GENOVA

Candidate: Chiara Alzetta

Supervisors: Prof.Ilaria Torre, Prof.Simonetta Montemagni

2021

# Abstract

Prerequisite Relations (PRs) are dependency relations established between two distinct concepts expressing which piece(s) of information a student has to learn first in order to understand a certain target concept. Such relations are one of the most fundamental in Education, playing a crucial role not only for what concerns new knowledge acquisition, but also in the novel applications of Artificial Intelligence to distant and e-learning. Indeed, resources annotated with such information could be used to develop automatic systems able to acquire and organise the knowledge embodied in educational resources, possibly fostering educational applications personalized, e.g., on students' needs and prior knowledge.

The present thesis discusses the issues and challenges of identifying PRs in educational textual materials with the purpose of building a shared understanding of the relation among the research community. To this aim, we present a methodology for dealing with prerequisite relations as established in educational textual resources which aims at providing a systematic approach for uncovering PRs in textual materials, both when manually annotating and automatically extracting the PRs. The fundamental principles of our methodology guided the development of a novel framework for PR identification which comprises three components, each tackling a different task: *(i)* an annotation protocol (PREAP), reporting the set of guidelines and recommendations for building PR-annotated resources; *(ii)* an annotation tool (PRET), supporting the creation of manually annotated datasets reflecting the principles of PREAP; *(iii)* an automatic PR learning method based on machine learning (PREL). The main novelty of our methodology and framework lies in the fact that we propose to uncover PRs from textual resources relying solely on the content of the instructional material: differently from other works, rather than creating de-contextualised PRs, we acknowledge the presence of a PR between two concepts only if emerging from the way they are presented in the text. By doing so, we anchor relations to the text while modelling the knowledge structure entailed in the resource.

As an original contribution of this work, we explore whether linguistic complexity of the text influences the task of manual identification of PRs. To this aim, we investigate the interplay between text and content in educational texts through a crowd-sourcing experiment on concept sequencing. Our methodology values the content of educational materials as it incorporates the evidence acquired from such investigation which suggests that PR recognition is highly influenced by the way in which concepts are introduced in the resource and by the complexity of the texts.

The thesis reports a case study dealing with every component of the PR framework which produced a novel manually-labelled PR-annotated dataset.

# ACKNOWLEDGEMENTS

As I am approaching the end of my PhD, I look back and feel lucky for the great deal of support and assistance I have received throughout my journey. It wasn't always easy, but there are many people I would like to express my gratitude to for helping me cut the finish line.

First, I would like to thank my supervisors, Professors Ilaria Torre and Simonetta Montemagni, whose expertise was invaluable in finding my way through the challenges of the research and for providing me with the tools for carrying on my work also throughout difficult times. Your complementary perspectives on the research were possibly the most valuable contribution to my work: you thought me that there are always many different answers to a question.

I would also like to acknowledge my thesis reviewers, Petya Osenova and Alessandro Lenci, for their valuable comments, thought-provoking observations and for pushing me to take my work to a higher level.

I can't help but thank all the members of the ItaliaNLP Lab (Institute of Computational Linguistics - CNR, Pisa) and Tel-DH Research Programme (DIBRIS, University of Genoa) for their patient support, precious advice and for making working together so fun and stimulating. I would particularly like to single out Giulia Venturi, Frosina Koceva and Felice Dell'Orletta: most of my research wouldn't have been possible without your contributions and ideas.

In addition, I must thank my family, who never lost the curiosity for my research (though they still get confused about it sometimes) and kept being supportive at all times. We passed through a lot of tough moments, but you were always there for me nonetheless. To those who are here and to those who went away, thank you for what you've given me.

Finally, I could not have completed this journey without the support of friends, who offered stimulating discussions as well as happy distractions to rest my mind outside of my research. In particular, I would like to thank Alessio, Ilenia, Lorenzo and Samuele, for being such wonderful and fun team-mates; Dominique, for your eye-opening comments during our lunch breaks; the guys of the 33° cycle, for surviving together the PhD adventure; Beatrice, Beatrice, Elena, Emanuela, Filippo, Ginevra, Giulia, Leonardo, Mimma, Simona and to all the others for the laughs and for making time pass so fast when we are together.

## Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

# Table of Contents

# List of Tables

# List of Figures

# PREAMBLE

The content of this dissertation, the challenges addressed and findings of the research, are the result of a collaborative effort of the Technology-Enhanced Learning & Digital Humanities (TelDH) Research Programme[1] (Department of Computer Science and Technology, Bioengineering, Robotics and Systems Engineering, University of Genoa) and other research teams, namely researchers from the ItaliaNLP Lab[2] at the Institute for Computational Linguistics "A. Zampolli" (National Research Council, Italy), and from P. Brusilovsky PAWS Lab (School of Computing and Information, Pittsburgh University)[3]. The collaboration between multiple groups resulted in a multidisciplinary teamwork where the perspectives of Computational Linguistics, Education and Computer Engineering all contributed to provide insights and expertise that greatly assisted the research here presented.

Our interest toward the topics of this research, namely propaedeutic relations and instructional content modelling, has a long history. It emerges from an ongoing research on the wide area of Information and Communication Technologies integration with formal and informal learning carried on by the TelDH Research Programme. The ultimate goal of the research is to aid and innovate the management and access to cultural content by exploiting new technologies as essential components and enablers of the learning processes. The first outcome of such research line was "ENCODE - ENvironment for COntent Design and Editing"[4], a tool to assists teachers in the design of lessons and/or learning paths. One of the limits of ENCODE consisted in its use of Educational Concept Maps (ECMs) as a way to formalise the knowledge structure of a subject matter by means of graphs representing key concepts and relations between them. Despite being a quite intuitive way of representing knowledge (as a graph of concepts and relations), manually building ECMs is time consuming and requires expert's knowledge. To overcome this limit, we directed our efforts towards the identification of propaedeutic relations between concepts. Broadly speaking, as ECMs should reflect the most effective organisation of concepts, we need to find a way to uncover domain concepts and their learning order from instructional materials. Started from a practical need (i.e., integrate in ENCODE a module for automatic ECM creation), our research on propaedeutic relations has took over our interests. Indeed, uncovering propaedeutic

[1] http://teldh.dibris.unige.it/

[2] http://www.italianlp.it/

[3] http://www.pitt.edu/~paws/

[4] See [3] and http://teldh.dibris.unige.it/encode/

relations from learning materials is still an open research problem, and it is exactly the issue that will be addressed in this dissertation.

The role of the author of this thesis in the research was to bridge the gap between the different fields involved by proposing and developing solutions that would incorporate the different perspectives to achieve the final goal. Although it wasn't always easy and in some cases we had to make compromises, the results we achieved so far are extremely valuable – in the author's view – as our multidisciplinary work constitute the solid foundation for fostering further research on the topic of educational materials modelling. Above all, it creates a common ground for researchers and scholars with different background to carry on the research on propaedeutic relations while mutually benefit each other with their findings.

## Published Papers

Parts of this dissertation are based on (or might refer to) the following publications.

- Alzetta, C., Miaschi, A., Dell'Orletta, F., Koceva, F. and Torre, I., 2020. Prelearn@ EVALITA 2020: Overview of the prerequisite relation learning task for italian. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop* (EVALITA 2020), Online. ISSN: 16130073
- Alzetta, C., Galluccio, I., Koceva, F., Passalacqua, S. and Torre, I., Digging Into Prerequisite Annotation. In *Proceedings of the Second Workshop on Intelligent Textbooks*, 06 luglio 2020, Online.
- Alzetta, C., Miaschi, A., Adorni, G., Dell'Orletta, F., Koceva, F., Passalacqua, S., and Torre, I. (2019) Prerequisite or Not Prerequisite? That's the problem! An NLP-based Approach for Concept Prerequisite Learning. In *Proceedings of 6th Italian Conference on Computational Linguistics* (CLiC-it), 13-15 novembre, 2019, Bari, Italia.
- Miaschi, A., Alzetta, C., Cardillo, F.A., Dell'Orletta, F. (2019) Linguistically-Driven Strategy for Concept Prerequisites Learning on Italian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (BEA), 2 agosto 2019, Firenze, Italia. ISBN 978-1-950737-34-5.
- Passalacqua, S., Koceva, F., Alzetta, C., Torre, I. and Adorni, G., 2019. Visualisation Analysis for Exploring Prerequisite Relations in Textbooks. In *iTextbooks@AIED*, pp. 18-29.
- Adorni, G., Alzetta, C., Koceva, F., Passalacqua, S., Torre, I. (2019) Towards the Identification of Propaedeutic Relations in Textbooks. In *Proceedings of the 20th International Conference on Artificial Intelligence in Education* (AIED), 25-29 giugno 2019, Chicago, USA. ISSN: 0302-9743 EISSN: 1611-3349 ISBN: 978-3-030-23203-0.
- Alzetta, C., Koceva, F., Passalacqua, S., Torre, I., Adorni, G. (2018) PRET: Prerequisite - Enriched Terminology. A Case Study on Educational Texts. In *Proceedings of 5th Italian*

*Conference on Computational Linguistics* (CLiC-it), 10-12 dicembre 2018, Torino, Italia. ISBN: 9788831978682.

- Alzetta, C., Adorni, G., Celik, I., Koceva, F., Torre, I. (2018) Toward a User-Adapted Question/Answering Educational Approach. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (UMAP2018) pp. 173-177, ACM, 8-11 luglio, 2018, Singapore, Singapore. ISBN: 978-1-4503-5784-5.

CHAPTER 1

# INTRODUCTION

Textual educational materials such as textbooks are traditionally meant to provide students with knowledge about a certain topic or subject matter. Currently, they are so largely available online that navigating them without the guidance of a domain expert might be overwhelming for a learner. A line of research at the crossroads between Education, Information Extraction and, more in general, Artificial Intelligence applied to Education deals with the development of automatic strategies to support students in their process of autonomously acquiring knowledge, possibly to be integrated into educational technologies. Such systems should be able to recover not only the most appropriate content based on students' requests, but also present it a way that prevents student misunderstandings and confusion. To this aim, being able to model the content of instructional materials is a fundamental task in order to acquire the knowledge structure underlying the resource. The aim of this work is to advance the research in educational content modelling by adopting an interdisciplinary perspective. More specifically, we draw on knowledge from the fields of Education, Information Extraction and Natural Language Processing to address the problem of uncovering the knowledge structure of educational texts with the goal of identifying the propaedeutic relationships between concepts in educational materials.
This introductory Chapter is meant to delineate the context of the research, the goals and research questions of our research, highlight its main contributions, and provide an overview of the thesis structure.

## 1.1 Overview of the Research Context

It is now a common practice to use technologies to support learning activities. Early examples of the collaboration between Education and new technologies appeared in the past century, but it's

only from the beginning of the current millennium that computers really began to infiltrate the education process at a wider scale, and it's clear that the process is still ongoing. Indeed, as the Web started housing massive quantities of educational materials and user-generated content, educational technologies became, on the one hand, more advanced in order to keep up with novel needs and, on the other hand, vital for organising and delivering such contents to the growing audience of learners. Indeed, while in the past learners were only (or, at least, mostly) young students in school classrooms, guided in their learning process by a certified teacher, nowadays a learner might be anyone who wants to acquire new knowledge, either in a traditional classroom environment or from home using the novel opportunities of autonomous learning. It goes without saying that different populations of learners might have different needs, that should be addressed in order to achieve the final goal of knowledge acquisition.

Recently there is a large discussion among the research community on the role of Artificial Intelligence (AI) in the process of learning and how AI can support educators in delivering individual and differentiated content to students. A recent survey about AI in Education reported that more than 60% of the published papers on AI applied to Education address academic and learning support at some level [293]. The ideal goal of such facilities would be not only to support knowledge acquisition, but also to enable learners to build better communication and problem solving abilities. Indeed, on the heels of the industrial revolution, many hope for the era of Education 4.0, where education is expected to prepare student for the novel world characterised by smart technology, artificial intelligence, and robotics, all of which already impact our everyday lives. In order to achieve this goal, it is not sufficient to simply present new trends to students in a traditional manner: Education itself must evolve in order to incorporate the novel tools and approaches [228]. Unfortunately, traditional learning at school seemed reluctant to go along with this change, remaining anchored to the classical lecture model of teaching. Such model mainly focuses on cognitive objectives and the main emphasis of this strategy is the presentation of content through lectures. However, for decades, there has been evidence that actively engaging students in the learning process produces better educational outcomes at all levels [227]. In the past decade, such idea was incorporated, at least in principle, into other forms of learning experiences, possibly alternative to in-class education, mostly located on the web. We refer to them as life-long and distant learning. Both address a different audience than traditional in-class education as they generally target adults which need to update their knowledge incorporating specific unknown competences and skills in a fast and effective way. In order to target such population of learners, Learning Management System (LMS) and Massive Open Online Courses (MOOCs) started spreading [217].

MOOCs really took off around 2010, when university courses started to move online to reach the general public. Andrew Ng, teaching machine learning at Stanford University, reports that the number of students enrolled in his class boosted from 400 to 100,000 when he moved the course online. Around the same time, also professors Sebastian Thrun and Peter Norvig offered

online the course of "Introduction to Artificial Intelligence" which saw approximately 1,600,000 students participating from 190 countries. Thanks to these events, models for online knowledge sharing started spreading: among them we could mention Udacity, edX and Cursera, just to name a few, still available today. Their underlying philosophy consists of providing learning content accessible from home, thus targeting a population of adults. Usually, LMS and MOOC platforms are employed for boosting knowledge about specific topics rather than attending complete university programs: although some universities provide a certificate once you complete a MOOC, this is frequently not an option, so these platforms just serve to expand the knowledge rather than providing a formal qualification. However, they were applauded as inclusive systems to provide free access to high quality learning materials for students coming from different parts of the world, with limited time or possibilities, eventually supporting the establishment of education as a fundamental human right accessible for all [217]. The increasing use of mobile devices and web applications observed during the course of the last decade made learning even more ubiquitous, allowing LMS and MOOC users to access educational content anytime and everywhere.

The great excitement around e-learning and distant learning was followed by a certain degree of disappointment over the years, at least to some extent. The potentiality of MOOC platforms were not always exploited at their best: deep interactions was only rarely deployed as most teachers simply uploaded lecture videos online where they talked to a camera delivering content. Instead of exploiting the new learning environment to design a novel way of teaching, most teachers simply used MOOCs as a container where storing digitised versions of their lectures. As a consequence, the teaching experience were deemed as cold and impersonal, resulting in low engagement and completion rates [235].

Eventually, although the research on the integration of novel technologies in education still continued [293], in the reality we kept observing an opposition between traditional in-class learning happening in educational institutions and informal learning, characterized by a low degree of planning. Informal learning refers to the process of acquiring knowledge from activities which are not undertaken with a learning purpose in mind and that take place outside of any organisational framework. An example of such king of learning might be what we acquire as new information by making a Web search. Clearly, such learning process saw a boost in recent years, fostered by the growing amount of online materials. However, the latter are frequently unorganised, accessed by individual learners as needed, with no actual supervision by an expert educator which could, on the other hand, guide the access to learning content in a way that maximises the learning experience. As a consequence, until recently, informal and distant learning were almost exclusively an activity for adults that wanted to update their knowledge about a certain already-known topic, while young learners remained anchored to more traditional practices of formal learning.

Such scenario drastically changed last year, when education at all levels was impacted by

the COVID-19 pandemic. The effect of the pandemic were so widespread over the globe and they changed our everyday lives so deeply that we can't ignore them when talking about education today. Since social distancing was adopted worldwide as first reaction to the spread of the virus, schools and universities were closed[1]. As a consequence, Education, together with the whole society, experienced a leap towards virtualization. Remote learning thus became a common practice and experience for millions of students regardless of their age, and we can expect that this will impact on future education and learning models. At this point in history, online learning is not an option any more, but it is something we have to do in order to keep acquiring knowledge in face of the new situation. Many wonder about the social and economical impact of school closures and, one year apart from the outbreak of COVID, the discussion in each country, as well as globally, led by organisations such as UNESCO[2], is very vivid.

However, the flip-side of such situation is the great excitement observed among the researchers and professionals working in the field of educational technologies and distant learning. What is most interesting from their point of view is investigating the opportunities and challenges opened by such new scenario. This is the context where the present thesis aims to offer its contribution.

**Challenges and Opportunities in the Novel Educational Setting** The crisis prompted by the pandemic has stimulated innovation within the education sector. Distance learning solutions were adopted to support education continuity and what first were only challenges associated only with MOOCs environments have become common to all teachers all over the globe. As we mentioned, the greatest challenge faced by distant learning through MOOCs was possibly the lack of engagement with students. Building a sense of belonging and community is what was mostly missed in e-learning with respect to in-class education, and this is possibly even more important when students can't meet in person. The sense of abandonment and loneliness experienced due to the social distancing deeply impacted mental health, also in young learners, as never before [74]. As a consequence, those who develop solutions to support learning activities must account for such issues and find solutions to overcome them. Demanding this job to teachers is impossible as it requires a one-to-one tutoring experience which is unfeasible, especially in online courses where the number of students is usually high. Technology might come to rescue and help teachers and professional instructors provide customised learning experiences by automatically selecting the most appropriate content for each student and adapt it to their needs. This might be helpful for all educational experiences. In school and university education, it might help students fill their knowledge gaps more effectively as a computer doesn't get tired in presenting the same content over and over until the student demonstrate to have acquired it.

In life-long and distant learning, a learner might want to tailor the learning process on its own needs, focusing on specific topics rather than on long learning sessions. Indeed, a MOOC

[1] UNESCO provides some very insightful interactive maps to monitor the global situation about school closures from the beginning of the pandemic at `https://en.unesco.org/covid19/educationresponse`.

[2] See the UNESCO Report 'Education during COVID-19 and beyond', August 2020

course generally tackles many concepts, trying to offer a view that is complete as much as possible on a subject. Sometimes it may happen that the user is not interested in the entire course but only on sections addressing a specific issue, so viewing all the contents of the course is not an efficient way to obtain her/his goal. Personalising the learning process is indeed one of the most exciting challenges of Education 4.0: it implies to face many novel issues pertaining to the role of teachers, the creation of instructional resources and the adaptation to students profile. Most of these issues are still open, although we see a great interest towards addressing them using new technologies.

It should be noted that due to the pandemic, next to challenges, we were also given opportunities. First of all, new learning approaches, such as blended learning, which were first applied in small and controlled environment mainly for research purposes, are now actually tried and tested by large groups of students. This might foster a deep renovation in the educational setting that might last beyond COVID [133]. Furthermore, and most importantly from our point of view, switching to online education has fostered the creation of a large amount of educational materials, in many different forms, freely available on the web [74]. Research in all fields, but especially in the wide area of AI, is eager of large amounts of resources: they provide evidence about the phenomena in the world and can be leveraged to develop theoretical models as well as educational applications to support learning. Since online resources are going to become a primary source of education, it becomes, accordingly, of paramount importance to take full advantage of them, so to enhance the learning experience and its effectiveness. However, since they were created in response of an emergency situation, these educational resources are frequently released on the web not as part of structured and well-designed learning plans but rather as unorganised, publicly available materials. As a consequence, the Web started turning into a rich but also chaotic educational environment in need of adaptive systems to support meaningful and effective navigation of its content [149]. Providing new services that allow to exploit this extraordinary asset by guiding students through their navigation would be a great opportunity to get the most out of a difficult situation.

Approaches based on Information Extraction and Natural Language Processing (NLP) could be fruitfully exploited in such scenarios in order to acquire the content of such large amounts of unstructured materials. Indeed, the opportunities of Language Technologies in the context of distance and e-learning, in particular for what concerns the development of educational applications and content acquisition from textual materials, have been acknowledged for many years [50]. As proof of such interest, it is worth mentioning the well-established ACL workshop on Innovative Use of NLP for Building Educational Applications (BEA)[3] or, among European projects, the Language Technologies for LifeLong Learning (LTfLL) [196], carried on between 2008 and 2011. Since 2003, BEA workshop represents an occasion for the NLP community to gather and discuss novel opportunities and challenges of NLP in educational applications. The

[3] https://www.aclweb.org/anthology/venues/bea/

LTfLL project, fostered by the large amounts of educational materials available online, specifically promoted the integration between Language Technology and Semantic Web to enhance e-learning with applications for education and training [197].

Language technologies indeed can be exploited to explore the content of textual materials, and this is extremely important from the perspective of educational technologies, as the content of educational materials should be organised in a way that allows learners to understand them [102]. For example, fundamental notions that are needed in order to understand the topic of the resource are generally mentioned at the beginning. This is usually done to introduce the reader into the subject matter by mentioning already-known familiar concepts, but the resource author might chose to not discuss them altogether and propose them as prerequisite of the remaining content. Being able to automatically retrieve such knowledge structure from educational materials could allow to evaluate the quality of an instructional resource, or connect them and integrate the content of multiple resources in order to dynamically build learning plans based on the learner needs. Next to this type of resources, on the web we also see an increasing amount of more traditional learning materials, such as textbooks, in their digitised versions.

Textbooks and learning materials designed by domain experts and experienced teachers still remain, also in the digital era, one of the most reliable and effective resources for acquiring new knowledge [54]. One of the main reasons for that is that the content of such materials is usually organised to represent a knowledge structure that guides readers-learners to the acquisition of the knowledge contained in the resource. Think, for example, of any high school textbook on Algebra. Ideally, the book will be organised in chapters and sections; the first chapters will deal with the most fundamental concepts, such as numbers and operations with numbers, whereas the last chapters will be dedicated to more advanced notions, such as functions and equations. Additionally, textbooks are generally of extremely high quality and explicitly target a specific population of learners with a certain level of expertise in the domain. However, the wide domain coverage of textbooks, if on the one hand provides highly valuable and complete information about the subject matter, on the other hand it could discourage a learner interested in a specific topic: revising the whole content of the resource might not be an efficient choice if the student only needs information pertaining to specific concept. If we were able to uncover knowledge structure of a textbook, we might provide the reader with the specific portion of textbook the student needs, while simultaneously also identifying gaps in the explanation that could be filled with extra materials in order to provide a richer discussion or learning aids to augment students' reading experience through automatically generated insights [39]. However, automatically capturing the knowledge structure from textual materials, either small knowledge pills or textbooks, is not a straightforward task which requires to deal with the identification of the concepts mentioned within the resource and also identifying the relations passing between them.

With such scenario on the background, the work presented in this dissertation aims to contribute to the faceted and wide area of educational technology by addressing, in particular,

the issues related to modelling the content of textual educational resources in order to obtain a text-bound representation of the instructional content. As it will further detailed along the thesis, *our goal is to define a general methodology for uncovering the propaedeutic relationships between the concepts mentioned in textual educational materials*. Such task is highly relevant to the scenario depicted above: it lays the foundations to assist the research on personalisation of the learning experience by offering necessary resources to surpass the "one size fits all" philosophy.

## 1.2 Knowledge Structures and Prerequisite Relations

Acquiring the knowledge structure of textual educational materials is a challenging, although extremely powerful, task. The goal of this task is formally representing the relationships between concepts by modelling the content of the resource under investigation. We can informally define concepts as small pieces of knowledge that we might represent by means of, e.g., keywords. In order to deliver knowledge to a student, a resource must present concepts in such a way that supports learners' understanding of the subject domain, and meanwhile avoids student's frustration, misunderstanding and disorientation [101], as we discussed above with reference to the Algebra textbook. Such basic principle must be preserved in every learning situation as it is fundamental to deliver knowledge to a student, and learning can't happen otherwise.

To better understand this idea, we might recall an experience which might be familiar to most of us: schematising the content of a textbook or lecture while studying in order to obtain a diagrammatic representation of the resource content. Those diagrams usually included domain concepts, i.e. the most fundamental ideas presented by the teacher or textbook author, and relationships between concepts. Exploiting again the example on Algebra depicted above, a diagram structuring the knowledge related to operations with numbers might include, for example, the concepts '*multiplication*' and '*exponentiation*', related to each other since '*exponentiation is a repeated multiplication*'. Knowledge structures very much resemble our diagrams as they too represent resource content by means of concepts and relations. But what do these relations represent? Do they have a specific meaning or do they refer to a generic relatedness? In truth, they can assume different meaning depending on the information we want to represent. For example, they might refer to taxonomic relations as well as semantic relations. With respect to taxonomic relations, we might want to express the fact that *"Algebra is a broad area of Mathematics"* or that *"Renaissance was a period of European history"*. In these cases, we might want to mark a relationship between the concepts "Algebra–Mathematics" and "Renaissance–European history", where the first mentioned concept is part of the area referred to by the other concept. Semantic relations, on the other hand, embrace a wide range of possibilities. Some example can be, just to name a few, "is caused by, "is based on", "is a property of", "is an instrument for", "is a material for". As we can imagine, many of these relations can only exist between certain concepts, thus they are often subject to restrictions depending on the domain. Among dependency relations in

the educational setting, however, the most relevant, possibly also overlapping with some of those mentioned above, are prerequisite relations.

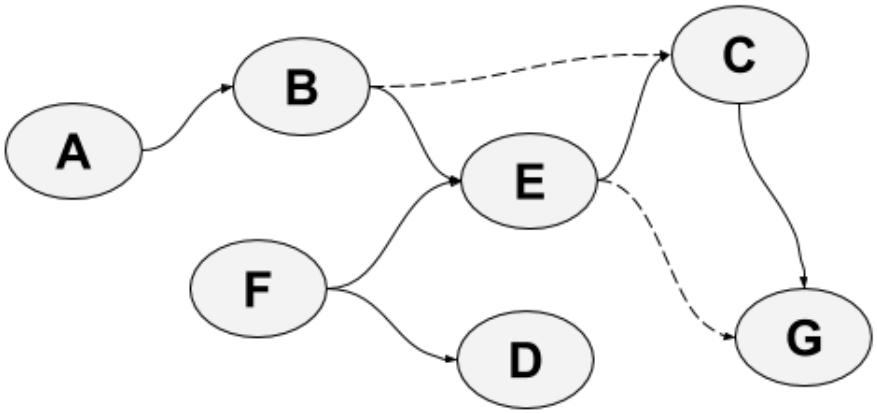



Figure 1.1: Sketch of a simple knowledge structure representing concepts as nodes and their prerequisite relations as edges.

A ***prerequisite relation (PR)*** is generally defined as a *binary dependency relation connecting a prerequisite and a target concept where the first has to be known in order to understand the second* [167]. In other words, PRs convey meaning about which knowledge is needed to understand and learn new knowledge. As a whole, PRs enable to identify which learning paths are most effective in order to obtain a full understanding of concepts. They are so relevant for knowledge structures that sometimes PRs are the only represented relation [46]. In order to understand how prerequisite relations can be represented through knowledge structures, consider Figure 1.1. The image depicts an example of a simple concept map [210] which formally represents a domain's knowledge. The graph makes explicit the prerequisite structure of the domain be means of prerequisite relations (represented as edges) between concepts (represented as nodes labelled with capital letters). The edges in the graph, e.g., between concept $A$ and $B$, shall be read as $A$ is prerequisite to $B$ ($A \prec B$). In order to make the example more concrete, we might recall the previous example on *'multiplication'* and *'exponentiation'*. The relationship between the two concepts can be easily represented in the graph: *'multiplication'* and *'exponentiation'* could be represented by nodes 'E' and 'C' respectively. We could go further and add another concept, 'addition', which we could associate to node 'B'. As we all know, multiplication is nothing more than a repeated addition, which is in fact represented by the edge connecting 'B' and 'E'. Interestingly, the map also depicts a dashed edge between 'B' (addition) and 'E' (exponentiation). This is an interesting property peculiar to prerequisite relation and learning paths, which we refer to as *transitivity*: the knowledge structure in the figure says that, if a student masters the concept of *'addition'*, (s)he already has most of the knowledge required to understand *'exponentiation'*. The PR property of transitivity, given its relevance to PRs, will be further discussed in the remainder of the thesis

## 1.3 Motivations

The relevance of knowledge structures representing PRs has been shown in multiple educational scenarios to the point that automatic identification of prerequisite relationships between concepts has been identified as one of the key requirements for modern, large-scale online education [103, 181, 264]. Technology-enhanced learning systems make use of such representations to incorporate knowledge about domains and enhance systems functionalities, such as automatic synthesis of study plans [4, 104, 299], reading list generation [108, 145], and automatic educational content creation [175]. Conversely, knowing the knowledge structure of a domain could allow to locate students' competence level (i.e., what they already know about the subject matter) with respect to the knowledge structure [212] and support them with personalised recommendations. The evidence acquired so far by existing research seems to confirm that prerequisite relations can support several adaptation and user modeling techniques [46]. Given their importance, we observe the interest of the NLP and AI communities towards automatically acquiring the knowledge structure from textual resources. Broadly speaking, such task can be broken down into two main sub-problems: *(1)* identifying the knowledge component discussed within the resource and *(2)* uncovering the possible relations passing between them. Manually enriching every educational resource uploaded on the web with PR would be extremely valuable, although also unfeasible since the manual annotation of PR in texts is a long and time-consuming process that can be done only by domain experts. As a response to such issue, the research has started dealing with the definition of method for automatically acquiring PRs.

The task of relationship extraction is a well-known task of Natural Language Processing and Information Extraction. Its main goal is to identify relations between entities in a document [27] in order to give a structured representation of the information conveyed by the text. However, uncovering educational relations from instructional materials requires to address specific issues due, e.g., to the peculiarities of the domain tackled by the resource of the final use of the acquired knowledge structure. Although few approaches for automatically acquiring prerequisite relations from educational materials without resorting to labelled datasets exist [281], most strategies rely on machine learning (ML) approaches exploiting annotated data for training the PR classification model [168, 264]. We argue that this might involve some problems and limits. First of all, these methods usually rely on external resources of structured domain knowledge, such as ontologies or Wikipedia. Relying on similar external graph resources is generally chosen because of their comprehensive set of knowledge items and their extensive relationships. However, they might suffer a limitation when applied to domains not well (or at all) covered by the external knowledge. Furthermore, as multiple strategies might be proposed for presenting the same topic (as testified by the existence of multiple textbooks for the same domain), relying on an external knowledge might not match with relations reported in a specific resource. If our goal is acquiring the **prerequisite structure** of an instructional resource we must rely on the textual content of the resource itself as the only available source of information. *Acquiring prerequisite relations*

*from textual materials without exploiting any external source of information* corresponds to our perspective on the task of PR identification. The need to adopt a similar approach arises from the observation that this would be: *a)* suitable for prerequisite learning also when external sources of structured information are not available; *b)* capable of inferring prerequisite relations directly from the educational material where concepts are described. Motivation b) represents a particularly desirable scenario, especially if we consider that a PR relation strictly depends on the writer's communicative intent and expository style. As an example, consider when an author decides to explain a topic starting from broad concepts and definitions as opposite to cases when (s)he starts from specific cases or examples: top-down and bottom-up approaches have a significant impact on the representation of prerequisites between concepts since they imply opposite relations (i.e. from general to specific or from specific to general).

However, regardless of the adopted perspective, datasets annotated with labels indicating PR relations are essential resources for training and testing systems for automating the extraction of such relations. Manual annotation is still a preferred practice over automatic annotation for obtaining reliable datasets reflecting humans' intuition about the annotated resource content. Unfortunately, we observe a scarcity of such resources, which instead are crucial for advancing the research in the field. The controversial nature of PRs and the difficulty of their identification play a role on that [91, 109, 283].Taking a deeper look at the few available datasets, we notice that, in most cases, annotation instructions tend to be absent or fairly basic, often relying on a naive definition of prerequisite relation. However, it is widely acknowledged by the community working on resource creation and corpora annotation that properly defining problems is vital when designing an annotation task: the clearer the definition of the problem, the better the data that will be collected, allowing for annotations less influenced by the subjectivity of the single annotator or by her/his interpretation of ambiguous instructions [127]. The consequences that arise from rough definition of the PR annotation task are low annotation agreement [56, 91, 109], difficulty to compare datasets annotated in different projects and performance variability of systems trained on such data. The presence of different PR identification approaches and the lack of guidelines defining good practices to encode prerequisites brought to the creation of datasets that are not easily comparable and that capture different aspects of the relation. Hence, defining annotation standard practices is essential at this stage. We try to fill this gap by proposing our approach for PR identification in educational texts.

## 1.4 Approach

The work proposed in this thesis aims to advance the research on prerequisite relation identification by proposing a methodology for dealing with PR between concepts in educational materials, both from the perspective of their manual and automatic identification in textual instructional materials.

The novel methodology incorporates the lesson learned from our experience and the observations emerging from the comparison with past works within the same line of research. In particular, we developed a multidisciplinary approach which combines the perspectives of Computational Linguistics, Computer Science, Engineering and Education to deal with prerequisite relations, from their definition and human identification in textual resources to automatic acquisition. Works expressing the point of view of Educational research provided the theoretical background for defining prerequisite relations, while Computational Linguistics and Computer Engineering gave us the tools for defining the principles of manual and automatic identification of such relations in texts.

The most fundamental – and also novel, with respect to other works – principle of our methodology is the following: *PRs should be uncovered from textual instructional resources on the basis of the resource content*. Although ideally simple, and possibly close to the way in which students acquire knowledge from learning materials, formally incorporating such principle into PR identification allows to integrate the multiple perspectives sketched above. More in detail, Education helped us understanding the characteristics of the relation passing between concepts showing a propaedeutic relationship; Computational Linguistics provided us with many approaches and solutions for identifying information contained in textual resources, both when manually annotating and automatically extracting them; Computer Science and Computer Engineering contributed by providing models for defining knowledge representations and when defining the strategy for automatically extracting PRs. By adopting a multidisciplinary perspective, we incorporated solutions borrowed from different research fields that, for the time being, mostly tackled prerequisite relations from different points of view and addressing different challenges. Our approach aims to combine them in order to build a shared vision around the task and a common ground of discussion for advancing the research.

On a more practical note, our methodology offers solutions for tackling manual annotation and automatic identification of PRs within educational texts. Each task is addressed separately by our methodology, but both are closely related by a shared common principle, i.e. considering the text of instructional materials as the main source of information for acquiring the PR structure of the resource. For what concerns manual annotation, the methodology provides instructions for building resources manually annotated with prerequisite relations. To this aim, we designed an annotation protocol, *PRErequisite Annotation Protocol* (PREAP), to support manual identification of PR relations while reading educational materials. The ultimate goal of this process is the creation of Gold-PR datasets, i.e., a manually labelled sets of items resulting from the annotation of a single expert or from the combination of all annotators' judgements.

The protocol is the result of our effort towards formalising the PR annotation process, which involved, for instance, providing a definition of educational concepts and indicating how to recognise a prerequisite relation along the text flow, and defining a strategy to evaluate the resulting datasets. The annotation of educational materials while following the principles of

PREAP annotation protocol is supported by PRET *(PRerequisite Enriched Terminology)* interface. Indeed, whereas PREAP defines the basic principles of the annotation, with PRET tool we provide all the functionalities that are needed to perform PR annotation on educational texts following those principles.

While the two above support manual annotation, PREL *(Prerequisite RElation Learning)* deals with PR automatic identification. PREL is an approach for automatically acquiring prerequisite relations between educational concepts. Also in this case, PREL is designed to be exploited in automatic PR learning scenarios where the text of the resource is the only source of information.

The set of instructions for annotation, the interface for text annotation and the model for PR automatic learning build up what we refer to as *'PR Framework'*. The framework systematically presents our methodology, as embodied by the three elements of the framework, namely the annotation protocol, the interface and the automatic PR learning system. All are meant to guide and support researchers along the different phases of their work, from dataset building to automatic extraction. On the one hand, by providing solutions specifically targeting each task, the PR Framework answers to a need of practical nature: each task, although dealing with a common problem, is affected by specific issues that must be dealt with individually. On the other hand, the framework also offers an advantage: one might not be interested with the whole process of going from the raw text to the prerequisite structure of a resource but only in, e.g., creating a manually annotated resource or applying the automatic extraction on its own annotated resources. Thanks to the framework modularity, researchers can rely on whatever element targets their needs without missing out information and, meanwhile, preserving the fundamental principles of the methodology.

## 1.5 Goals and Research Questions

The work presented in this thesis tackled multiple tasks associated to prerequisite relations. While defining the strategies for addressing them, we preserved the same perspective: using textual instructional materials as sole source of information. As a consequence, our wide goal consisted in defining a methodology for uncovering the knowledge structure of an instructional material exploiting only the content of educational texts. In practice, and recalling the title of this thesis, our aim can be framed as *defining a method for obtaining a formal representation of the prerequisite structure underlying the content of an instructional material relying on its raw text*. Although it might seem trivial at a first glance, this issue was actually neglected, or only marginally addressed, by most existing literature. Finding a solution to such problem represented the ultimate goal of our study. In order to pursue this goal we had to address many open challenges and issues that we investigated throughout our research.

We summarise our main research goals (G) and related research questions (RQ) as below.

(G1) Defining a methodology for uncovering the prerequisite structure of educational textual

materials in order to obtain a representation of the resource content without relying on any external structured knowledge base.

RQ1) How should we define concepts and prerequisite relations if our goal is searching for them within the content of instructional materials?

RQ2) What is the level of granularity of concepts and how can we uncover the relations between them?

RQ3) Which are the advantages of employing our approach as opposed to currently frequently adopted ones?

RQ4) Does the linguistic complexity of the instructional material play a role in the identification of prerequisite relations?

(G2) Defining annotation guidelines that could set a standard practice for manually annotating prerequisite relations.

RQ1) Which are the most suitable resources for identifying the PR structure of a subject matter?

RQ2) How should we compute the agreement between text-bound PR annotations?

RQ3) Which is the most suitable approach for combining different annotations in order to obtain a gold standard PR-dataset?

RQ4) Which annotation interface can we use to achieve the goal of obtaining the annotated resource? Should we use the same interface also when the annotators is not familiar with textual annotation practices?

RQ5) Can we use the PR-annotated resource, which models the content of instructional materials, to allow both linguistic and educational oriented analyses, aimed at exploring how prerequisite relations are instantiated within texts, as well as to train and validate automatic PR learning approaches?

(G3) Developing a model for automatic extraction of prerequisite relations which doesn't need information acquired from structured knowledge representations to be used as proxy for prerequisite relation identification.

RQ1) Which textual features can we rely upon in order to train a machine learning model able to identify prerequisite relations between concepts based on the content of an individual instructional material?

RQ2) How can we acquire the knowledge related to a specific concept from the whole textual resource content, which traditionally deals with concepts in a fluid manner presenting them in continuous explanatory presentations?

The first set of research questions, related to the research goal (G1), addresses issues related to the general approach and methodology for uncovering PRs from texts. These questions will be dealt with by Chapters 2 and 4, where we outline our perspective on concepts and prerequisite relations and discuss the methodological issues of our approach. Research goal (G2) and related questions deal with the practical implications of implementing an annotation methodology for PRs based on the identification of relations on the basis of the content of instructional materials. They will be discussed in Chapters 5, 6 and 7. In those chapters we will present the basic principles of our annotation methodology, how we implemented them in an annotation tool and report an annotation project aimed at building and exploring a PR-annotated dataset. Research goal (G3) addresses the issue of automatically identifying PRs from texts without resort to structured domain knowledge, which is actually a novel approach with respect to existing ones. Our answers to the research questions associated to (G3) will be provided in Chapter 8.

## 1.6 Contributions and Research Challenges

The main contribution of the work of this thesis is represented by a **novel methodology for dealing with prerequisite relations in instructional textual materials**. We systematised the methodology within the PR Framework, which is aimed at supporting researcher when dealing with prerequisite relations in educational materials. Defining a methodology for addressing the tasks of PR identification in texts brought us to tackle many challenges connected to the research fields dealing with PR manual annotation and automatic extraction. Here below we outline our main contributions and the related research challenges we had to face.

1. **Definition of a Research Problem**: as we will discuss in chapter 2, we observed a varied and, to a certain extent, conflicting definitions of prerequisite relation and between which items it can occur. Our first challenge was defining our perspective and setting the boundaries of our study. This task opened the way to other related issues:

   a) Formalising our definition of concept and prerequisite relation, backed up by theory from Linguistics, Knowledge Representation and Pedagogy. Defining a problem is a crucial issue of the research as it will guide all further investigations. Eventually, we delivered our own definition of concepts and PRs as educational items that naturally occur in educational textual materials and must be searched for within the content of the instructional resource. Our effort towards providing a **problem formulation** of the task we tackled represents one of our main contributions as it provides the grounding for formally addressing the task also in future researches.

   b) Evaluating the **effects of text variety and linguistic complexity** on the identification of PRs based on the content of textual instructional materials. This investigation is actually novel in the line of works dealing with prerequisites as so far none explored the

relationship between the complexity of texts and the identification of PRs, at least to the best of our knowledge.

2. **Annotation in Context:** the main novelty of our methodology is that we propose to uncover PRs directly from textual instructional materials without resorting to knowledge bases or structured knowledge representations (from text to prerequisites, with no intermediate steps). Such approach allows to produce a text-driven and context-anchored annotation, i.e. experts annotate relations actually conveyed by the text itself and not by other sources, including their background knowledge. The inserted relations are thus anchored to the linguistic context where they take place, at least according to the opinion of the expert who annotated them. Such representation of the annotated information fosters analyses and researches not otherwise possible. For example, it allows to investigate the contexts where PRs take place at different levels of linguistic analysis, possibly gaining a deeper understanding about how the phenomenon is instantiated in texts and eventually confirming or discovering insights regarding the linguistic features we should take into account for the automatic identification of PRs.

3. **Annotation Protocol:** the main challenge we addressed when first tackling prerequisite relations was defining a systematic strategy to uncover PRs from the content of educational materials without suffering the influence of annotators' background knowledge. We eventually came to define an annotation protocol which provides a documentation containing example and clear guidelines to carry out the annotation according to the defined principles. The protocol also incorporates and it is grounded on the findings of the investigations on the role of complexity in the identification of PRs. In order to define the principles of our protocol, we addressed the following challenges:

   a) **Annotation Task and Manual:** we put our efforts towards defining a novel annotation task which is based on a shared definition of what a concept and prerequisite relation are. One of the protocol goals is to limit the disagreement between annotators which might easily occur due to the subjective nature of the PR relation. Such goal is particularly challenging: as mentioned, PRs cover a fuzzy area of semantic relations thus can easily undergo multiple interpretations depending on the reader interpretation of the text. We developed an annotation manual which specifically addresses this issue.

   b) **Annotation Recommendations:** while testing our annotation protocol, we realised that some choices might be project- dependent. They concern, e.g., the approach for annotation consolidation to create the ultimate gold standard, or the use of certain non-compulsory steps, such as annotation revision. Considering our goal of providing the community with a set of recommendations for creating their own PR annotated datasets that could be widely adopted and applied to different projects and texts, we accompanied the annotation manual with a set of recommendations and good practices for carrying

out the annotations following the protocol principles and for adapting the methodology to different scenarios and goals. Rather than developing a strict annotation protocol which provides a fixed solution for any possible situation, we preferred to offer multiple options for certain tasks involved in the annotation process in order to make the protocol adjustable to different needs. In our view, the only aspect the must be preserved at all costs and in all situation is the text-bound annotation approach.

4. **Annotation Evaluation**: the deep analysis of the PR phenomenon in instructional texts brought out the limits of the current approaches when evaluating the output of manually produced PR annotations. In order to overcome this limits, we propose a novel methodology for evaluating the annotations. Our evaluation approach accounts for the distinguishing properties of the prerequisite relation formalised within the problem definition and integrates them into existing approaches in order to adapt them to the PR scenario.

5. **Annotation Interface:** another contribution of this research concerns the development of a novel annotation interface which implements the principles of the annotation protocol. The interface allows to create manually annotated PR datasets and also to analyse them relying on build-in analysis functionalities specifically designed to support the exploration of PR realisations in textbooks or other textual instructional resources. Developing the interface brought us to tackle the following challenges:

   a) Fostering the annotation of prerequisite relations for a **wide audience**, involving both experts in the process of textual annotation and researchers in the field of Education, which are usually not familiar with text annotation tasks;

   b) Creating a complete annotation tools which not only allows to obtain PR-annotated datasets, but also **supports annotation analysis and use**. This is done to promote the dissemination of our annotation protocol.

   c) Support **collaboration** between multiple users on the same project but at the same time foresee the **supervision** of a project manager to guarantee the quality of the data.

6. **Multi-purpose PR Annotated Resource:** we produced a gold dataset manually annotated with prerequisite relations between pairs of concepts occurring in a text. Even if limited in its dimension, this dataset is available for the community and for further investigation on the PR phenomenon. The dataset is versatile as it can be helpful in several tasks. For example, it can be used as a dataset to train machine learning algorithms, as gold standard for the evaluation of PR extraction algorithms, or to generate novel learning materials.

7. **Textual Analysis of Prerequisite Relations:** along this dissertation we report the analyses carried out on PR-annotated datasets aimed at investigating the realisation of prerequisite relations within instructional materials. This is the first time, to the best of our

knowledge that such analyses are carried out on a educational resource manually annotated with prerequisite relations. The annotation protocol and resulting dataset allowed a linguistic-based analysis aimed at verifying whether different concept roles (prerequisite or target) are also associated with different linguistic contexts or characterised by diverse linguistic structures.

8. **Automatic PR Identification:** we propose a novel text-based extraction of PRs from textbooks. Consistently with our idea that PR relation is largely affected by the characteristics of the text where it is found, we propose an extraction approach that, contrary to many others, aims to extract PR from unstructured data (text) without resorting to external structured knowledge. This contribution involved dealing with:

   a) The definition of **text-based features** for training the model which have the characteristics of being acquired from the raw text of the resource only.

   b) Experimenting with different **architectures** of the model in order to identify the best performing one.

   c) Defining a novel approach for **automatically creating the units of learning** associated to each concept mentioned in the text.

## 1.7 Chapters Guide

As said, this dissertation proposes a novel methodology for dealing with the manual and automatic identification of prerequisite relations from educational textual materials. The methodology is based on a thorough investigation of the main issues related to uncovering the prerequisite structure of an instructional resource relying on the content of the resource. For this reason we carried out preliminary explorations aimed at understanding the role played by text, as a mean to vehicle content, in the identification of PRs. To pursue the goal of reporting our findings and presenting our methodology, the remainder of the thesis is organised as follows.

The first part of the dissertation discusses the related work and background research. Specifically:

- **Chapter 2** presents the basic notions related to concepts and propaedeutic relations, which are at the center of inquiry of the research reported in this dissertation. After briefly presenting the definitions provided by different fields and the literature, we introduce in this chapter our perspective and point of view. We also outline the main properties of concepts and relations: we take them into account along the research for defining our methodology and provide examples to clarify how concepts and PR relations might be represented in different instructional materials.

- **Chapter 3** examines the work related to our research. In particular, we deal with two main research areas: the science of annotation and the literature related to prerequisite relations. This chapter is meant to provide an overview of our methodological points of reference with respect to the development of the annotation protocol, but also where we stand with respect to the existing research on prerequisite relations.

After presenting where we stand with respect to the existing literature, we introduce our preliminary studies carried out to explore the different factors intervening in the identification of PRs withing texts.

- **Chapter 4** discusses the principles of our methodology and tests their feasibility and implications in a crowd-based experiment aimed at investigating the role played by text and linguistic complexity in the manual identification of PRs based on the content of small instructional texts extracted from different resources. The results of our experiment provide the grounding for our methodology and PR annotation protocol, thus in this chapter we also discuss the practical implications of our approach, how they impacted the definition of the methodology and which challenges they open.

The following part deals specifically with our methodology for uncovering PRs. Each chapter presents the methodology through a different component of the PR framework:

- **Chapter 5** introduces our protocol for PR annotation on textbooks PREAP. We first discuss the iterative development process that brought to the definition of the current version of the protocol, as well as its compliance with the desiderata and requirements for annotation tasks outlined in the literature review. Then, we present the guidelines and documentation that we release along with PREAP in order to reapply the annotation process to other projects.

- **Chapter 6** presents PRET, the annotation interface designed to support the application of PREAP principles on corpora. We discuss the modules and functionalities of the annotation tool and present a usability test carried out to evaluate the usability of the tool from the perspective of different populations of users.

- **Chapter 7** reports an annotation project carried out on PRET tool and following PREAP principles devoted to produce a PR annotated resource which models the content of the chapter of a computer science textbook. The chapter also shows some analysis aimed at investigating the realisation of PRs in instructional materials.

- **Chapter 8** discusses PREL, the model for automatic prerequisite relation learning from instructional texts, and reports two different experiments aimed at evaluating the performances of the model on a textbook and the iterative design process which produced different versions of the PR resource.

The last part of the thesis concludes the work:

- **Chapter 9** summarizes the contributions of this dissertation, and discusses limits, future improvements and applications of this research.

CHAPTER 2

# BASIC NOTIONS

Propaedeutic relationships, as established in instructional materials, are at the center of inquiry of the work proposed in this thesis. These relations occur between pieces of knowledge of the subject domain, which we can refer to as educational concepts. Although our research mainly addresses the issue of identifying propaedeutic relations and leaves aside the task of identifying concepts in texts, we still need to define both items in order to set the boundaries of our research.
The aim of this Chapter is to provide an overview of the basic notions that will be employed with respect to concepts and propaedeutic relation along this thesis.

## 2.1 Concepts

The term "*concept*" refers in general terms to an abstract and general idea conceived in the mind. Such informal definition is extremely broad and it can be declined in multiple contexts, although the strict relationship between concepts and human cognition is common to most of them [53].

Many research fields investigate the nature and essence of concepts. Early reflections on the nature of concepts were carried out by Philosophy. Plato, for example, defined concepts as the essences of things, assigning them an abstract and ideal nature, whereas, in the Aristotelian view, concepts are representations of classes of objects, symbols, or events sharing common properties. More recently, most of the discussion has been centered around the relationship between concepts and the human interpretation of the world [96]. Going into details about the philosophical debate about concepts is beyond our scope. Suffice to say in this context that the debate concerning the nature of concepts and their role in human reasoning and world interpretation has been inherited by many other fields, such as Linguistics and Education. These two disciplines investigate the nature of concepts from different perspectives. Linguistics explores the relationship between

concepts, meaning and lexicon, while Education is interested in the role played by concepts in the process of learning. At a first glance, the two issues might seem unrelated, however, focusing on the process of acquiring knowledge from textual educational resources (e.g., textbooks) requires to take into account also the way in which concepts might be represented through language.

As a matter of fact, Linguistics (and Semantics in particular) has deeply investigated the relationship between concepts, lexical entities, meaning and phonological and grammatical representations and many theories of meaning have contributed to the discussion. We will not review all of them, but we believe it is worth at least mentioning Conceptual Semantics [130] which, in our view, is the most relevant for us as it provides a theoretical foundation for identifying concepts as corresponding to lexical items within a textual resource [200]. More than a linguistic theory, Conceptual Semantics is a framework defining how humans express their understanding of the world by means of linguistic utterances. According to Conceptual Semantics, concepts are mental structures representing the world that interact with formal aspects of language, i.e. phonological and grammatical. As a result, concepts might be seen as corresponding to words (more precisely, lexical entities) in texts that can be used to communicate about new knowledge to someone that has to acquire and internalise it.

On a different note, Education has investigated concepts with respect to the process of knowledge acquisition. Here, concepts are frequently referred to as 'Knowledge Components', generic pieces of information that can be used to accomplish tasks [143]. Their role in the learning process is two-fold: on the one hand, the instructional designer has to properly use concepts when referring to objects, events or entities; on the other hand, the learner can take advantage of concepts to organise the new knowledge into categories sharing common properties [190, 238, 266], thus reducing the cognitive effort [136].

Next to these theoretical reflections on concepts here briefly overviewed, we must mention also other fields which provide more concrete interpretations of the term 'concept'. Knowledge Representation, for instance, a field of Artificial Intelligence whose goal is the description of a state of the world using a machine–readable formal language [40], refers to concepts as atomic and discrete components in a knowledge structure (e.g., a concept graph or ontology) that represents a subset of a domain [111, 209]. Consider again the knowledge graph of the previous chapter (Fig. 1.1). In that case, concepts are represented by the nodes denoted by letters. Note that the level of granularity and interpretation of a concept may vary between knowledge models on the basis of what is intended as minimum component of a knowledge structure [208]. For example, modelling the content of textual instructional materials by means of domain concepts is traditionally carried out at section-level: domain experts manually index sections of the considered resource with the set of domain concepts discussed within it [66].

Our perspective on concepts inherits some of the most fundamental intuitions of the perspectives sketched above, however it also takes the distance from them at many levels, as we will discuss shortly. In between the theoretical and practical interpretation of concepts, we feel

our view is more closely related to a branch of Linguistics which deals with specialised terms: Terminology. As a discipline, Terminology investigates the set of specialised words (as well as their associated meanings and inter-relations) related to a specific domain [52, 200, 242]. Accordingly, the terminology of a text corresponds to the set of domain terms which have a specific meaning and associated knowledge within the domain. One of the main characteristics of domain terms is being very precise and less ambiguous as possible, rarely presenting cases of synonymy: building a terminology consists of identifying as many lexical units as there are concepts in the domain while finding unique correspondences between a concept and a term [106]. Therefore, the terminology of a subject domain ends up reflecting the conceptual organisation of the discipline and tends to provide as many lexical units as there are concepts in its subspace [242]. In line with the work of [57], we argue that concepts are better associated to lexical entities (keywords, in [57]) rather than text sections, as proposed by knowledge representation theories: such approach allows to obtain fine-grained knowledge structures representing the content of instructional materials. As a consequence, we see concepts as domain terms, mentioned and discussed within a text, unambiguously associated to pieces of domain knowledge. This is why we believe that Terminology is the line of research which better matches our perspective. By relying on this interpretation, we can say that a knowledge structure consists of variously interlinked *terms*, as displayed in Figure 1.1.

Before going into the details of the nature of the links and relations established between terms, we will formalise our definition of domain concepts and briefly discuss how they can be acquired from texts.

### 2.1.1 Defining Concepts

As we discussed in the previous section, defining concepts is quite complex as different disciplines declined the term according to their needs. Based on what outlined above, we we can highlight the following issues related to the definition of concepts as the most relevant to our research.

i) Concepts can be represented in the language by means of lexical entities.
ii) Lexical entities embody pieces of knowledge denoting the set of properties owned by the concept they represent.
iii) Learning consists, among the other things, of acquiring the information associated to concepts. Acquisition happens when concepts are coherently stored in the mental representation with respect to previously acquired knowledge.

Our definition of concept relies on the above assumptions and combines multiple definitions presented above, although we adopt a more operational definition in order to address our needs. In particular, we share some space with the definition of concept defined in the context of the work of Chau *et alii* on concept annotation in textbooks [57]. We define concepts as follows:

- A concept can be instantiated in text as corresponding to a text fragment (i.e., a term) of an instructional resource.
- A concept embodies a piece of knowledge of a subject matter, thus it has a specific meaning in the considered domain, but it could also have a different meaning in other domains or in every-day language.
- The set of concepts of a subject domain corresponds to the domain terminology.

Practically speaking, we represent concepts as domain entities corresponding to single or multi-word domain terms mentioned in textual instructional materials. Note that, since each concept is used to represent a piece of knowledge, in the educational setting each term of the terminology implies the domain knowledge associated to the corresponding concept, which constitute what the learners should acquire from the resource. However, we are not interested at this stage of the research in formally representing such knowledge (or the properties owned by a concept) in the knowledge structure.

It is important to note that, as a consequence of such definition, *the set of domain concepts emerges as the list of domain terms mentioned in a textual resource*, rather than abstract and resource-agnostic entities. This is a crucial aspect of our work: while the domain terms of a subject matter are potentially infinite (as the discipline may introduce new concepts as the research goes on), the terminology of a document is a finite and fixed set corresponding, in the most inclusive case, to the whole set of nouns mentioned in the text. Most frequently, the text terminology is instead a sub-set of nouns which are identified as particularly relevant for the domain. To put this idea into practice, consider the following short text extracted from Wikipedia[1]:

> Addition is one of the four basic operations of arithmetic, the other three being subtraction, multiplication and division. The addition of two whole numbers results in the total amount or sum of those values combined.

Domain terms, as manually identified by us, are underlined in the text. As can be noted, this terminology contains both single terms (e.g., addition, operations, division) and multi–word terms (whole numbers), but not all nouns are included: 'amount', e.g., was excluded from the terminology since, according to our view, it is not a domain–relevant term. However, one might rightfully wonder how to distinguish a domain term from a common noun. We will discuss this issue in the next section.

### 2.1.2 Uncovering Concepts from Textual Documents

The effective approach to acquire the set of domain terms from a document is to involve a domain expert into the process [57]. Based on its expertise and domain knowledge, a human expert will

[1]Page about 'Addition': `https://en.wikipedia.org/wiki/Addition`

be able, by reading a text, to select which are the most relevant terms (and, as a consequence, also the concepts) mentioned in the document. However, when large amounts of documents are considered, it is unfeasible to perform this task manually. Automatic strategies might come to aid to speed up the process by automatically selecting a set of candidate domain terms to be manually revised or to completely replace the expert.

Concept extraction from documents is now a well-established research issue in Information Extraction [24]. As a matter of fact, also in the field of Knowledge Representation part of the research aims at finding strategies for (semi)automatic creation of ontologies by learning concepts from different information sources [22, 49, 116, 300], proving the relevance of this task in multiple scenarios. As we will show, the automatic identification of domain terms from educational texts is only marginally covered by our work since we demanded concept extraction to an existing tool. However we believe that, for a matter of clarity, it is worth briefly reviewing the literature concerning this topic and the main approaches available in order to clarify where we stand and provide the elements for understanding why the approach we chose is the best suited for our research.

Among the methods addressing concept extraction from unstructured resources, e.g. texts, we distinguish two main approaches: those exploiting external resources and those relying on the information available within the document. Entity linking, falling into the former group, consists of identifying mentions of entities in a text and linking them to their corresponding entry in a Knowledge Base (KB) [185]. In such approach, first the relevant entities are matched against a dictionary containing the entities of the KB, than a disambiguation step finds the right correspondence between the mention and the identifier [250]. When it comes to educational concepts, Wikipedia is the most widely used KB for defining concepts [215, 283, 301] due to its wide coverage, easy connection with other KB such as DBPedia or Wikidata and richness of metadata for each entry. The limit of such approaches becomes apparent when the domain of the resource is too narrow and/or not well covered by the KB: exploiting multiple KBs could solve the problem, but since a commonly agreed definition of 'entity' is still missing, specialised settings remain an open challenge [185].

Approaches relying only on the information internal to a specific resource are valuable alternatives in these scenarios. Computational Linguistics and NLP strategies are generally employed to perform terminology extraction. They are often distinguished in *i)* pattern-based linguistic approaches, which employ syntactic parsing to identify domain terms among short noun phrases in the text [92, 107, 118], and *ii)* statistical approaches, that assign a *termhood* degree to words relying on distributional properties [233, 261, 298] or on sentence-level contextual information [63, 81, 273]. It should be noted that while the above methods address the task of general terms extraction, only a small number of projects have considered a textbook corpus for extracting domain–relevant concepts [149, 281]. This is possibly due to the fact that scientific literature usually contains more complicated sentence structure compared to other resources

generally used for key phrase extraction, such as news articles, and there is also a noticeable variation between scientific literature of different domains and targeting different audiences [198]. On the other hand, such rich structured information embedded in scientific literature could be utilized in key phrase extraction to acquire more robust and rich information.

## 2.2 Prerequisite Relations

It is widely acknowledged that, when imparting knowledge, properly introducing concepts, not only in terms of associated notions but also with respect to other concepts, plays a crucial role in supporting learners' understanding of the subject domain while simultaneously avoiding student's frustration, misunderstanding and disorientation [67, 101, 241]. For example, it is common practice to introduce the concept of "addiction" before discussing "multiplication" as it might be useful to refer to the former when introducing the latter. In general terms, such condition is generally expressed through the notion of propaedeutic knowledge: a propaedeutic notion represents *the piece of knowledge a student has to learn and master before approaching novel content*. We often see this notion used to express the prior knowledge required to access the content of a book or of a lecture. For instance, in their 1999 book "Foundations of Statistical Natural Language Processing" C. Manning and H. Schutze explicitly say that a student reading their book is assumed to have "prior programming experience, and has some familiarity with formal languages and symbolic parsing methods. [...] The student may have already taken a course on symbolic NLP methods, but a lot of background is not assumed" [180]. Paraphrasing the authors, this means that the basics of programming, formal language and symbolic parsing won't be covered by the book since the reader is supposed to know them already, while the fundamentals of NLP are not taken for granted. In order to express a propaedeutic relationship between learning concepts more concretely, we can exploit *prerequisite relations* (hereafter also referred to as PR).

In general terms, we define a ***prerequisite relation PR*** as a *binary dependency relation connecting a prerequisite and a target concept where the first has to be known in order to understand the second* [167]. In other words, the knowledge associated to the **prerequisite concept** covers the prior knowledge required to understand the **target concept**. Although the definition of "prerequisite" provided above, making reference to the prior knowledge, seems intuitive and grounded in our common-life experience, formally defining prerequisite relations and prerequisite knowledge is actually difficult.

As noted by [124], the term "prerequisite" seems to bear at least two meanings: on the one hand it expresses a pedagogical relationship between two elements that the student should learn, on the other hand it indicates a formal mechanism that can be used to partially order two units of instruction (concepts, pages, exercises or similar) inside a sequence of learning materials. From a cognitive perspective, the prerequisite relation exists as a natural dependency among concepts in

cognitive processes since it denotes which concepts, or skills, a student has to learn before moving to a new topic [168]. This leads us to an interpretation of the prerequisite relation as related to the process of organising and ordering concepts when designing instructional materials.

First evidence about the importance of properly sequencing concepts was reported since early studies in instructional design carried out by Robert M. Gagné (see e.g. [101, 102]) and David Ausubel [25]. Both stressed the importance of prior knowledge in being able to learn about new concepts, positing that learning happens in a sequential manner and builds upon prior knowledge. According to this view, the act of recalling prerequisite concepts becomes one on the main events of learning. Gagné's work had a huge impact on instructional design and pedagogy, paving the way to the development of instructional theories that, although different in their use of the terminology, share the same interest for the role of prior knowledge in new knowledge acquisition [189]. Ausubel's work, on the other hand, comprises the foundations of Novak's concept mapping theory, discussed at the end of this Section (subsection 2.2.4).

### 2.2.1 Properties of PRs

Since Gagné's and Ausubel's works, where relations between concepts can be expressed by means of tree– or graph–like structures, represent our theoretical background, we account for PRs as *dependency relations* between *concept pairs*. Formally, they are represented as *directed relations* semantically expressing *learning precedence*. Although we are aware that other interpretations of PRs are possible as opposed to Gagné's hierarchical structures, we rely on such interpretation as it allows us to formally define the following PR properties generally used to describe graphs. Here below we go through them in detail. Note that, in what follows, we will use the notation $\prec$ to refer to the prerequisite relations. For example, $A \prec B$ shall be read as $A$ is prerequisite of $B$. More concretely, recovering our example on 'addition' and 'multiplication', we will write $addition \prec multiplication$ to express that (i) there is a relation between the two concepts and (ii) *addition* is the prerequisite, while *multiplication* is the target.

- *Complete and Directed Relations*: if the pair of concepts $A$ and $B$ shows a PR, there must be a relation connecting them and the relation must be directed, thus the relation either relates $A$ to $B$, or $B$ to $A$. In other words, if it's true that *addition* and *multiplication* are related with each other, then either $addition \prec multiplication$ or $multiplication \prec addition$.

- *Irreflexivity*: PRs can only appear between pairs of distinct concepts. Indeed, it is logically impossible for a concept to be prerequisite of itself. Thus, if a PR relation exists between two generic concepts $A$ and $B$, then $A$ must be different from $B$. In other words, $A \prec A$ (e.g., $addition \prec addition$) is not a valid prerequisite relation.

- *Asymmetry*: being directed relation, PRs reflect in which sequence concepts must be learned. In order to preserve the above irreflexive property, sequences can't involve cycles and loops,

otherwise a student will never reach the final target concept. In practice, if $A \prec B$, the opposite cannot be true (e.g., if *addition* $\prec$ *multiplication*, *multiplication* $\prec$ *addition* can't be valid);

- *Transitivity*: a concept inherits its previous prerequisite concepts as it would be impossible to move on in the process of learning unless all background notions are properly acquired and stored in the learners' mind. This fact implies that every target concept has at least one direct prerequisite concept, but it could also have some indirect prerequisite concepts inherited from the learning sequence. In other words, for every $A$, $B$, and $C$, if $A \prec B$ and $B \prec C$, then $A \prec C$. Expanding our example with the concept of *'power'*, we can apply the above property by saying that if *addition* $\prec$ *multiplication* and *multiplication* $\prec$ *power*, then *power* inherits the prerequisite of its prerequisite concept and *addition* $\prec$ *power* is a valid PR.

### 2.2.2 Paradigms to Uncover PRs

A question that still remains open is *how can we uncover prerequisite relations between concepts in a subject domain?* The question is far from trivial and it implies to answer some other related questions. For example, should we consider PRs as absolute relations that are always true in a domain regardless of the way a teacher presents concepts? Or are they bonded to the way concepts are organised in instructional materials, such as lecture notes or textbooks? We refer to these two opposed paradigms as *ontological view* and *pedagogical view*. While the former considers PR as absolute relations of the subject domain, the latter considers PRs as fluid relations that can vary with respect to the way they are presented in a resource. These views do not necessarily match, or at least not perfectly, because the latter is bonded to a specific organisation of concepts, while the ontological view aims to develop a wide–coverage resource–independent representation of the domain. Indeed, although there could be some shared practices suggested by our common sense, the order of concepts largely varies depending on the resource. Indeed, bonding PRs to a specific resource implies taking into account how the resource expresses relations between concepts. Consider, for instance, two opposite, but both widely adopted, explanatory approaches: top-down and bottom-up. The former tends to explain a topic starting from broad concepts and definitions, while the latter introduces specific cases or examples before discussing the bigger picture. As a consequence of presenting the concepts using one or the other approach, PRs entailed in the explanation will have opposite direction, i.e. from a general concept to a specific concept, or vice-versa.

To better appreciate this distinction, consider, for instance, two classic programming language books, one for C[2] and the other for C++[3]. The former explains *while loops* first and then *for loops*

[2]Ritchie, D.M., Brian W. Kernighan, and Lesk M.E. *The C programming language*. Englewood Cliffs: Prentice Hall, 1988.

[3]Stroustrup, B. *The C++ programming language.* Pearson Education, 2000.

(because *for loops* can be rewritten if you know *while loops*); in the second case, *for loops* are explained before (as a more general iteration statement) then *while loops* (as a specific case). Imagine we give each of these two books to a different person, both novices in Computer Science and programming, and then ask them which one, between while and for loops, has to be known first (in other words, which one is the prerequisite and which one is the target). Since the two persons won't have any other knowledge of the domain apart from the book they were given, they would possibly provide two different answers to the question, each based on the content they read. We could address the same question to two domain experts without showing them the books, thus asking them to order concepts based on the ontological view. In this case it is possible (although not certain) that the two experts would provide the same answer as emerging from extensive reading about the topic carried out over the years. Given the scenario depicted above, could we claim that one ordering of concepts is correct and the other is wrong? Assuming that both books allow a learner to properly acquire the concepts of for loops and while loops, we can only say that one ordering is more typical while the other represent a less frequently adopted organisation of concepts, but we can't make assumptions about correctness. As a matter of fact, both must be considered as valid options for introducing the concepts of 'loops' in programming languages. The motivation for the different directions can be traced back simply to the explanatory approach adopted by each resource.

We also empirically tested the above intuition in an experiment involving 60 subjects of different ages but all with a high educational level (i.e., ongoing or completed university education). The subjects where asked to order 10 triples of concepts taught in elementary education, some of which even commonly used in every-day communication (e.g., "Geometry, circle, cone", "point, line, angle", "number, integer, polynomial")[4]. Concepts of the triples were known to be related by a PR (i.e., $concept1 \prec concept2$ and $concept2 \prec concept3$), and subjects were presented only with randomly shuffled terms representing concepts to be re-ordered according to their own individual sensitivity. Despite the experimental setting was reflecting the ontological view to acquire PRs, none of the 10 triples was unanimously ordered by all subjects: although the majority of answers (on average, around 60%) converged toward a commonly agreed sequence defined by domain experts, different sequences were also proposed. Such result confirms our previous intuition: multiple organisation of domain concepts can be legitimately proposed. A limit of this study is that we can't know why a sequence was proposed, unless we explicitly ask the subject who created it. If we want to investigate what motivates the identification of a PR or, e.g., its direction, it seems reasonable to *prefer the pedagogical view as it would allow to explore the context of concepts and motivate relations based on the way concepts are presented in the instructional material*.

Relying on the pedagogical view requires to add another piece to the definition of PR provided above: *a PR emerges from the comprehension of the context where it is introduced.* Obviously, although more informative, there some consequences of acquiring PRs from the content of

[4]We verified through direct question whether subjects actually knew the knowledge associated to each concept.

resources. The implications of employing the pedagogical view, as opposed to the ontological view, to acquire the prerequisite organisation of a domain will discussed in Chapter 4.

### 2.2.3 Overlap with other Relations

Semantically, PRs cover a fuzzy area, partially overlapping with other kinds of relations. For example, we notice that that a PR frequently overlaps with lexical relations. For instance, if $A \prec B$, there is some probability that $A$ is also an hypernym of $B$ (e.g., *fraction* and *improper fraction*) as this relation easily recalls a taxonomic hierarchy. Such overlap happens more frequently in typical top-down explanations: here, prerequisite concepts tend to be more general (representing a broader class) than target concepts (representing a narrower class), and it would be the opposite in a bottom-up explanation where specific concepts are used to explain more general ideas. PRs are frequently associated also to holonym-meronym relations (better known as *part-of* relation in Knowledge Representation). A new topic can be presented in general terms in the first place, then each of its components can be described. Think, for instance, when introducing the *human body* as comprising multiple types of *cells* that together create *tissues* and subsequently *organ systems*[5].

Prerequisite relations may also coincide with semantic relations that are inferential in nature, such as causal and temporal relations (specifically with the precedence, or *before*, relation). Causality is relevant to many domains: in medicine, for example, causal concept maps have been proposed for helping anatomy students to handle the complexity of the subject [142]. When presenting historical events, the explanation of a phenomenon or an historical event is generally followed by a discussion about its effects (e.g., the assassination of the Archduke Franz Ferdinand on 1914 in Sarajevo is frequently presented as the *casus belli* of the First World War). Indeed, it is known that temporal and causal relations interact with each other at many levels [199, 206]. Consider again the example of the First World War, but many others could be made: by definition, an event can't precede its cause, so it fallows that if the death of the Archduke caused the war, the event must have happened before the conflict started. It should be noted that temporal/causal relations share also some common formal properties with PRs: they are both *binary, directed and transitive relations* [206]. Such shared properties fostered the use of temporal relations also in prerequisite relation identification [1, 214, 240].

### 2.2.4 Prerequisite Concept Maps

The PRs occurring between the whole set of concepts of a domain can be used to represent the prerequisite structure of the subject matter. Indeed, graph structures can constitute a straightforward approach for representing the knowledge components contained in an educational resource by means of concepts and relations between them.

[5]This is also another interesting example of top-down versus bottom-up presentation of medical concepts.

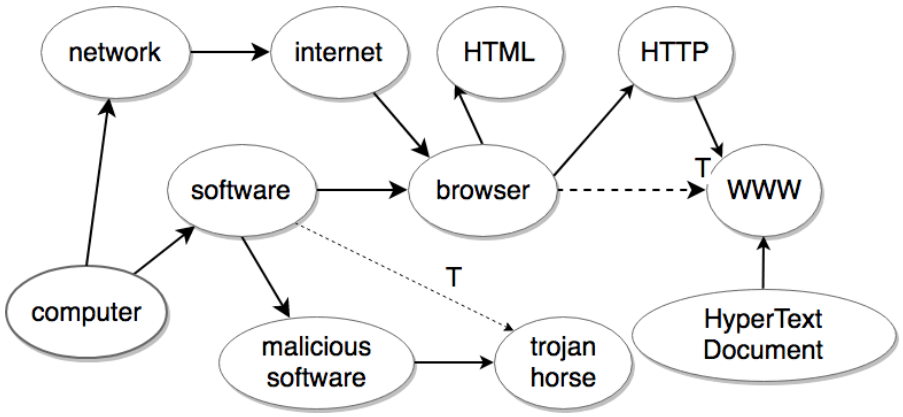



Figure 2.1: Prerequisite concept map of computer science concepts. Dashed edge represents transitive PRs.

In principle, graphs are structures made of vertices (nodes) and edges (explicit relations between nodes). Based on the properties of PRs described above, directed graph structures seem the most appropriate representation for modelling learning content as they allow to build acyclic graphs where domain concepts are depicted as nodes and relations are represented as edges. Such structure very much resembles *concept maps*, which indeed are diagrammatic approaches to represent human knowledge, for example by means of acyclic graph structures enforced by explicit relations between its components [209, 210] (see an example of concept map in Figure 2.1). Although concept maps have a long history in instructional design and Learning Management Systems (see [137] for a survey), and they are used to support educational activities (e.g., automatic lesson plan generation [3, 163] or automatic assessment [283]), in what follows we will use the term 'concept map' to generically refer to graph structures representing knowledge.

Concept maps are generic enough that any type of relation between concepts can be included. However, the simplest way to define a domain structure is to use only one possible relationship. Among them, we note that it is quite common to build concept maps including only prerequisite relations (see, e.g., [47, 155, 166, 240, 282]). We refer to such structures as *prerequisite concept map* as a specific case of graph depicting only prerequisite relations. Historically, concept maps result as a shift from the hierarchical representation of knowledge fostered by Gagné's work on learning hierarchies [102] through the work of Ausubel, which promoted instructional strategies (e.g. maps representations) as a mean to facilitate learning and retention of new information and to encourage students to find connections between new and previous materials [25]. While hierarchies impose a sequential ordering or concepts, graph representations are more flexible as they allow multiple paths between two concepts, accommodating students' needs and interests when designing learning sequences [47]. Thus they allow to surpass the limit of having only one valid concept sequence, while still preserving the idea of concept ordering thank to the directed graph representation.

## 2.3 Discussion

In the previous sections, we defined **PRs** as binary directed relations between pairs of domain terms denoting relevant **concepts** that can be represented through graph structures which we refer to as **prerequisite concept maps**.

To better understand how these elements collectively allow us to represent the prerequisite structure of a subject matter, consider the sample map depicted in Figure 2.1 referring to the computer science domain, reporting in particular concepts related to networks and the Web. As can be noted, nodes in the graph are labelled with terms (single or multi-word) denoting the domain concepts (e.g., *computer*, *browser*, *HTML*). The edges between the nodes represent PRs occurring between them[6]. This map can be used to explore the propaedeutic relationships between the included concepts.

According to the graph, for example, we see that *computer* is a prerequisite concept of *software*. It is interesting to note that such representation is effective to reflect the PR properties defined above. In fact, the relationship between *computer* and *software* has an explicit direction from the former to the latter. Additionally, all concepts involved in the graph are connected (meaning that all concepts discussed together must be somehow related from an instructional point of view) and no cycle is included in the graph, meaning that you can only navigate the paths connecting concepts going forward adding novel knowledge. For what concerns the transitive property, consider for instance the concept *trojan horse*. This concept has only one explicit prerequisite (i.e., *malicious software*), however it also has one transitive prerequisite, namely *software* (the transitive edge is labelled as $T$). Such representation tells us that, in order to understand what a *trojan horse* is, learners necessarily have to understand the concept of *malicious software* first. But at the same time, in order to have a clear understanding of the bigger picture in which these concepts play a role, students also need to know what a *software* is.

[6]Note that the map in Figure 2.1 was manually build by a domain expert as an example of a prerequisite concept map.

CHAPTER 3

# STATE OF THE ART AND RELATED RESEARCH

There are multiple research lines related to the work described in this dissertation. The background research dealing with *concepts* and *prerequisite relations* has been tackled in Chapter 2. Here, we will discuss two main research areas related to the following topics:

- *Manual annotation of textual corpora*: Sections 3.1, 3.2, and 3.3 will deal with issues related to textual annotation and annotation protocols;
- *Identification of prerequisite relations*: Sections 3.4 and 3.5 will focus on existing approaches for building datasets annotated with prerequisite relations and for automatically acquiring PRs respectively. Section 3.6 will provide a short selection of educational applications exploiting PRs to provide different services.

The goal of this section is to provide an overview of existing research related to our work and that inspired us, but at the same time we also want to underline that much works still needs to be done in order to address the challenges of PR identification.

## 3.1 Good Practices in Manual Annotation Tasks Design

The long standing research area devoted to corpus annotation deals, among the other things, with the definition of the best practices to design **annotation tasks** and develop annotated resources. The study of such processes is known as "science of annotation" [122, 127]. Its goal is to define general methodologies and recommendations for defining an annotation task that allows to carry out successful **annotation projects**. In what follows, we define an annotation project as the set of activities involved in the process of obtaining an annotated corpus, e.g. recruiting annotators, labelling the text applying the task guidelines, revising annotations, combining

multiple annotations to create a gold standard, etc. The annotation project is distinguished from the annotation task since, while the task consists of the general approach for annotating a certain information on a certain corpus, the project involves the practical activities for obtaining the annotated dataset from a specific corpus.

*Corpus annotation* can be broadly conceptualized as the process of enriching a corpus by adding linguistic (and other) information, inserted by humans or machines (possibly manually revised) in service of a theoretical or practical goal [122]. In our case, we wanted to design a protocol for manually annotating PRs in textual educational resources which could be used to pursue multiple purposes, such as perform corpus exploration of PR instances as well as training systems to automatically extract those relations from educational materials. One might wonder why we should put time and effort into the manual annotation of texts, especially nowadays where language models reach state-of-art performances in most downstream NLP tasks without using any labelled data in training. Indeed, recent advances of unsupervised models might lead into thinking that the need of labelled data that we experienced a decade ago is now surpassed. However, even unsupervised systems rely on manually annotated data for evaluation purposes: consider for example the GLUE benchmark datasets [279] and the datasets produced for shared tasks. As a matter of fact, labelled data are still of some relevance: in addition to being useful to compare the performances of systems addressing the same task adopting different strategies, they can also be exploited to conduct (linguistic) inquiries about language structure and the realisation of the annotated phenomenon into the language [78].

However, when using annotated corpora, one should always be aware that their representation of the phenomena is limited by the choices made during the annotation phase [110]: if some phenomenon was beyond the scope of the annotation, it won't be represented. Take for example our scenario on PR annotation: ideally, a corpus annotated with prerequisite relations will be well-suited to study how concepts and instructional content are organised in it, but that same corpus will be useless for exploring, e.g., the interaction of students with the text (unless we annotate it). This might seems obvious, but it is actually a crucial point of annotation and has very strong implications on the data we produce.

Indeed, the way a phenomenon is formally defined has direct implications on its annotation, and consequently on the information that can be acquired when exploring the annotation, as well as on the accuracy of systems designed to automatically identify the phenomenon on unlabelled data. To put this concept in perspective, consider a scenario where we define the prerequisite relation as occurring between chapters of a textbook. Such type of annotation won't be able to represent how individual concepts are organised within a textbook unit (as the minimum component of the annotation is the book chapter), and a system trained on such dataset would as well show poor performances when addressing a more fine-grained task. As a matter of fact, annotation is an extremely delicate task that requires to deal with multiple issues involving clearly defining its scope and goals. In what follow we will discuss what should be done, as good

practice, when designing annotation tasks and which are the general desiderata for annotated corpora.

### 3.1.1 Annotation Protocols

In order to achieve our goal of obtaining a PR-annotated corpus we had to pass through a fundamental stage of annotation tasks: design an annotation protocol, i.e. define, though instructions and recommendations, how the annotation task should be carried out [230]. Given the great relevance of annotated data both to system training and corpus exploration, a certain effort has been put toward defining how to obtain good quality annotated corpora. We will now review some works discussing good practice for annotation task definition which have been incorporated into the canon of what is considered best practice for corpus annotation. Adopting standardised procedures when designing annotation protocols and tasks is of great relevance since it would promote accuracy of the annotations, speed of application of the annotation schema and faster analyses on the annotated datasets.

Leech [158] was among the firsts to define general recommendations for building annotated corpora. Focusing on the task of developing annotation schema, the author provides a set of rules of thumb to keep in mind in order to guarantee the success of the annotation project and the good quality of the final product. Leech recommendations are quite general and refer to high-level issues of annotation tasks. Specifically, they concern: (i) the annotation and corpus: the two must be separate and independent from each other; (ii) the annotation scheme: formally represented through symbols and corresponding definitions, it should be based on a theory-neutral analysis of the data and shouldn't be presented as the only way to encode the annotated phenomenon; (iii) the annotation project documentation: it should always be available for future corpus explorers. By following such general recommendations when designing the annotation task, one should be able to reduce the effort of both annotators, when applying the scheme, and corpus explorers, when using the annotated corpus for their analysis.

Similarly to [158], also [213] offer their recommendations for achieving consistent and good-quality annotations by listing which elements must be taken into account when designing annotation schema. This time, instead of presenting good practices for designing the task, the authors identify a series of issues that must be accounted for in the task development phase in order to produce good annotations when eventually applying the scheme. Since we took into account those principles when designing our PR annotation protocol, we will discuss how these are relevant to our scenario.

First and foremost, according to [213] an annotation task should *define which phenomena will be annotated*. This might seem trivial, but defining the scope of the annotation in advance and setting clear boundaries of what the annotation will encode and what it will not be included is of great importance. Subsequent uses of the annotated corpus will rely on the annotated dataset to perform supervised training for automatic extraction of the same phenomenon, or data-driven

analysis. Clearly declaring which information can be found within the annotation and which, on the other hand, is not included is essential. This also depends on the *corpus selected to perform the annotation*: depending the goal of the annotation, certain corpora could be more or less suited for the task. As an example, consider our scenario about the annotation of prerequisite relations: corpora broadly pertaining to the educational setting, such as textbooks, video transcripts of lectures, scientific papers, are clearly more suited for the task than, e.g., cookbooks, which in turn might be more appropriate to explore how local recipes changed along the centuries. A more detailed discussion on the desired characteristics of corpora will be carried out in the next sub-section. When dealing with the actual process of annotating the data, the authors recommend to carefully evaluate which *annotation tool* adopt since poor tools can have negative impact on both quality and quantity of annotated data. Considering the specific issues of our task, we developed our own annotation tool within the PR framework. Existing text annotation tools are reviewed in Section 3.3 of this chapter. Furthermore, the corpus might require *pre-processing* in order to remove mark-ups or make the text easier to annotated, which might also be done automatically or on the tool. After the manual annotation phase, the authors suggest to carefully evaluate the *inter-annotators consistency and intra-annotators homogeneity of annotations*. We deal in detail with this issues in Section 3.2.

Both [158] and [213] became well-known good practices of corpus annotation, integrated in the general practice. In particular, we see them incorporated into other models for annotated corpus construction, such as the general pipeline of annotation described by [122] and the MATTER methodology for creating annotation projects and applying them to machine learning algorithms [229, 230]. Both share some common ground in the sense that both present a general sequence of steps that must be carried out in order to obtain a good and re-usable dataset. However, while MATTER is specific to dataset creation for machine learning purposes, the pipeline is aimed at representing a general annotation process in NLP: it constituted an attempt to formalise the steps involved in the process of corpus annotation to call the attention to its methodological challenges.

While we employed MATTER as methodological framework for developing our annotation protocol, thus we discuss it in a dedicated Section, [122] pipeline draw our attention to certain issues that weren't addressed in the two works discussed above. In particular, we were inspired by the idea of writing a Manual (or Codebook) containing the annotation instructions. The *annotation manual* is intended as a resource explaining the motivation of the work and the underlying theory which provides the basis for the creation and definition of the annotation categories. Apart from being a valuable resource for re-applying the annotation scheme on novel unannotated resources, the manual also represent an instrument for protocol developers to keep track of emergent ideas about the annotation practice. Another issue outlined by [122] regards the level of inter-annotators agreement that should be considered as satisfactory for the task. Indeed, the desired level of inter-annotators agreement might vary depending on the

task or on the settings of the project where the annotation was carried out [30]. Knowing the agreement obtained on similar tasks could help in better evaluating the results obtained, as well as considering possible changes in agreement values due to the refinement of protocol principles.

#### 3.1.1.1 MATTER Annotation Cycle

MATTER is a general methodology for carrying out annotation projects aimed at obtaining annotated datasets that could be used in a machine learning experiment setting. It was first introduces in by Pustejovsky in 2006 as the 'Annotate, Train, Test' model [229], and subsequently expanded and revised in [230]. The classical workflow of MATTER is represented in 3.1.

As can be noted, the circular flow, which gave it the name of annotation *cycle*, is articulated into six phases. The first two (represented by the letters M – model – and A – annotate – of MATTER) deal with defining the annotation schema and actual corpus annotation process. In particular, the Model consists of the set of tags and attributes added to the corpus and their interpretation, whereas the Annotate phase includes all the actions needed to perform the actual annotation, such as recruiting the annotators, defining the guidelines and applying the model to the corpus. The following steps of the cycle are aimed at evaluating the effectiveness of the obtained annotation with respect to the project goal and, if not satisfied, revising the model and re-doing the annotation. Hence, phases 1 and 2 could be considered the most relevant of the workflow, or at least those having the higher impact since their outcome influences all other phases. This is why the model is frequently presented as a combination of two sub-models: the Model-Annotate cycle (MAMA), comprising steps 1 and 2, and the Training-Evaluation cycle, dealing with the Training, Testing, and Evaluation stages.

In principle, MATTER and its sub-cycles only provide a set of guidelines for the process of creating an annotated corpus and using it for machine learning techniques. Indeed, MATTER doesn't deal with issues of defining the context, setting and goal of annotation projects, which is why it can be easily integrated with other existing annotation standards presented above.

### 3.1.2 Desiderata of Annotated Corpora

Once the annotation protocol is defined and the task of applying it to the data is completed, we should have obtained an – hopefully good – annotated corpus. We will now review which are the desired characteristics that any good annotated corpus should have.

First of all, the collection of textual data that was used in the annotation project should have been chosen as *representative and balanced* with respect to the annotated phenomenon. A corpus can be deemed as representative and balanced with respect to a certain phenomenon if it shows a good approximation of the distribution of the phenomenon we aim to investigate into the real language use, thus allowing generalisation of the results [187, 203]. This is a critical issue of corpus linguistics since some advocate that corpora are only limited observations of the actual nature of the language and thus can't be used to draw general conclusions about phenomena

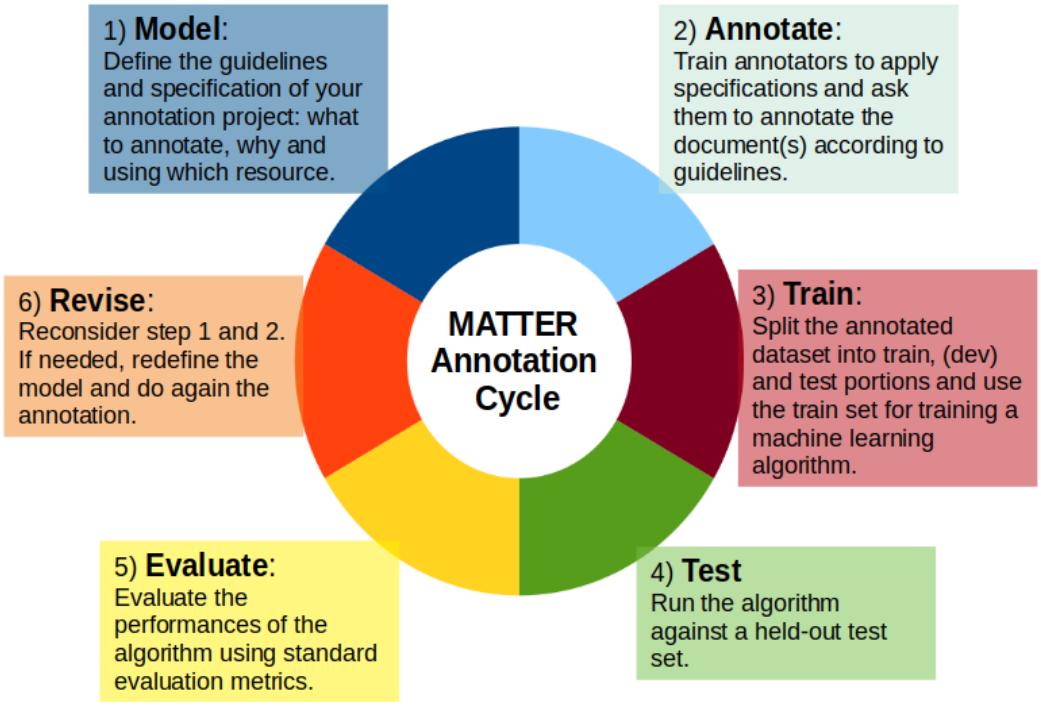



Figure 3.1: Classic workflow of the MATTER cycle.

(see, among the others, [61]). A large body of literature focuses on this topic, see for example [23, 35, 159, 253] among many others. Going into details about the discussion concerning the limits of corpora and how to obtain representativeness is beyond the scope of this work. However, it is worth mentioning that corpora are tools constructed by selecting and collecting data that should allow investigations about a specific phenomenon at the center of the inquiry. We need to keep this in mind when choosing the corpus to be annotated, as mentioned in the previous section, but also when exploring an annotated corpus: we must remember that they represent a specific point of view on the language use and knowing their building criteria is essential when exploring annotations.

Corpus building criteria might be included in the annotation documentation (sometimes also called annotation manual). The annotated corpus should always come with a documentation explaining choices and theoretical motivations behind both corpus construction and annotation schema, as well as instructions for carrying out the annotation [85, 158]. The latter are known as *annotation guidelines* and they are an essential tool to support dataset construction, exploration and use. Indeed, as pointed out by [85], annotation manuals (including the guidelines) must serve the needs of multiple users: those who use the annotated corpus to carry out linguistic explorations or develop NLP tools might be interested in the manual in order to understand which information (and how) in encoded in them. From the annotators perspective, guidelines

are essential instruments to learn annotation criteria. For this reason, guidelines must be clear, simple and precise in order to achieve the goal of reducing annotation discrepancies [297]. Some work addressing annotation in a highly specialised domain such as Biology, for example, showed that annotation discrepancies increase in the absence of guidelines, also if the annotation is performed by domain experts [65, 138, 265]. The consistency goal could be achieved by repeated definition, testing and validation of annotation guidelines, as suggested by [122, 230], or by measuring the appropriateness of annotators to the task [297]. As we will discuss further, we decided to include both the above practices in the definition of our PR-annotation protocol.

Another property that annotated corpora are expected to preserve is the independence between the original text and the annotation level(s). As claimed by [188], annotation is an added-value to corpora because it makes them usable for those who want to extract linguistic information from them, and also ideally multi-functional. As a matter of fact, annotation task might be designed to make explicit certain well-defined information but, once distributed, one might use the resource also for purposes not originally planned by its authors. In order to support the latter goal, it is important that the original text remains recoverable and that annotation is performed at consecutive stages and is accumulated as multi-level annotation. One of the most effective ways to easily distinguish the text level from the annotation level is to make use of *stand-off annotations* as opposed to *inline annotations*. These two represent the most common approaches to attach annotations to the text: while inline annotation is directly included into the resource, thereby changing the primary data (think, e.g., of a tokenized text), stand-off annotations are stored separately from the primary data they refer to [267], thereby leaving the primary data untouched. Although the choice of which approach is the best suited depends on the information to be annotated, the stand-off approach presents multiple advantages with respect to data reuse and information representation. First of all, having the annotation level and the text stored on separate files, one can apply multiple annotation layers to the same text (also at different times) without interfering with each other [177]. The multiple levels can be used to describe different complementary but co-existing information contained in the texts (e.g., syntax and semantics), fostering analyses bases on the interaction of such multiple annotation levels. Second, the annotation can always reference to the original text through pointers, but also to other annotation levels [302], which could be useful for, e.g., comparing different representations.

The above principles were formalised into the International Standards Organization (ISO) 24612 Linguistic Annotation Framework (LAF) [128, 129], developed over the past decade to provide a comprehensive and general model for representing linguistic annotations. This framework in large part simply brought together existing best practices from a variety of sources in order to allow for variation in annotation schemes while at the same time enabling comparison and evaluation, merging of different annotations, and development of common tools for creating and using annotated data. Defining community-standards for annotating textual resources was felt as a need since the 1980s: having a standard and commonly-shared representation of the informa-

tion improves re-use and interoperability of systems [125, 126]. Indeed, not having a commonly agreed definition and representation of prerequisite relations makes different automatic-PR learning systems difficult to compare (see Section 3.4). Although community standards for highly specific phenomena are difficult to reach, there are some standard representation of linguistic information. Review them in detail is beyond our scope, but in the next sub-section we will briefly introduce one of the standard formalisms to represent morpho-syntactic information, Universal Dependencies, as we adopted it in our research.

**Universal Dependencies Initiative** The Universal Dependencies (UD) initiative[1], originating from the evolution of pre-existing tagsets [77, 222, 294], is a project aimed at developing cross-linguistically consistent treebank annotation for many languages. It is currently the the most prominent linguistic representation formalism, with nearly 200 treebanks in over 100 languages [296]. Since the project offers a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages [207], it has prompted numerous multilingual studies (see, e.g., [68, 277, 295] and, among our works [8, 9, 11, 12]). The large number of gold resources annotated at the morpho-syntactic level under a shared annotation schema provides useful linguistic evidence for use in NLP tools [225].

The format employed within the UD project to represent linguistic information is a revised version of CoNLL-X format [48], called CoNLL-U, a column-based format where each sentence is separated from the following by a blank line and each token of each sentence is represented on one line through the 10 fields, separated from each other by a TAB[2]. The UD formalism was chosen to represent the linguistic information underlying PRs for two main reasons: (i) it is the most prominent formalism currently used for linguistic analysis, and (ii), being multilingual, potentially allows multilingual analysis on PRs and the development of multilingual PR annotated datasets.

## 3.2 Approaches for Annotation Revision and Agreement Evaluation

Once the annotation process is completed, the effort is generally addressed towards the creation of a Gold Standard Dataset to be used in future analysis. Gold standard datasets are ground truth annotated datasets intended to provide a generally accepted annotation of a phenomenon that can be looked at as accurate and reliable reference [286]. Given the importance of such benchmarks for both manual analysis and automatic system training [94], in order to maximise coverage these resources are usually built by a pool of experts by performing a shared annotation or, more frequently, by combining their single annotations. Only rarely the annotation of a single expert is used as reliable reference, possibly when the phenomenon to be annotated is highly

[1] `https://universaldependencies.org/`

[2] Refer to `https://universaldependencies.org/format.html` for more information about CoNLL-U representation.

subjective and extremely hard to identify. In order to improve gold datasets quality, preliminary steps to their creation are aimed at refining and evaluating individual annotations, as discussed below.

### 3.2.1 Annotation Errors and Revision

Manually annotated corpora are known to be error-prone, as well-recognised in the literature [84, 98, 123, 201]. Also so-called gold standard datasets encoding well-defined linguistic information contain a consistent amount of errors: the POS assignment in the widely used Wall Street Journal corpus [183], for example, has an estimated 3% error rate [192]. Errors in annotation could be due to multiple reasons, either caused by annotators distraction, misunderstanding of the guidelines, gaps in the annotation specifications or genuine ambiguity of the data. The negative impact of even few errors in benchmark datasets has been shown to create problems for both computational and theoretical linguistic uses of annotation [120], from unreliable training and evaluation of NLP technology [10, 148, 271] to low precision and recall of queries for already rare linguistic phenomena [192].

As discussed in 3.1 above, clear and complete annotation manuals, dealing with as many potentially critical issues of the annotation as possible, are one of the most effective instruments to prevent errors as they provide guidance to annotators during the labelling process [97]. However, also when the project is accompanied with high-quality guidelines, some instances of errors might be introduced. [224] distinguish two types of annotation inconsistencies, whose boundaries are difficult to determine in most cases, namely *proper errors* and *hard cases*. The former are items annotated incorrectly according to the guidelines, thus it is possible to identify their correct annotation, which will be different from the one assigned by the annotator. In our scenario dealing with PR annotation, an example of proper error can be found when an annotator inserts a self-prerequisite (i.e., $A \prec A$): since PRs are dependency relation, it is not possible that a concept is prerequisite of itself, hence this type of relation must be removed. Cases belonging to the latter group, namely hard cases, are usually more difficult to find and correct. In [32], hard cases are defined as instances which are difficult to decide upon because they happen in cases where annotator preferences come into play. What makes them difficult to handle is the fact that such annotations are not necessarily incorrect, but can be naturally ambiguous or not covered by the guidelines. PR-wise, such error types could correspond to distant transitive relations. In our scenario, by annotating PRs based on the content of educational materials, the annotators' interpretation of the texts could either lead to explicitly express the presence of a relation between two concepts through the creation of a PR or to leaving the relation as a transitive implicit relation because too weak. Indeed, such cases may have multiple correct answers, depending on the subjective interpretation of the instance context, or it may even not be obvious what the label should be. An annotation revision step, to be carried out after the annotation is completed, could allow to identify errors, correcting them and thus improve the

overall homogeneity of the annotations [224].

### 3.2.2 Metrics for Agreement Evaluation

The reliability of manually labelled data is usually measured through metrics aimed at capturing how many items received the same label by different annotators, either as a percentage over the total number of items (raw agreement) or accounting for the agreement expected by chance or arbitrary coding (inter-rater reliability measures) [19]. Indeed, it is a common practice to evaluate annotation reliability be means of annotation consistency: if different annotators produce consistently similar results, then we could claim that they share similar understanding of the annotation guidelines and we can expect them to perform consistently under this understanding [19, 20]. Consequently, measuring consistency can help evaluate both guidelines clarity and annotators performance. However, disagreement among annotations produced by different annotators is natural to occur: it is very rare, and for some tasks even almost impossible, to have multiple human annotators to completely completely agree with each other on what and how to annotate [123]. Disagreement might occur between annotators due to involuntarily included errors (see 3.2.1), or difference between annotators knowledge and experiences [123]. We also commonly observe intra-annotator disagreement [70], namely discrepancies occurring within the annotation of the same annotator, possibly due to a decreasing level of interest and motivation may drop and level of fatigue rises as the annotation process continues [112]. Although raw agreement captures fairly well a dataset characteristics and identifies potentially problematic areas in the annotation, inter-rater reliability measures are considered more robust since they account for the possibility of agreement occurring by chance.

Among inter-rater reliability measures, Cohen's Kappa coefficient ($k$ [64] is the most widely adopted in linguistic and semantic annotations to compute the agreement between pairs of annotators. The value of $k$ is computed according to the following equation:

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{3.1}$$

In equation 3.1, $P_o$ stands for the *observed agreement*, i.e. the probability for an item to receive the same annotation by both raters, computed as a ratio between the items receiving the same label by both annotators and the total number of items in the annotation task. $P_e$ denotes the *agreement expected by chance*, i.e. the probability of each individual category, computed as the number of items receiving one or the other label over the total number of items in the task. When the annotation is performed by more than two raters, Fleiss' $k$ should be used since it accounts in equation (3.1) for the number of rater-rater pairs agreeing on each item [95].

Although being so popular to become a *de-facto* standard for manual annotation evaluation, $k$ coefficients are affected by well-known issues regarding the annotations and the metric interpretation. About the former, [90] points out that $k$ value is strongly affected by the distribution of the

phenomena into the set of items being annotated: $k$ is computed as a ratio between probabilities, so if the distribution of categories in the data is skewed, i.e., with prevalence for some categories, the expected probability is higher and thus the $k$ score lower. This condition is called "prevalence problem" and could be caused by *i)* a tendency for raters to identify certain labels more often than others, *(ii)* a quality defect when building the dataset or *(iii)* a truly unequal frequency of the labels. By contrast, the "bias paradox" in Cohen's metric causes $k$ to obtain untruthfully high values when the two raters produce substantially different label distributions in their annotation, as it happens, e.g., when one rater largely uses one category while the other rater applies it less frequently [51]. Both prevalence and bias issues might come across as a problem when dealing with PR annotation: due to the nature of PRs, the number of prerequisite relations between the concepts of, e.g., a lecture is much lower than the number total of all possible relations between the mentioned concepts. There were attempts to address these limits [33, 247, 252], but to date none corrects for both prevalence and bias paradox problems.

The qualitative interpretation of the $k$ score is also a critical issue. The scale used by [150] proposes different thresholds of $k$ to evaluate agreement strength, ranging from 'poor' to 'almost perfect agreement', but which threshold should represent the lower bound for guaranteeing reliability of the data is debated [147]. For example, restricting acceptable agreement values to those above the conventional cutoff point of 0.80 (with 0.67 to 0.8 tolerable) is claimed as too restrictive by the NLP community, especially for semantic annotations, although still widely adopted [236]. Furthermore, $k$ metrics are well designed for categorical data, but they don't fit more qualitative and grained types of annotation; in addition, they are not suited for the annotation of open sets of items [285, 287]. To address these limits, [56, 216, 283] tackled the task of PR annotation between pairs of concepts as a labelling annotation task using a small set of predefined categories on a given set of items (i.e., pairs of concepts): given a pair of concepts, decide whether or not they show a PR, which allows the straightforward use of equation (3.1). Our PR annotation approach, on the other hand, is not based on a pre-defined fixed set of items to be labelled, as annotators have to identify PR relations while reading the text. This requires to adapt $k$ and to account for the transitive property of PRs which becomes relevant in our scenario.

## 3.3 Tools for Text Annotation

Manual text annotation is frequently performed exploiting annotation tools which provide an interface simplifying the process of adding labels to texts. According to our knowledge, the only interface designed to specifically address the issues of annotating prerequisite relations was developed by [109]. In their scenario, the interface aimed to facilitate the manual evaluation, performed by experts, of candidate domain topic pairs matched by an automatic system as showing a prerequisite relation. For each topic, the interface shows the list of related topics in grey scale (i.e., most relevant items coloured in black, fading to white as the association strength

with the topic decreases) and the documents where they are described. Experts are then asked a set of questions to evaluate the coherence of the topics and the strength of association between the topics in the pair. Although this can be of some support for the annotation, it can't be exploited in PR annotation scenarios where, contrary to [109], annotators have, e.g., to annotate prerequisite pairs among concepts directly on educational texts.

Choosing or defining the most appropriate annotation tool to accomplish an annotation project is actually a delicate matter because of the impact that the interface might have on the annotation process. Currently, there is a wide variety of annotation tools available: some of them are general-purpose and try to address the most common needs of annotation projects, while project-specific ones were designed to meet the requirements of a specific project [94]. A set of general requirements of annotation tools were discussed by [86]. In broad terms they match with (but also were expanded by) the set of commonly found features of annotation tools outlined by [94]. Even without going into detail about their lists of requirements, we can say that they refer to the ability if supporting multiple types of data (written or spoken), multi-layer annotation, inclusion of new tagsets, annotation quality evaluation, corpus analytics and simplicity of use. While a tools showing such properties might be suited to support the fundamental requirements of many annotation projects, general purpose tools usually miss some key functionalities, leading researchers to modify existing tools or even develop new ones to support their needs and thus reduce time and effort for creating annotations [237]. Such tools are in opposition to crowd-sourcing platforms (e.g. Amazon Mechanical Turk, Figure Eight), usually employed when a large number of annotators is required and the annotation task is guided by simple rules or common sense [89]. Yet, a crowd-sourcing annotation approach needs to be replaced by single or multi-user annotation tools when experts' knowledge is crucial to properly recognise a phenomenon [254], as it happens in the case of PR annotation.
Hereafter, we review a selection of commonly used tools for manual text annotation that inspired our work and highlight the reasons that prompted us to design our own PR-annotation interface. For more comprehensive surveys, refer to [36, 94, 186, 204, 245].

BRAT [258] is currently the most prominent tool for creating annotated corpora: it is an open-source web-based system integrated with NLP technology and optimised to support the annotation of local relations between spans of text, such as dependency structures. Being capable of a handling a wide variety of annotation tasks, BRAT has become the first choice for many annotation projects (see e.g. [226, 231, 272]). Nevertheless, some of its features made this tool not suitable for our scenario. PRs are a specific type of dependency relation that, emerging from the content of a textual instructional resource, can occur also between concepts that do not share the same textual context. This means that connecting close spans of text in the text is limiting for us as it misses some cases of prerequisite relationships that we want to identify. Moreover, we want to save in the annotation the coordinates of each PR, i.e. for each relation, we want to be able to retrieve where in the text the annotator decided to enter it. Instead, BRAT tool returns

only the annotated pairs without providing information about, e.g., in which sentence of the text the annotator added them. Lastly, although creating a gold dataset is a common procedure of annotation projects, BRAT doesn't provide an adjudication interface to combine multiple annotations in order to obtain a gold standard dataset or to compute agreement measures, which are essential in most annotation projects. The latter two features were implemented in Webanno [75, 290], a general purpose open-source web-based annotation tool that exploits BRAT visualisation and supports the same set of annotation types. Other missing features of BRAT implemented in Webanno include web–based configuration of the tagset and agreement computing.

Some annotation tools support the annotation by integrating modules for semi-automatic annotation of data. This is the case of Tagtog [55], a web based collaborative annotation tool for entities and relations. The tool is presented as particularly suitable for annotating large texts as it allows to perform automatic annotation through Machine Learning while supporting also the human revision phase. It must be noted that automatically annotating data, if on one hand makes the annotation process much faster, on the other hand suffers from the anchoring effect bias [268]: when revising the automatically produced annotations, humans' opinion tend to flatten towards the automatically assigned labels. For this reason, automatic annotation is recommended only when its accuracy is known to be generally high. Moreover, semi-automatic annotation is not useful in cases of medium-small datasets, where letting annotators freely express judgements results in much richer data. Tagtog also supports the automatic creation of a dictionary by assigning an *id* to each entity (concepts, in PR annotation) in the text, thus conducting all synonyms to a unique base form. We argue that such feature is helpful mostly when annotating non-scientific texts, such as news, where entities can be referred to in many different ways. Specialised languages are by nature less ambiguous in the use of terms [52], thus this feature is less relevant in our scenario or even misleading in some cases: terms that can be considered as synonyms in non–specialised contexts might not be real synonyms in a specialised one (e.g. *network* and *web*).

The aforementioned tools are designed to ideally support any annotation task, but we observe that they can't be customised or extended, nor they address the specific requirements of PR annotation, as it frequently happens in highly specialised tasks. Additionally, but not less important, a complete annotation tool should support also corpus analytic and annotation pattern analysis, i.e. provide quantitative information about the annotated corpus (e.g. number of relations and in which part of the text they appear) and comparisons between multiple annotations performed by different experts or automatically extracted. Such analyses are highly project-specific and it's hard to find a tool which perfectly fits the needs of each project.

## 3.4 Datasets Annotated with Prerequisite Relations

The raise of interest in Artificial Intelligence for automatic prerequisite learning has fostered the development of datasets annotated with labels expressing the presence of a prerequisite relations between two concepts. Such datasets are valuable resources for training and testing Machine Learning algorithms (see Sec. 3.5) or to validate whatever kind of extraction method against a gold dataset [1, 165]. Despite being time consuming, creating manually annotated datasets is an effective practice and produces gold resources annotated with PRs, which are still rare and mostly limited to the English language, with the exception of two Chinese datasets [176, 301].

Most of existing PR datasets consist of pairs of educational concepts enriched with a binary label expressing the presence or the absence of the prerequisite relation [56, 109, 283]. The PR representation usually obtained by such resources resembles knowledge graphs where educational concepts are represented by nodes and PRs are represented as graph edges. Indeed, the final goal of those annotation projects is generally to build knowledge structures representing a certain domain knowledge [91, 283] rather than encoding the information contained in educational resources, which is instead our perspective on the PR annotation task. We notice that educational resources used to build such pairs are generally acquired from two distinct types of data: *i)* course materials (e.g., MOOCs [56, 215, 216, 240, 299] or university websites [162, 167, 289]); *ii)* educational materials in a broader sense, such as scientific databases [108] and, more frequently, Wikipedia pages [103, 193, 244, 264, 301]. Using external resources to build PR-annotated datasets, as in the works mentioned above, could be effective if the goal is acquire domain knowledge since using knowledge bases tends to bind PRs to their ontological relations in the subject domain. However, this approach might return poor results when applied to domains not well covered by the external knowledge. Textbooks, on the other hand, are rarely used for PR annotation [14, 149, 176, 283], possibly due to the challenges of interpreting the author's didactic choices. However, it should be noted that textbooks are self-contained, meaning that they cope with every concept a learner has to know in order to understand the book content. Since this is exactly the goal of our annotation (i.e., modelling the content of educational materials), we argue that textbooks are instead one of the most suitable resource for PR annotation.

With respect to PR manual annotation, we see a tendency towards the manual validation of all pairwise combination of pre-defined concepts [56, 162, 283, 301] or of a random sample of that set [103, 108, 216]. Asking annotators to autonomously create concept pairs based on their knowledge about the topic, as in [176], is less common and mostly employed for modelling domain knowledge rather than educational resources. To the best of our knowledge, [264] is the only case where crowd–sourcing is employed for PR annotation. Here, the authors acquired candidate PR by exploiting hyper-links in Wikipedia and use crowd-sourcing to validate those relations. In most cases, the annotators recruited to perform the task of PR annotation are domain experts [91, 165, 166], or students graduating in the field [215, 281, 301] since domain novices are generally not suitable for the task [14].

Among the datasets mentioned above, the one presented in [283] (further expanded as AL-CPL dataset in [168]) is the one we consider closest to our work as it shows prerequisite relations between relevant concepts extracted from textbooks, thus we share a similar level of granularity. The dataset consists of a manually constructed set of binary-labelled concept pairs collected from English textbooks on different educational domains: data mining, geometry, physics and precalculus. Concepts are retrieved from textbooks as domain terms matching a Wikipedia page title, while prerequisite relations between them were annotated by three domain experts based on their background knowledge. The final annotation allows to produce a concept map for each of the four domains, a knowledge structure that represents key concepts of subject matter organised by means of pedagogical relations (here, prerequisite). In AL-CPL expansion, the dataset was augmented in order to feature also negative, irreflexive and transitive pairs automatically acquired from existing pairs [168].

The presence of different annotation approaches and the lack of guidelines defining good practices to encode prerequisites brought to the creation of datasets that are not easily comparable and that capture different aspects of the relation. Indeed, the number of PR annotated resources is still too limited, possibly affecting the advancement of the research in the field.

## 3.5 Automatic Prerequisite Relations Learning

Automatic Prerequisite Relation Learning (hereinafter in this section referred to as APL) consists of acquiring prerequisite relations between educational items (concepts) using automatic strategies. APL is usually exploited to acquire the knowledge structure of a domain with the purpose of automatically build concept graphs reflecting the prerequisite structure of the concepts in the subject matter [56, 109, 174, 215]. Such structured representations of knowledge can support many educational applications. Currently, APL has been applied to curriculum planning [4], course sequencing [276], reading list generation [91, 108], automatic assessment [282], domain ontology construction [152, 303] and automatic educational content creation [175].

Just like the learning process involves (at the minimum) the content to be learnt and a learner, existing approaches dealing with APL address the task by leveraging information either referring to educational materials or to their users. Next sections will review approaches falling into each of the two categories.

### 3.5.1 Leveraging Learners Data

A line of research on APL is aimed at estimating prerequisite structures from students behaviours [59] or acquired skills (i.e. knowledge mastered by the student) as collected from the student interaction with Intelligent tutoring systems [45, 115, 278]. Such models are based on the assumption that changes in a student's knowledge (e.g. acquisition of new concepts) can be inferred from students' performances during assessment events [143, 246]. As early application

of this idea, [87] proposes to infer prerequisite graphs based on the results obtained by student on multiple tasks (questions, problems solving, etc.): if a student is able to correctly solve a task A (e.g., finding least common multiples) and a task B (e.g., adding fractions with unlike denominators) but not the other way around (i.e., many students that can find common multiples fail at adding fractions), then A must be prerequisite of B. Subsequently, the performances of learners when testing their skills were used to create student models representing an estimate of skill proficiency at a given point in time [83]: since prerequisite concepts should be acquired before advanced ones, student models can be used to estimate precedence knowledge acquisition, i.e. prerequisite relations [60].

Another line of research that deals with modelling students knowledge while they interact with coursework or learning materials is *knowledge tracing* [66] or, as it was called since machine learning methods took over classical Bayesian approaches [219, 292], Deep Knowledge Tracing [223, 288]. Knowledge tracing in some ways is close to student skill modelling, as it addresses a similar task, but it also offers a complementary perspective since the two diverge in terms of scope: while capturing skill acquisition level at a certain time is generally used to trace back what a student has already learnt, modelling student knowledge over time, as in knowledge tracing, is aimed at predicting how students will perform on future interactions in order to provide personalised new content. As a consequence of the different perspectives, PRs can be used in knowledge tracing as constraints when modelling students' knowledge [58], but not as an outcome of the modelling task.

Although taking students' behaviour into account when performing APL is an attractive perspective, such type of information is not available unless we capture it exploiting, e.g., learning management or intelligent tutoring systems. Instructional events through learning systems represent only a fraction of the overall learning experiences, thus restricting APL only to the data collected in those situations is limiting. Another known challenge of such approaches is that students' data are generally sparse, namely they result from irregular use of the tutoring system, thus it is quite difficult to use them to generalise and accurately represent the domain knowledge [58]. In order to handle those cases, one could rely on a APL approach to leverage the information from educational resources rather then their users. This represent also our approach for tackling the task, as we will discuss in chapter 8.

### 3.5.2 Exploiting Instructional Resources

Different sorts of instructional materials have been used to train and evaluate APL systems. Among them, Wikipedia is undoubtedly the most used resource [104, 165, 166, 264, 301], possibly because of practical reasons: it is freely accessible, heterogeneous in terms of topics covered, widely used and multilingual. However, we see also works leveraging other types of resources, such as scientific papers [109, 162], knowledge units of MOOC courses [56, 162, 240, 289], transcripts of video lectures [6, 26, 215]), Learning Objects repositories [103], DBPedia [181] or

textbooks [281]. Depending on the resource considered, the APL model might address the task relying on a different strategy.

#### 3.5.2.1 Graph-Based Approaches

A line of research dealing with APL exploits graph-based approaches [109, 134, 167, 181, 289]. As said, knowledge graphs can be used to represent domain knowledge and dependency relations between concepts by depicting concepts as nodes and relations as edges. These methods exploit graph theories to infer unknown PR edges from knowledge graphs. For example, [167] proposed an optimization based framework to discover concept prerequisite relations from course dependencies; [216] developed a graph-based propagation algorithm to order latent representation of concepts automatically acquired from video transcripts of MOOC courses. Other works exploit directed graphs [173, 215] or multidimensional knowledge graphs [251]. Although graph-based methods show great potential as they generally reach good results, they rely exclusively on formal properties of the graph topography and do not take into account the information contained in the content of the resources. As most educational resources assume the form of textual materials (think, e.g, of textbooks, but also of Wikipedia pages), NLP approaches can be exploited to overcome this limit and leverage information contained the content of instructional materials [56, 103, 109, 162, 282].

#### 3.5.2.2 Resource Content and Structure

We notice that information extracted from educational textual materials are used either to define features to train APL models based on machine learning [103, 173, 182, 264] or exploited to define unsupervised approaches that do not need to learn from labelled examples [1, 161, 165]. For the purposes of this dissertation, we are mostly interested in discussing the different types of information used to uncover PR rather than going into detail about the employed models. For this reason, the remaining of the discussion will be focused on presenting existing strategies for tackling APL distinguishing the type of information each considers informative to uncover PRs.

Indeed, regardless of the employed model, we distinguish between information that can be acquired from the raw text and information that refers to formal and structural properties of the resource. While the former can be acquired from plain text, as we will discuss below, the latter need to access some extra-textual information of the resource in order to define PRs predictors. What can be considered as a structural property, however, depends on the resource we are looking at. [283], for example, dealing with textbooks, exploits the book organisation into chapters and sub-chapters to create a structured dependency-based representation of text portions. By relying on such representation, the authors propose to uncover PR by computing the relation strength between concept pairs as a distance between the sub-chapters where they appear (e.g, if the concepts ‘addition’ and ‘multiplication’ are mentioned in chapters 3.1 and 3.2 of the book respectively, their distance will be equal to 1). As apparent, this method can

be applied only if relying on a resource that shows a table of content or, at least, chapters and sub-chapters that can be used to recreate a hierarchical structure of the resource content. For [109], performing APL on the ACL Anthology, scientific papers can be organised into a relational structured representations by considering citations and papers similarity (in terms of overlapping content). Knowledge bases such as Wikipedia or ontologies are by definition structured knowledge representations by means of manually entered relations (in the case of ontologies) or hyperlinks and categorical structure (in the case of Wikipedia): this makes the acquisition of structural properties even more straightforward. As a matter of fact, there is a whole line of works dealing with APL that rely on Wikipedia graph and categorical structure to extract concept relational features [79, 103, 282, 301].

Among the works exploiting Wikipedia, [264] was the first to adopt machine learning to tackle APL. The authors of the paper presented a Maximum Entropy classifier to predict prerequisite relations between Wikipedia pages by exploiting three types of features: Wikipedia hyperlinks features (i.e. random walk with restart (RWR) score between two pages and PageRank score), edits of pages (i.e. RWR score on edit information), page content (i.e. category assigned to the pages, presence of a link between them, mention of concept in the page text). Similarly, [103] use Wikipedia's hierarchical category structure and hyperlinks as features for a Multilayer Perceptron classifier. [301] as well experimented with different classifiers trained with a rich set of features capturing concept relatedness by exploiting links between Wikipedia pages, overlap of pages categories and of pages content. Contrary to previous methods, [165] did not exploited a machine learning method, but still relied on Wikipedia to acquire information. In [165], the authors presented a very intuitive yet robust link-based metric to uncover PRs based associative strength between concepts, the Reference Distance (RefD) metric. As the name suggests, the focal point here is the notion of (co-) reference: indeed, RefD models the prerequisite relation by measuring how differently two concepts refer to each other using TF-IDF. Although the RefD is thought to be quite generic and can be computed considering mentions in books, citations in scientific papers, etc., the experiments described in the paper present a Wikipedia-based Ref-D implementation exploiting hyperlinks between the Wikipedia pages of the concepts and computing TF-IDF on pages content. The RefD metric was also used by [283] in a method that jointly extracts relevant concepts and prerequisite structure from textbooks exploiting also external knowledge from Wikipedia.

What is common to the above methods is that they all require to access the explicit structure of the resource. Although they do take into account the information contained into the pages (to compute, e.g., TF-IDF and to acquire concept mentions), such information is always associated with other features coming from the resource hierarchical structure (links, metadata, page history). If on the one hand these methods are to be distinguished from the graph-based ones discussed above as they take advantage of both the pain text describing a concept and structural features of the dependency representation of concepts, they still mostly rely on structural features,

also reported as the most informative [166, 167]. However, this condition is neither the most common or the most natural. First of all, there are some well-known limits related to the use of Wikipedia for acquiring domain knowledge [113]. For example, emerging fields (for with students might be more interested in finding resources) tend to be poorly represented in Wikipedia since it takes time before the community starts building a consistent number of pages referring to the topic, thus obtaining a solid representation. Plus, the actual coverage and quality of Wikipedia are frequently questioned. Moving to a different knowledge structure won't solve the problem either as structured representations might be rare for some domains. Moreover, students dealing with a new topic generally relies only on the information that can be leveraged from the content (s)he's reading. Hence, automatically uncovering PRs relying exclusively on the content (plain text) of the educational material is a challenging, though undoubtedly possible, scenario.

#### 3.5.2.3 Plain Text Information

Models that rely on plain text information to acquire PRs have the advantage of returning a PR structure that reflects the actual content of the educational resource taken into account. As a consequence of leveraging information exclusively on the textual content, relations not present in the text won't be acquired. For what concerns the work related to the present dissertation, a consistent part of the research has faced the task of automatically acquiring PR from plain texts as we identified it as a quite neglected issue. It goes without saying that the work represented a joint effort with other researchers: we developed solutions which were presented to the research community to drive the attention towards the potentiality and the benefits of the in-text setting of APL. This section won't discuss our approaches, but they will be addressed in the last chapters of this thesis. The remainder of the section will, on the other hand, discuss the textual information employed by other models, although frequently combined with structural features.

The most simple approach when searching for PRs in plain text is relying on lexical and lexico-syntactic patterns [283]. The patterns are aimed at identifying a lexical relation between two terms, which might underpin a PR. Consider as an example the pattern "NP2 such as NP1", that might appear in, e.g., a computer science text instantiated as "Wireless networks such as 4G". According to the example, "4G" is related to "wireless network" via the "such as" pattern, which reveals that 4G is a specific type of wireless network. In most cases, showing a "type-of" relation overlaps with having a prerequisite relationship: in order to understand the most specific item (4G in the example), the student first has to master the knowledge related to the super-ordinate item (wireless network). Clearly, pattern-based methods are not robust as they require to manually define the patterns in advance (which might apply quite well to a resource, but fail on another) and they also miss to identify relations between concepts not co-occurring in the same sentence. Relying on external lexical resources, such as WordNet[3], to acquire hyponyms-hypernyms might turn out effective to overcome these limits, but, as above, the

[3] `https://wordnet.princeton.edu`

results might drift away from the content of the resource and return ontologically valid relations not actually expressed in the considered text.

As mentioned, unfortunately exploiting only plain text information is a task frequently neglected by the existing literature on APL. Indeed, features based on lexical overlap between pages [108, 282] (measured as shared terms or using similarity measures) and concept co-occurrence [109, 165] are combined with the structural features described above. Co-occurrence, for example, is at the core of many approaches for PR relation identification [165]. However, while co-occurrence is an intuitive condition for PR, high co-occurrence is not necessarily a measure for PR strength, since it could identify also other types of relations, such as taxonomic relations, complex relations, general associations or co-requisites. Therefore, a reasonable assumption is that co-occurrence of two concepts is likely a necessary but not sufficient condition to identify a prerequisite relation that needs to be further refined with other information. In general, in fact, high co-occurrence frequency (i.e. counting how many times two concepts occur together in a certain span of sentences) seems a good indicator of relatedness, thus it could underpin other kinds of relations besides PR.

Before ending the section, it is worth mentioning the set of features proposed by [166] to train their APL model. In this work, the authors defined a set of proper textual features to be acquired from the content of Wikipedia pages to be combined with graph-based features to train a PR classifier. In addition to simple mentions, the authors rely on topic modelling performed on pages, Jaccard similarity and embedding representation of concepts acquired using Word2vec [194]. We feel this work is the most closely related to our approach, and we also took inspiration from the their set of features when designing our model.

## 3.6 Educational Applications Exploiting Prerequisite Relations

In this Section we will provide a brief overview of selected educational applications enriched with prerequisite relations knowledge.

*Automatic lesson plan generation* is possibly the most straightforward application of PRs [26, 169]. Within this line of research, [4] proposes Socrates, a tool for automatic synthesis of study plans. Given a set of concepts, Socrates is able to organise concepts respecting the prerequisite constraint in order to obtain effective study plans exploiting a graph-based approach on the prerequisite concept graph. Socrates was tested in a study involving 193 Physics concepts and the automatically produced study plans where deemed as good quality by a pool of expert teachers, although the authors say that more extensive testings should be performed. More recently, [71] developed a model for identifying a prerequisite-aware curriculum plan aimed at acquiring the knowledge required to fit into the ideal profile of specific job offers.

Similarly, *customised learning materials generation* deals with automatically building novel educational resources that reflect the prerequisite structure of the domain. [164] propose to build

personalized learning resource similar to a textbook using BBookX system which automatically collects and organizes online resources related to user input. [108] presented a system for generating reading lists based on inferred domain structure and learner models.
In [282], prerequisite concept maps are used to perform *automatic assessment*, namely assessing learning achievements and providing feed-backs to learners. The system, enriched with a prerequisite concept map automatically extracted relying on textbooks and Wikipedia, provides learners with questions about concepts in the map. If a student answers correctly to the questions, the systems recommends to move forward in the map to a more advanced concept, otherwise the learner is recommended to revise some prerequisite concept in the map.

CHAPTER 4

# METHODOLOGICAL ISSUES OF UNCOVERING PREREQUISITE RELATIONS FROM EDUCATIONAL TEXTS

Previous Chapters discussed background and research related to our work. In particular, we dealt with the definition of *concept* and *prerequisite relation* in Chapter 2 and presented a literature review on textual annotation, prerequisite relation annotation and automatic extraction in Chapter 3.

It should be clear at this point that uncovering, annotating and extracting PRs are not trivial tasks. As a consequence of such difficulty, we notice a lack of consensus on the nature of concepts and rough definitions of annotation tasks, which frequently leave unspecified the distinctive features that can be used as clues to identify PRs. As a result of the aforementioned, we identify the following main consequences: *a)* low agreement values between annotators [56, 91, 109], *b)* difficulty to directly compare existing datasets and *c)* performance variability of systems trained on such data [16]. The main causes of points *a)* and *b)* could be traced in fairly basic and vague PR annotation guidelines, which leave annotators with naive and innate definitions of the PR relation, leading to heterogeneous annotations. As a consequence, different annotation initiatives produce datasets encoding the same phenomenon according to different principles. Despite disposing of diverse datasets might be useful to test the robustness of, e.g. an automatic PR extraction system, the lack of consistency in annotation shows that the community working on PRs misses a commonly shared definition of the phenomenon and standard procedures for dealing with it.

We try to fill this gap in the literature by proposing a *novel methodology for uncovering PR relations in textual instructional materials. The* ***goal of our methodology*** *is to provide a systematic approach for modelling concept relations from the educational point of view, thus supporting the development of a shared interpretation of PRs among the research community*.

The methodology was systematised in the PR Framework introduced in Chapter 1.4 and further discussed in the next Chapters, which comprises multiple components, each dealing with a different issue of PR identification. The first issue we tackled when developing our methodology and framework was the manual annotation of PRs in instructional materials.
Note that, from now on, when we mention 'instructional materials' we refer to educational resources in form of textual materials, if not otherwise specified.

As discussed in 3.1, developing annotation protocols is a challenging task that requires, first of all, to clearly define the phenomenon to be annotated and how the annotation should be carried out. As we already stated when introducing prerequisite relations (2.2), among all interpretation of PRs we are mostly interested in the view proposed by the *pedagogical perspective*, which focuses on modeling resource content rather than domains. In the following Sections, we will explore the issues, benefits and challenges relating to the adoption of the pedagogical view to uncover PRs in instructional texts as opposed to the ontological perspective. Specifically, Section 4.1 will discuss our perspective on the task of prerequisite identification and the challenges it poses; Section 4.2 will present the results of crowd–based experiment where we investigate the influence of language complexity on the task of manual PR identification; Section 4.3 introduces our novel methodology by discussing how we incorporated the results of the crowd-based experiments and the issues we addressed. In the last Section summarises the chapter.

## 4.1 Tracing Prerequisite Relations with the Pedagogical Perspective

As discussed in Section 2.2.2, there are two main paradigms which deal with the relationship between prerequisite relations and domain knowledge: the ontological and pedagogical views. The methodology described in this dissertation relies on the latter.

The basic idea behind the pedagogical view is that there are multiple valid ways to organise concepts within a subject domain. These are generally reflected in instructional materials, whose content ultimately mirrors the author's view about the domain structure and how it should be taught. It is true that, for some subject matters, teachers and the authors of learning materials tend to introduce concepts to students according a predicable order. For instance, this happens in foreign and second language teaching, where the sequence of acquisition is an ascertained notion used for describing a possibly fixed and universal order in which all learners tend to acquire grammatical features of the target language [146]. However, this predictable and shared order is not common to every subject: the content of instructional materials is generally organized on empirically based design strategies and not according to conventional discipline structures [171]. As a matter of fact, although the teaching of a subject may eventually experience a process of standardization, the order of contents to be presented is still largely a matter of the author's preference [266] (see the example on the two Computer Science textbooks in Section 2.2.2). Hence,

based on the pedagogical view, it is more appropriate to model the prerequisite structure of instructional resources rather than creating absolute and resource-agnostic knowledge structures.

However, the ontological view is by far the most widely adopted paradigm both for building resources annotated with PRs and for acquiring prerequisite relations relying on automatic strategies [103, 244, 264, 301]. Such approaches, heavily relying on structured knowledge bases (e.g., Wikipedia, DBpedia or ontologies) and domain experts' background knowledge, might return poor results when applied to domains not well covered by the external knowledge, or they might bind the PRs to a specific ontology of the subject matter showing relations which might not be reported in all instructional materials. The opposite might also happen: we might overlook and fail to find a relation because it is not mirrored in the knowledge structure we rely upon (or in the experts' ideal representation of the domain).

Consider as an example of the latter phenomenon the following text excerpts taken from a computer science textbook[1]. Both discus the family of devices used to connect different networks. In the first text (text *1)* we underlined the first occurrences of the relevant concepts (identified by a domain expert), while in the second text (text *2*) underlined entities correspond to existing hyperlinks to other Wikipedia pages.

> *1)* The simplest of these [devices] is the repeater, which is little more than a device that passes signals back and forth between the two original buses (usually with some form of amplification) without considering the meaning of the signals. A bridge is similar to, but more complex than, a repeater. Like a repeater, it connects two buses, but it does not necessarily pass all messages across the connection. [...] The connection between networks to form an internet is handled by devices known as routers, which are special purpose computers used for forwarding messages. Note that the task of a router is different from that of repeaters, bridges, and switches in that routers provide links between networks while allowing each network to maintain its unique internal characteristics. [42]

> *2)* In telecommunications, a repeater is an electronic device that receives a signal and retransmits it. Repeaters are used to extend transmissions so that the signal can cover longer distances or be received on the other side of an obstruction. Some types of repeaters broadcast an identical signal, but alter its method of transmission, for example, on another frequency or baud rate.

In text *1*, *repeater*, *bridge* and *router* equally denote a sort of device with a similar function. During the flow of the exposition, the author naturally presents these concepts in a sequential order, offering for each of them a definition based on similarities and differences with respect to the others. As a result, understanding one of them, for the reader, becomes very useful to

[1]It is the same textbook we used to build the PR-annotated dataset as presented in Chapter 7 and the Wikipedia page on 'Repeater'[2]

understand the next in line, thus suggesting the presence of a prerequisite relation between these concepts. It should be noted that, in an ontological representation, these three concepts would be arguably encoded as sibling nodes (not-hierarchical organisation), possibly children of *device*. If we consider text *2*, what is most striking is that there is no concept overlap between the two texts: apart from not being present in the presented excerpt, *bridge* and *router* are never mentioned also in the remainder of the Wikipedia page. What clearly emerges is that, although the two texts discuss roughly the same topic, they adopt two different perspectives and explanatory strategies.

If we were to adopt the ontological approach to enrich the two above texts with PRs, we might end up adding some relation that is ontologically legitimate but not really present in the text. Given these issues, if our goal is exploring and modelling how prerequisite relations are expressed and organised in educational materials, strictly bounding the identification of PRs to a specific text is the most appropriate strategy: tracing prerequisite relations along a text, as in the pedagogical view, rather than acquiring them from structured knowledge representations would be more effective in uncovering the PRs actually expressed within it. Essentially, uncovering the PR structure of the text would consist of finding how the author decided to organise the content of the resource. This approach more naturally mirrors the human learning process. In fact, even in the case of a poorly conceived learning material, that may lack important dependencies or where concepts are presented in an awkward order, it is reasonable to identify the relations that are expressed in that specific text: a final user (e.g. a learner studying the material) will eventually cope with these relations and not with those included in a domain ontology or reflected in a particular expert's background knowledge.

### 4.1.1 The Holistic Process of Identifying PRs within Texts

Ideally, the interpretation of an educational text on the part of the learner should perfectly match the communicative intent of the writer. However, moving beyond the surface level of the language and dealing with semantic interpretations requires to activate some inferential mechanisms that allow the comprehension of the content of a piece of text [69]. This might result in multiple interpretation of the text content and, consequently, on diverse prerequisite structures. Investigating what causes slight of significant divergences of interpretation among text readers is beyond our scope. What is interesting from our perspective is finding the consequences of the interplay between text and reader. As a matter of fact, one of the implications of adopting a pedagogical view is that we must put great attention on the text of the resource we are using as some prerequisite relations might be implicitly entailed in the text.

Given the observations above, we refer to the PR identification (and thus manual annotation) activity as an *holistic process*, which involves not only identifying the relation, but also having a clear understating of the context where concepts are described as some relations might emerge from reading even large text fragments rather than individual sentences. Indeed, the meaning of words can very depending on the context where they appear. In the educational context, we argue

that what changes with respect to the context is not concept meaning, but concept relationship with other pieces of domain knowledge: first a concept might be just mentioned or introduced, then used inside its definition and later recalled to explain some new information or to explain another concept. To better understand this idea, consider the examples below, showing short texts extracted from Wikipedia. In the excerpts, we underlined three of the relevant concepts.

> *3)* Malware is any software intentionally designed to cause damage to a computer, server, client, or computer network. A wide variety of malware types exist, including computer viruses, worms, Trojan horses, ransomware, spyware, adware, rogue software, wiper and scareware.[3]

> *4)* In mathematics, an equation is a statement that asserts the equality of two expressions, which are connected by the equals sign. [...] The most common type of equation is a polynomial equation in which the two sides are polynomials. The sides of a polynomial equation contain one or more terms.[4]

Based on the content of these short texts, we might be able to identify the following PR relations occurring between the underlined concepts. In example *3)*, a *malware* is introduced as a type of *software*, thus we could create the relation $software \prec malware$[5]. As a *Trojan horse* is in turn described as a specific type of *malware*, we could also add the relation $malware \prec trojan\ horse$. Consider now example *4)*. Here, *equation* is presented as a generic statement including more specific types of equation, such as the *polynomial equation*, which in turn is composed of two sides comprising *terms*. Such explanation justifies the PRs $equation \prec polynomial\ equation$ and $polynomial\ equation \prec terms$. Although we might end up creating different sequences if considering other texts, these above are motivated by the presentation of concepts in this specific case. The above are direct and explicit PRs as we could identify lexical cues helping us spotting their presence. However, these texts also entail two mediated PRs: $software \prec trojan\ horse$ and $equation \prec terms$. These two relations could be easily motivated as they result from the interpretation of the overall textual context where they appear, although there is no connective clue hinting for the PR between the two: in example *4)*, for instance, $equation$ and $terms$ are even mentioned far away from each other (in terms of inter-occurring sentences). These are clear examples of *transitive PR relations* (see Sec. 2.2.1) which happen quite frequently between educational concepts: two concepts might be related because they are part of the same learning path, but not directly connected or discussed within the same text fragment.

Such type of implicit relations are usually vague and easily influenced by the interpretation of the text [275]. Indeed it seems that familiarity with the domain plays a crucial role on the task of finding concept relationships within educational texts. For instance, [220] suggests that

[3] From Wikipedia page about 'Malware' (`https://en.wikipedia.org/wiki/Malware`).
[4] From Wikipedia page about 'Equation' (`https://en.wikipedia.org/wiki/Equation`).
[5] Following the notation used in Chapter 2, we represented the relation "A is prerequisite of B" as $A \prec B$.

readers generally perceive a text as more complex if they are not familiar with the domain, and a higher perceived complexity usually results in a lower number of identified relations (not all necessarily legitimate) between mentioned concepts. This fact sheds more light on the fact that, in the pedagogical view, we should pay attention not only to texts content, but also to linguistic complexity. Ultimately, it might be that choosing a more or less complex instructional text might affect the overall correctness and homogeneity of manually annotated relations, resulting in better or worse annotations overall.

As our goal is defining a systematic methodology for dealing with PRs, also when manually annotating them in texts, we can't carry on our research without before addressing the following question, left – in our opinion – unanswered by the existing literature on PRs: *are all instructional materials equally suitable resources to uncover prerequisite relations between educational concepts?* In order to investigate the impact of linguistic complexity on PR manual recognition, we carried out a crowd-based experiment on concept sequencing, described in the Section below.

## 4.2 Impact of Linguistic Complexity of Texts on Prerequisites Identification

The goal of this Section is investigate whether there is an impact of textual complexity on the task of manual identification of prerequisite relations between concepts. To pursue such goal, we setup a crowdsourcing experiment on *prerequisite concept ordering*.

In spite of some well known limits of crowdsourcing [254], such approach has become a widely used paradigm in NLP to collect human judgments about linguistic phenomena. Its mostly appreciated advantages are being fast and allowing to collect a wide variety of judgments. Among the vars amount of works employing crowdsourcing to collect human judgments, it is worth mentioning [44, 76, 153], which specifically tackled textual complexity. Among the works dealing with PRs, crowdsourcing was employed for annotating PR relations between Wikipedia pages by [264]. Here, the authors inferred a set of PRs between concepts exploiting Wikipedia hyper-links and then used crowd-sourcing to validate those candidate relations.

In principle, the task of prerequisite concept ordering proposed in this experiment consists of creating sequences of concepts motivated by prerequisite relations based on the information contained in randomly presented short concept descriptions[6]. Contrary to [264], our goal here is not to create a manually PR-annotated dataset, but to investigate whether different concepts descriptions convey different concept sequences. For this reason, we ask subjects to create prerequisite sequences rather than validating pre-arranged concept pairs. Our investigation is motivated by the fact that it is widely acknowledged that linguistic complexity has a direct relationship to readers' comprehension (see e.g. [21, 139, 211, 262]). Following from such evidence, our intuition is that complexity might have an effect also on the task of structuring the knowledge

[6]In what follows we will use the terms 'orderings', 'order' and 'sequencing', 'sequence' interchangeably.

contained in a resource by means of PRs. To test the impact of textual complexity on prerequisite identification, we built three parallel versions of the ordering task varying with respect to the texts used to describe concepts. Specifically, texts were acquired from three different sources, each targeting a different audience and representing increasing degrees of complexity.
Details about the experiment setup and the obtained results are presented below.

### 4.2.1 Experimental Setup of a Crowd-based Concept Ordering Task

#### 4.2.1.1 Prerequisite Concept Ordering Task

Prerequisite concept ordering, in general terms, consists of manually creating a sequence of concepts reflecting their prerequisite structure. In the current experiment, we asked to perform the task by sequencing randomly ordered concepts triples. Each triple consists of a short learning path, thus, in order to be valid from an instructional point of view, the proper learning precedence of concepts within the triple must be preserved in the re-created sequence.

For the purposes of the current experiment, each concept is represented by a short text containing its description. Hence, prerequisite concept ordering here consists of sequencing the short texts knowing that each text refers to a different concept. More formally, given three randomly ordered concepts $A$, $B$, and $C$, each represented by a short text, referred to as $t_A$, $t_B$ and $t_C$, we ask to create the ordered triple $T = (t_A \prec t_B \prec t_C)$, which conveys the following meaning: $A$ is prerequisite of $B$ and $B$ is prerequisite of $C$. To guide subjects during the sequencing process, we asked to order the texts so that the final sequence reflects the following properties:

1. $t_A$ contains the knowledge required to understand $t_B$ and $t_C$;
2. $t_B$ can be understood only if $t_A$ is known;
3. $t_B$ is required to understand $t_C$;
4. $t_C$ can be understood only if $t_A$ and $t_B$ are known.

In order to guarantee that concept triples show the prerequisite relations between the concepts involved, we retrieved gold concept pairs from AL-CPL dataset [168]. As detailed in Section 3.4, the dataset contains manually annotated prerequisite pairs of concepts belonging to four domains, namely data mining, geometry, physics and precalculus. We claim that concept triples represent short learning paths as we selected them if the dataset reported $A \prec B$, $B \prec C$ and $A \prec C$ as positive pairs (i.e., showing a prerequisite relation). Given the high correlation (83%) reported by the authors between the three experts while manually annotating AL-CPL pairs, we assume that the prerequisite relations contained in the dataset aren't ambiguous. Anyways, we manually inspected the texts in order to verify whether the sequences proposed by Al-CPL were reflected in the concept descriptions.

The four domains of the dataset have different sizes (considering the number of prerequisite-annotated pairs), thus we selected 3 concept triples from the larger domains, i.e. geometry,

precalculus and physics, and 1 triple from the smallest, i.e. data mining. Eventually we obtained 30 concepts spread over 10 triples.

#### 4.2.1.2 Concepts Descriptions

In order to investigate whether different concept descriptions affect subjects' identification of prerequisite relations, we collected the short texts describing concepts of the experiment from three different sources varying with respect to linguistic complexity. Linguistic complexity is traditionally defined with respect to the intended audience, i.e. as a measure to determine how challenging a text is for a reader on the basis of multiple linguistic factors, thus affecting the reader's ability to access text content [72]. Accordingly, here we define linguistic complexity with respect to the intended audience targeted by the source that we used to acquire the concept description. Specifically, concept descriptions were acquired from the following sources: *i)* Simple English Wikipedia[7], *ii)* English Wikipedia[8] and *iii)* encyclopedias (i.e., Encyclopedia of Mathematics[9] for precalculus, data mining and geometry concepts, Encyclopedia of Physics[10] for physics concepts). All sources are works of encyclopedic scope since they are organised in entries (articles) and they provide factual information about the concept covered by the article. However, they differ with respect to the target audience, as discussed below, thus they can be used to represent different levels of reading difficulty, ranging along the complexity spectrum from simple to complex.

More in detail, the three sources can be described as follows:

i *Simple English Wikipedia* is an online free encyclopedia written at a basic level of English to foster learning for children, adults with learning difficulties, students and English Language Learners. Its goal is to help readers understand hard ideas or topics through easy-to-read content [132, 270], thus we use this source to represent the ***simple*** variety of texts.

ii *English Wikipedia* is a highly popular online encyclopedia with more than 3 hundred million visits each day and more than 39 million pages[11]. Wikipedia is an open project where volunteer contributors are not required any specific qualification and whose goal is to deliver knowledge to everyone freely. Although the quality of Wikipedia articles in terms of content is known to be low for some pages, [105] showed that the level of accuracy of hard science articles approached that of encyclopedias curated by experts. For this reason, we relied on this source to represent the intermediate complex variety, which we refer to as ***neutral***.

iii *Specialised Encyclopedias*, unlike the previous two, focus on a single domain and target an audience of experts of the field. Those resources often assume that the reader already

[7] `https://simple.wikipedia.org/wiki/Main_Page`
[8] `https://en.wikipedia.org/wiki/Main_Page`
[9] `https://encyclopediaofmath.org/wiki/Main_Page`
[10] Besancon, R. (2013). *The Encyclopedia of Physics*. Springer Science and Business Media.
[11] `https://stats.wikimedia.org`

masters the fundamentals of the discipline and aim to convey in-depth discussions about the relevant knowledge of the subject domain. For precalculus, geometry and data mining concepts we selected their descriptions from the "Encyclopedia of Mathematics Wiki", an open access graduate-level resource designed by experts for the mathematics community. Physics concept descriptions were selected from "Encyclopedia of Physics", an encyclopedia written by international authorities in the field of physics. Considering that this resource is intended for specialised contexts, we use it to represent the ***complex*** variety of texts.

In summary, for each concept $A$ of our experiment, we have three short descriptions (3 to 5 sentences long, around 100 tokens on average) $tS_A$, $tW_A$ and $tE_A$, referring to the descriptions acquired, respectively, from Simple Wikipedia, Wikipedia and encyclopedias.

#### 4.2.1.3 Crowdsourcing Design

Considering our goal, i.e., investigate whether different concept descriptions affect subjects' identification of prerequisite relations, we defined three crowdsourcing tasks, one for each linguistic complexity level, i.e. simple, neutral and complex. The crowdsourcing tasks were all administered through the Prolific platform[12], which allowed us to select how many subjects involve in each task and their desired characteristics. We decided to recruit 20 different subjects for each task among native English speakers.

Each crowdsourcing task consists of a questionnaire $Q_x$ where subjects are asked to solve the prerequisite concepts ordering task on 10 concept triples $T_x$, with $x = \langle S, W, E \rangle$ depending the complexity level of the texts used to describe concepts. Note that we used the same 10 concept triples in all questionnaires, presented in the same order both with respect to the order of questions in each questionnaire and of the internal random order of concepts in each triple. Concept triples and their correct prerequisite ordering as reported in the AL-CPL dataset are displayed in Table 4.1. Questionnaires only vary with respect to the complexity level of the texts used to describe concepts. In practice, the triple (Arithmetic $\prec$ Multiplication $\prec$ Power) is represented in $Q_S$ through the descriptions of concepts *arithmetic*, *multiplication* and *power* acquired from Simple Wikipedia, while in $Q_W$ and $Q_E$ through the descriptions acquired from Wikipedia and the encyclopedia respectively. This way, questionnaires are homogeneous with respect to linguistic complexity, meaning that each triple $T_x$ of $Q_x$ is represented by texts acquired from the same source.
Such setting allowed us to *investigate the influence of linguistic complexity on the task of prerequisite concept ordering without suffering the influence of other intervening factors*. We further included in each questionnaire 2 control questions to identify non-reliable subjects.

Before ordering the concept triples, subjects were provided with task instructions and an example of a solved question (Figure 4.1), supported by a brief motivation for the proposed

[12] `https://www.prolific.co/`

| ID | Concept A | Concept B | Concept C | Gold Sequence |
|---|---|---|---|---|
| 1 | Geometry | Cone | Circle | A-C-B |
| 2 | Line | Angle | Point | C-A-B |
| 3 | Addition | Summation | Arithmetic | C-A-B |
| 4 | Gravity | Gravitational field | Physics | C-A-B |
| 5 | Skew Lines | Line | Parallel | B-C-A |
| 6 | Acceleration | Speed | Motion | C-B-A |
| 7 | Sample | Statistical significance | Confidence interval | A-C-B |
| 8 | Polynomial | Number | Integer | B-C-A |
| 9 | Function | Mathematics | Limit of a function | B-A-C |
| 10 | Deformation | Hooke's Law | Elasticity | A-C-B |

Table 4.1: Concept triples administered to subjects recruited for the experiment. Each concept A, B and C is represented in the questionnaire by means of a short textual description and the term referring to the concept is masked as described in the experimental design Section. The 'Gold Sequence' column reports the gold prerequisite ordering of the three concepts as acquired from the Al-CPL dataset.

ordering. As can be noted, the name of concepts in the texts is covered by masks represented by alphanumeric codes. Masks were introduced to avoid the influence of subjects' background knowledge in solving the task. In fact, a preliminary experiment aimed at defining the experimental setting revealed that, once identified the concept described in each of the three text, subjects tended to ignore the descriptions and to create sequences based on their prior knowledge abut the domain. Given our goal of observing the impact of texts on the prerequisite ordering task, we decided to *conceal concept names in the descriptions and asked subjects to avoid trying to guess which concept is hidden behind the mask* since, by doing so, they could create an ordering based on their prior knowledge about the topic rather than on the texts' content.

#### 4.2.1.4 Task Example

In order to better understand the experiment setup and desired output, we will now look again at the example in Figure 4.1.

Text A, B and C describe a different concept each by means of a different short text. The multiple choice grid can be used to report the sequence of texts believed to reflect the most appropriate prerequisite sequence. The ordering proposed in the example as correct solution is the following: $(C \prec A \prec B)$. This solution is motivated by the fact that text C describes a concept which is a prerequisite of the concept covered by the mask *j7o* described in text A: according to the text, *q48* consists of studying *j7o*. In order to understand *3s0* (explained in text B) the student should know *j7o* ("*3s0* corresponds to *j7o*") and consequently text B should be placed as third in the ordered sequence. To satisfy the curiosity of the reader, we can now reveal that j7o is a mask for *multiplication*, q48 for *arithmetic*, and 3s0 for *power*. It could be argued that *arithmetic* isn't a prerequisite concept of the other two, but rather a cover term used for referring to the subject

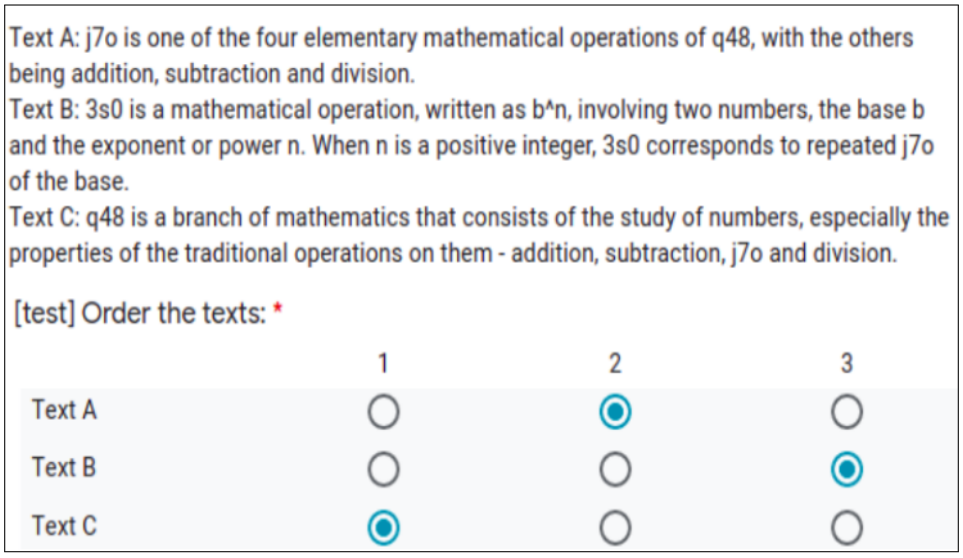



Figure 4.1: Test question presented to subjects.

domain, thus representing a taxonomic relation with the others more than a PR. Although in the current experiment we simply used the concepts proposed by the dataset we relied upon, in our PR annotation protocol we accounted for such observation, as we will discuss in 5.2.2.2.

### 4.2.2 Linguistic Complexity Evaluation

As preliminary analysis, we measured the complexity of the texts extracted from the three sources, namely Simple Wikipedia, Wikipedia and encyclopedias. To pursue this goal, we relied on Profiling–UD[13] [43], a web–based application that performs linguistic profiling of texts for multiple languages. By relying on the linguistic annotation of texts at mopho-syntactic level, Profiling–UD extracts more than 130 features modeling lexical, grammatical and syntactic phenomena that, all together, contribute to characterize language variation within and across texts [43, 114]. The set of linguistic features monitored by Profiling–UD are extracted from the different levels of annotation (the texts are annotated according to the UD formalism, see 3.1.2) and capture a wide number of linguistic phenomena, such as *a)* raw text properties, *b)* lexical variety, *c)* morpho-syntactic information, *d)* verbal predicate structure, *e)* global and local parse tree structures, *f)* syntactic relations and *g)* use of subordination. The complete list of features monitored by Profiling-UD and their description is reported in Table B.4 of Appendix B.

We exploited Profiling–UD to analyse the texts used in the questionnaires, representing different text varieties identified on the basis of the target audience, and compared the results. In order to explore whether there is an association between the values of the features extracted using Profiling-UD and the text variety represented by our texts, we performed Mann-Whitney

[13]Demo available at `http://linguistic-profiling.italianlp.it/`

U Test correlation analysis on the features values of each group of texts (texts are grouped on the basis of their original source). See the results in B.1 (Appendix B). Mann-Whitney U Test analysis revealed significant differences between the texts, concerning in particular the *parse tree* and the *verbal predicate structure*. The parse tree structure varies in particular for what concerns the use of prepositions, either verbal arguments or nominal and adjectival modifiers sharing the same nominal head: confirming our original distinction, Simple Wikipedia sentences are characterised by few embedded complement 'chains', which make the sentences simpler overall. On the other hand, sentences extracted from encyclopedias, representing the complex variety, show a richer subordinate structure and higher 'verbal arity', i.e. a feature capturing the average number of dependency links (covering both arguments and modifiers and excluding punctuation and auxiliaries) sharing the same verbal head. Refer to B.1 in Appendix B for consulting the average values obtained by each group on significantly varying features.

Since these properties are among the most predicting features for sentence complexity [82], the fact that they vary significantly between groups shows that the sentences belonging to the three text varieties represent different degrees of complexity. As proof, consider the visualisation displayed in Figure 4.2 obtained using Principal Component Analysis (PCA). PCA is a classical data analysis method that reduces the dimensionality of the data while retaining most of the variation in the data set by identifying principal components, along which the variation of the data is maximal [135]. We computed PCA on all sentences in the experiment (represented as the vector of significantly different feature values) by considering 2 principal components and plotted the results to visually assess similarities and differences between them. PCA visualisation (Fig. 4.2) reveals that, regardless of the major aggregation at the center of the plot, the sentences tend to group in a way that reflects their expected complexity degree: Simple Wikipedia and Encyclopedia sentences tend to stay in the lower and higher part of the plot respectively, whereas Wikipedia sentences have a sparser distribution. The latter result suggests that, rather than neutral, the complexity of Wikipedia texts is mixed and highly variable from sentence to sentence, as observed also by [132].

Based on such results and thanks to the ordering answers collected in the crowd-based experiment described above, we aim to test the following two hypotheses.

- **HP1)** Textual complexity affects the human identification of prerequisite relations between concepts. In particular, since a student can only acquire information about unknown concepts from their description in educational materials, we expect easier-to-read texts to express more clearly and unambiguously the propaedeutic relations between concepts. Conversely, prerequisite ordering judgments expressed on the basis of the information presented in more difficult-to-read concepts might be less homogeneous, revealing a higher difficulty in abstracting the relations from the texts.

- **HP2)** The pedagogical role of concepts influence the complexity of texts describing them. In other words, the most fundamental concepts in learning paths, corresponding to the

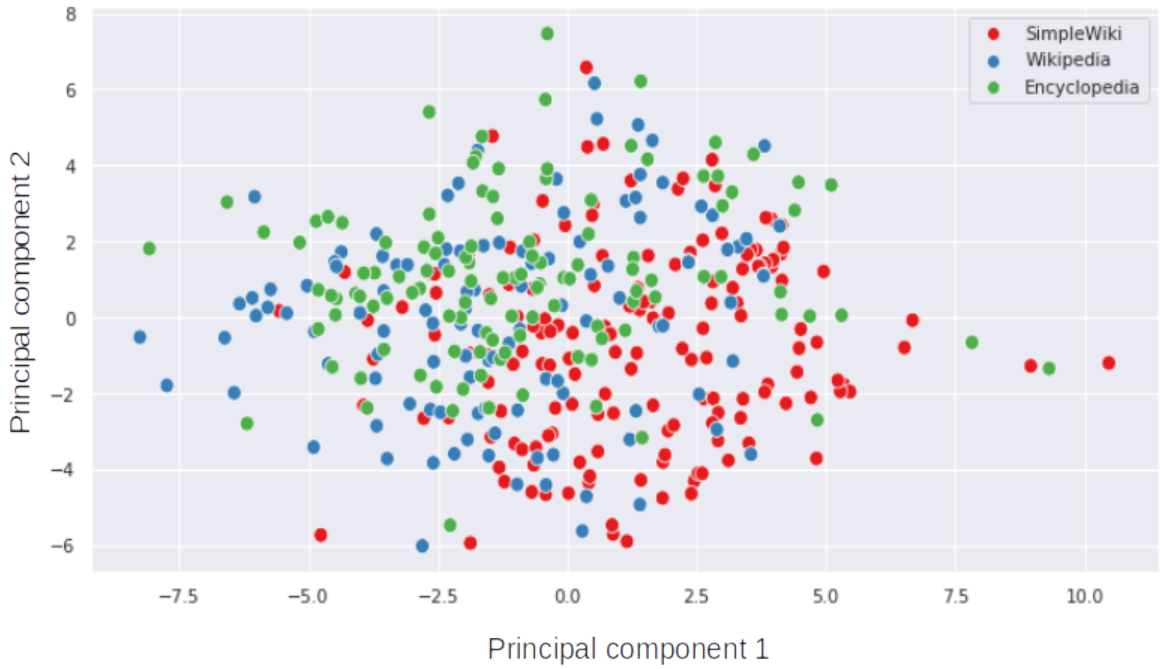



Figure 4.2: PCA visualisation of the sentences (each dot corresponds to a sentence) contained in the texts used for the experiment. Different colours are used to indicate to which group of texts the sentences belongs.

first element of the prerequisite sequence, should be described in a easier and simpler way than the subsequent elements as they convey more basic knowledge about the domain. Vice-versa, concepts representing more advanced and detailed knowledge (i.e., the final items in the sequence) might be described through more complex texts.

In order to investigate HP1 and HP2 above, we compared the subjects' answers collected on the three questionnaires. Our intuition is that, if HP1 is true, the orderings of the Simple Wikipedia texts will be more similar to the gold sequence than the orderings of more difficult-to-read texts. HP2, on the other hand, could imply that the first elements of the concept sequence are easier to identify, regardless of the textual variety of the texts. The results of the analyses addressing HP1 and HP2 are presented below in Sections 4.2.3 and 4.2.4 respectively. Note that we restricted our analyses to the answers provided by 15 subjects. Specifically, we excluded from the study those subjects who failed the control questions and who took less than 5 minutes to answer the questionnaires (which we empirically evaluated as minimum time).

### 4.2.3 Linguistic Complexity and Concept Orderings

The impact of linguistic complexity on the task of prerequisite concept ordering has been assessed here with respect to the orderings produced by subjects in the three questionnaires. First,

we computed the accuracy of the answers collected for each question of each questionnaire. Answering a question correctly consists in this case of re-creating the same triple sequence proposed in the AL-CPL dataset. Consequently, ordering accuracy has be computed as a ratio between correct answers and total answers collected (corresponding to the number of subjects taking the questionnaire, i.e. 15, as questions can't be left blank). The accuracies values of each questionnaire are reported in Figure 4.3.

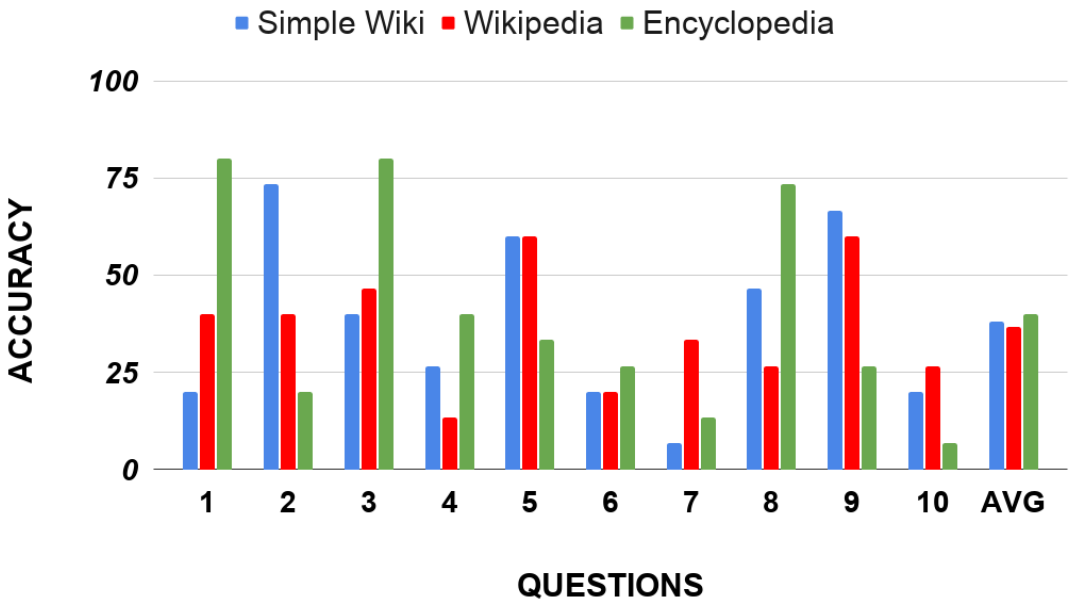



Figure 4.3: Ordering accuracy for each question and overall ('AVG' column) for the three questionnaires.

The bar graph of the figure is interesting in several ways. First, it's worth noticing that no question received a 100% accuracy: the highest reported accuracy (80%) is observed in only two cases, namely question 1 and 3 in the Encyclopedia questionnaire, which received a correct answer by 12 out of 15 subjects. Such result is in line with our previous observations about the difficulty of organising concepts on the basis of their prerequisite relationship (see 4.1). Second, if we consider the average questionnaires accuracy reported in the column 'AVG', we observe a minimal difference between the three textual variety: the questionnaire accuracy corresponds to 38%, 36.7% and 40% for Simple Wikipedia, Wikipedia and Encyclopedia questionnaires respectively. If on the one hand the lower accuracy observed on Wikipedia texts could be explained considering that Wikipedia pages are collaboratively edited by volunteers, possibly with little or no training in writing specialised texts [73, 172, 249], the difference between Simple Wikipedia and Encyclopedias is more striking. Based on HP1, we expected that Simple Wikipedia texts, addressing young learners and readers with cognitive impairments, would have conveyed the relationship between concepts in a clear way, causing only few cases of ambiguity. The picture

emerging from the accuracies comparison depicts a different scenario: in general, subjects were slightly better at identifying the correct sequences of concepts when relying on the descriptions of the more complex texts. This might be due to the fact that, addressing an audience of domain experts, encyclopedic concept descriptions are more precise and accurate, thus they reveal concept relationships more clearly.

The higher complexity of encyclopedia texts is anyways reflected in the average time required to complete the questionnaire (displayed in Figure 4.4). As a matter of fact, if we take time into account, we notice that subjects completed the Simple Wikipedia questionnaire more quickly (13 minutes and 26 seconds on average) than the Encyclopedia questionnaire (20 minutes and 50 seconds on average). The Wikipedia questionnaire, which shows the higher variations with respect to subject times, required 16 minutes and 36 seconds on average, showing a value in between the simple and complex variety. These results suggest that, despite a slightly lower accuracy of the answers, ordering concepts after reading simpler texts is easier than doing the same task based on the knowledge acquired from complex texts, which in turn takes more time. This intuition is supported by the answers provided on a post-questionnaire interview.
During the interview, we asked to rate the task difficulty on a 5-point Likert scale where 1 means "very difficult" and 5 means "very easy". The three groups of subjects all reported the task as difficult, regardless of the questionnaire they were given. However, we notice a variation in the difficulty score assigned by the three groups: the group taking the Simple Wikipedia version of the questionnaire reported an average 2.4 on the Likert scale, whereas the other groups reported an average 1.93 and 2.0 for Wikipedia and Encyclopedia respectively. Note, however, that Prolific platform only reports the amount of time taken to finish the entire questionnaire, thus we are not able to distinguish between the amount of time required to read the texts and to provide the answers. Future analyses, capturing question times, could further investigate this aspect.

In order to further investigate the correlation between subjects' answers and the texts used to describe concepts, we computed Pearson correlation coefficient (PCC) on the questionnaires answers. The highest correlation (PCC=0.54, p-value $< 0.05$) are observed when we consider the answers of the questionnaire involving Wikipedia texts and compare it with the other two questionnaires. Simple Wikipedia and encyclopedia-based answers actually show a slightly weaker correlation (PCC=0.45, again p-value $< 0.05$). Considering that the correlations are all significant, we can highlights two factors: *1)* these results provide a further evidence about the difficulty of the task, which emerges regardless of the texts used for describing concepts; *2)* the impact of the linguistic complexity on the task is significant. In fact, although the average accuracies of answers produced after reading Simple Wikipedia and Encyclopedias are quite similar, they do not converge on the same concept triples. If we consider again the picture depicted by Figure 4.3 in light of the PCC values, we can notice that those questions where the accuracy is high in the Simple Wikipedia variety obtain instead a low accuracy in the Encyclopedia variety, and vice-versa. This result is interesting and should be further investigated in the future by,

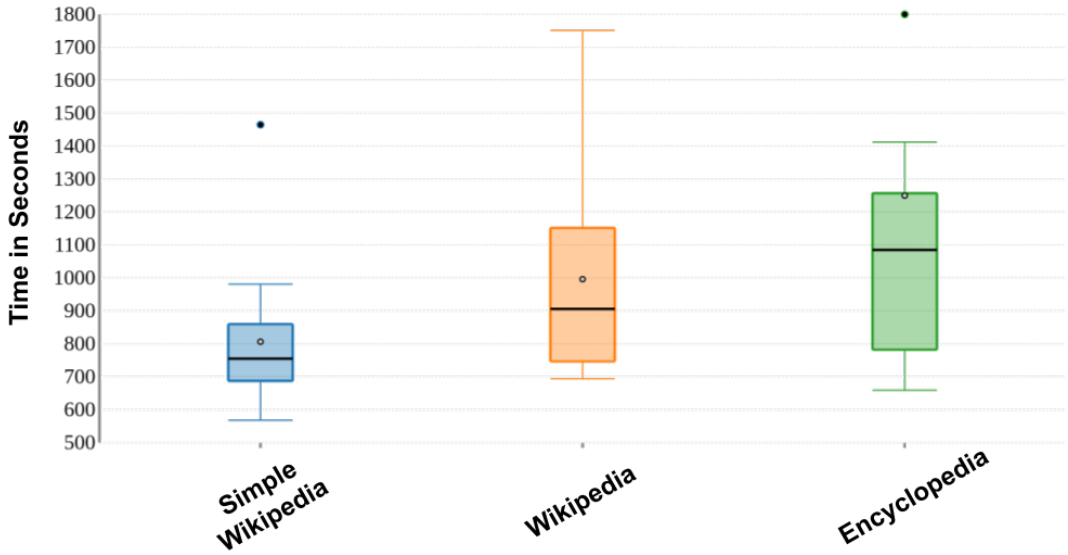



Figure 4.4: Box plot reporting time (in seconds) employed by subjects to complete each of the three questionnaires.

e.g., employing parallel descriptions: presenting the same text adapted to comply with different complexity degrees could allow to explore which constructions make PR recognition more or less difficult. Data collected in the current experiment do not enable us to investigate this aspect in more depth.

### 4.2.4 Influence of Concepts Pedagogical Role

In this Section we address HP2, namely *does the pedagogical role of concepts influence the task difficulty?* To this aim of this analysis, we informally define the 'pedagogical role' as corresponding to the type of information they convey with respect to the overall prerequisite structure of the subject matter. In practice, we say that the first items in the learning sequence play the role of primary notions and they convey basic and fundamental knowledge, whereas the last concepts in the sequence play the role of learning outcomes as the knowledge they convey corresponds to the ideal outcome of the learning path[14]. Our intuition is that fundamental concepts are easier to identify as basic notions, whereas more detailed and advanced concepts are more difficult to understand. To tackle this issue we analyse the crowd-based answers on prerequisite concept ordering to investigate whether subjects are able to find initial concepts of the sequences more easily then final concepts. Initial concepts are those occupying the first position in the triple sequence, while final concepts are those that should be placed in the third position. Our intuition is that it is easier for a reader to access the content of a initial concept description since they are

[14]We take the distance here from the research addressing Learning Objects and educational ontology construction, which used the terms 'Primary Notion' and 'Learning Outcome' with more precise and formal interpretation.

usually more broad and thus should contain less complex syntactic structures. This idea was already proposed by [182] and [18] who introduced readability and complexity-based features for automatic prerequisite learning. However, neither of them tested the impact and significance of those features.

As first step, we compared the accuracy (computed as above) of initial and final elements of the concept sequences: if our hypothesis is true, the ordered concepts should show a higher accuracy on the first elements of the sequence rather than the last ones. The results confirm our hypothesis: regardless of the complexity level, initial concepts show a higher accuracy than final ones. Specifically, in Simple Wikipedia questionnaire, initial concepts are correctly identified in 68% of cases, while final concepts in only 45.34%. Similarly, in the Encyclopedia questionnaire we notice a noticeable difference between the accuracies of initial and final concepts, with the former showing 71.34% of accuracy and the latter only 54.67%. In Wikipedia questionnaire the gap is smaller, although still present, with initial concepts showing 58.67% of accuracy and final concepts 51.34%.

To better investigate the reasons behind such differences, we exploited Profiling-UD to analyse the linguistic properties of the texts describing initial and final concepts. This time, we performed the Profiling-UD analysis at document level (i.e. one document for each concept description) rather than on individual sentences as above. This was done in order to better reflect the human process of understanding a concept: in the experiment, subjects are supposed to read the whole description before creating the sequences. Our goal here is to verify whether texts describing initial concepts are generally easier-to-read (regardless of the textual source) than final concepts descriptions. If confirmed, this might be the reason that made initial concepts more accurately identified.

The complete results of this analysis, as well as the average values obtained by each group on significantly varying features are reported in B.2 in Appendix B. The linguistic profiling analysis on the documents showed again significant differences concerning the verbal predicate structure and inflectional morphology. If we use the two dimension PCA to visually inspect the differences between documents describing initial and final concepts (see Figure 4.5), we see that the documents tend to separate, having final-concepts descriptions mostly on the left part of the plot and initial-concepts descriptions on the right and central side. The dispersion plot in Fig. 4.5 suggests that the descriptions concerning final concepts tend to be more similar and, considering their linguistic features, equally complex. On the other hand, initial concepts descriptions complexity seems more variable. This rather mixed picture emerging from the central part of the graph is not completely surprising if we consider that, in this analysis, we didn't distinguish the texts depending on the original source. What is interesting however is that, eventually, although having only few items (60 documents) and a quite sparse visualisation, the results hint for the presence of complexity differences associated to the pedagogical role of the concepts. We further investigate these difference on the PR-annotated gold dataset described in

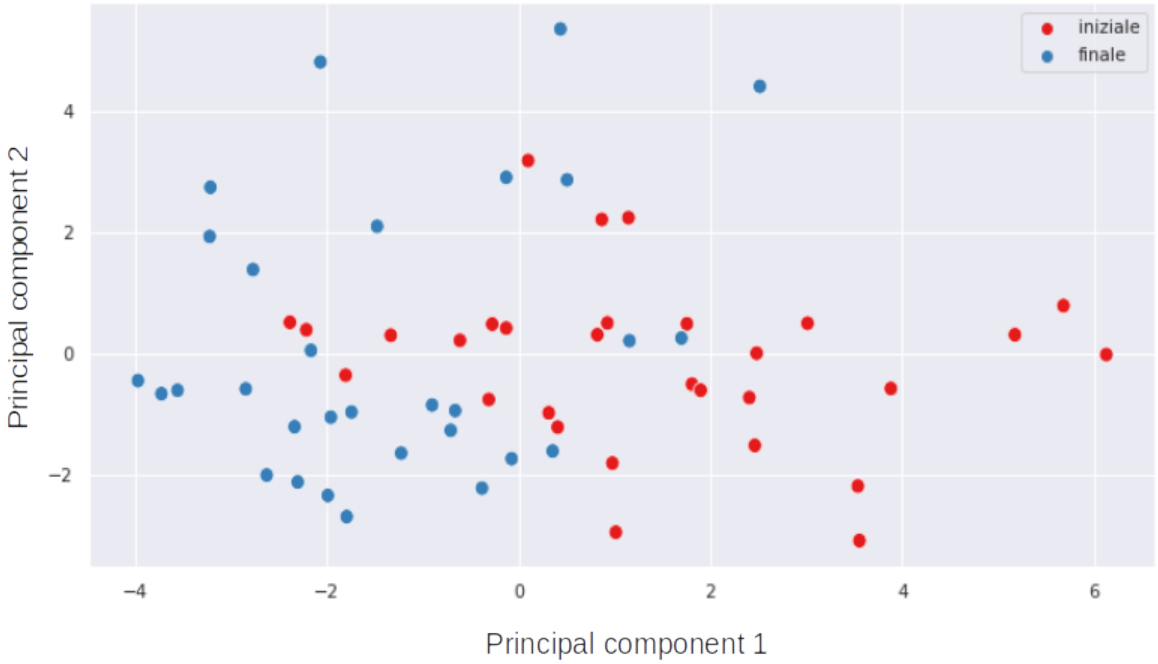



Figure 4.5: PCA visualisation on initial and final concepts of the PR sequences. Each dot represents a document, i.e. a concept description.

Chapter 7 (Section 7.5.2).

## 4.3 Towards a Novel PR Identification Methodology

The results presented above offer some insightful evidence about the tight bond between the task of PR identification and the content and expository style of educational materials. Specifically, we observed that, although prerequisite identification is a rather difficult task overall, the complexity of the text where relations are searched for can make the results more or less reliable. This fact shows from two intertwined results: *a)* easy-to-read texts allowed faster and more accurate orderings; *b)* texts addressing domain experts are possibly more precise and this makes PRs emerge more clearly and unambiguously, although it takes longer for readers to deliver a concept sequence. Such results provide further motivation for considering PR identification as an holistic process, involving multiple simultaneous tasks: textual content comprehension, relation extraction and logic inference. As already mentioned in Sec. 3.4, most PR-annotated datasets tend to neglect these facts, which in the reality are highly impactful and should be kept into account. Our effort toward the definition of a methodology for dealing with PRs which incorporates the findings of our text variety analysis brought to the conclusion that *uncovering PRs from educational texts* is the most viable option.

As we anticipated in chapter 1, our methodology has been systematised in the PR framework

which incorporates a protocol for manual annotation of PRs in texts and a model for supervised automatic extraction of PR relations which exploits the manually annotated examples previously obtained. In what follows we will summarise the motivations and challenges of our methodology for uncovering PRs, focusing in particular on the issues related to manual annotation. Indeed, it is widely acknowledged by the community working on resource creation and corpora annotation that properly defining problems is vital when designing annotation tasks: the clearer the definition of the problem, the better the data that will be collected, allowing for annotations less influenced by the subjectivity of the single annotator or by her/his interpretation of ambiguous instructions [127]. The effects of a well-defined and motivated annotation protocol will have positive impacts also on automatic extraction.

### 4.3.1 Text-Bound Annotation Approach

When we designed our novel PR annotation task, we decided to rely on the pedagogical perspective. Uncovering PRs based on instructional materials content is a key point of our methodology as it allows to capture the instructional design knowledge, namely the author's organisation and arrangement of concepts, while simultaneously bounding the annotated PRs to the text. Such result is actually novel and it can be achieved by designing an annotation task which dictates to *annotate PRs directly on texts while reading it, thus reflecting the reader's understanding of the prerequisite structure entailed in the text*.

By adopting an annotation approach strictly bound to the text being read we lay the foundations for performing analyses not possible otherwise aimed at identify which textual contexts triggered PR relations. In practice, our annotation approach requires to add labels indicating the presence of a PR directly on the text. Consider, as an example, the short text (taken from the example in Fig. 4.1): "Arithmetic is a branch of mathematics that consists of the study of numbers, especially [...] addition, subtraction, multiplication and division". Based on our approach, one would need to express in the annotation the this particular instance of the concept *'arithmetic'* is described in such a way that a learner first needs to know what *'mathematics'* is (assuming that both are concepts). The fact that such relation is expressed by this particular occurrence of *'arithmetic'* is relevant to us as a different occurrence in another part of this same text might underlie a different relation. More detailed information about the formalisms we adopted to encode such highly specific information and the description of the output of the annotation process will be provided in the next chapter. For now we simply want to underline the fact that, although our approach might seem trivial as text-bounded annotations are common practice in corpus annotation tasks, anchoring PR to text spans is actually novel in PR annotation.

As a consequence of this choice, the manual annotation will encode the following information:
*(a)* a PR is found in the text;
*(b)* the PR occurs between concepts $A$ and $B$;
*(c)* the PR was found while reading a certain span of text which ideally contains the information

suggesting the presence of the relation.

Point (c) is particularly promising and novel: once we know where in the text we should search for a PR, we can explore such text spans searching for linguistic cues hinting the presence of the PRs. Such investigations are not possible in ontologically-based annotations as there is no reference text and the motivation of a PR could be retrieved only by interviewing the annotators. In order to achieve our goals, we designed *PR annotation as a corpus annotation task where PRs must be anchored to the text portions where they are identified by the human annotator*. Such text-bound annotation can be used to motivate PRs by exploring the context surrounding the annotated text fragment. Additionally, it could be used to investigate whether PRs appear in recurrent linguistic structures. In order to achieve this goal, the corpus needs to be linguistically annotated in advance, as we will discuss in 5.1.2. It should be mentioned that such approach could be adapted also to generalise captured knowledge: when comparing multiple annotations produced on different texts tackling the same topics, it could be used to compare different teaching approaches for the same subject matter.

#### 4.3.1.1 Educational Texts

Text annotation of PR relations, intended as the task of manually enhancing educational resources with explicit propaedeutic relations between concepts, has been performed on multiple types of educational materials (see 3.4). Among them, *textbooks* might be one of the most suited resource to build a corpus for annotating PRs which has a good coverage of the topics of a subject domain and, at the same time, it shows a clear explanatory approach. Despite being rarely used for PR annotation (if not for acquiring concepts as in [149, 176, 283]), textbooks constitute the most prominent learning resource in traditional classroom-based settings [154]. Indeed, they contain narrative and expository text, from which students are expected to acquire novel knowledge after careful reading [184]. Being self-contained and curated instructional materials, textbooks generally cope with every concept a learner has to know in order to understand the book content. This fact involves multiple advantages when using these resources in PR annotation. First of all, we can assume that the concepts appearing within a book are all related with each others. Based on this assumption, we can expect that PRs identified on the basis of the textbook content as defined above would allow to obtain rich annotations in terms of number of identified relations. Second, textbook authors generally curate their materials so that they already show a quite apparent arrangement of concepts, possibly also highlighting most relevant information in order to guide learners through the reading. As a consequence, in a text-bound annotation, annotators should simply identify the pre-organised learning path, rather than re-arranging concepts according to their sensitivity and prior knowledge. The resulting annotation would reflect the explicit content structure as well as the implicit inner structure of the book (such as transitive relations). Note however that the prerequisite structure can be acquired from any source of instructional material: as long as it relies on textual content, our approach can be

applied. Textbooks simply represent a perfect ground for challenging our methodology as they contain long and, possibly, coherent concept descriptions which should make the annotation process more onerous to carry on. It the methodology works well on textbooks, it can be easily applied on many other resources.

### 4.3.2 Challenges with Annotation Evaluation and Automatic PR Learning

The principles guiding the design of our PR annotation protocol allow to create datasets that simultaneously encode multiple information, referring both to the text content and possibly also to its linguistic structure. However, such approach also involves some challenges and peculiarities that need to be acknowledged and taken into account.

#### 4.3.2.1 Comparing the Manual Annotations

Previous work mostly addressed PR annotation and evaluation as a pairwise combination of concepts [264, 283, 301]; we also used this approach in [14]. However, we believe that such approach over-simplifies the result of the annotation process and may result in misinterpretations of the relations contained in the text. Indeed, the final output of the annotation could be represented as a directed graph in which each path is an interpretation of a relation that arises from reading the whole text and should be evaluated accounting for those peculiarities. The commonly adopted pairwise evaluation of PRs misses the interdependence between concepts involved in a PR path and does not take into account the PR-annotated graph as a whole [170]. Temporal relations may represent an interesting ground of comparison for such cases, since precedence relation also shows a transitive sequential nature and are reported as difficult-to-annotate relations [248]. Similarly to what we see happening in PR evaluation, also temporal relations evaluation usually misses to consider the overall result of the annotation (or automatic extraction) in favour of but sub-results, such as individual pairs of successively described events [179, 274] or even same-sentence events [151]. Indeed, researchers in both fields encounter similar limitations using traditional performance metrics used in information retrieval, e.g. precision, recall [269]. A common scenario in both fields is when three items $A, B, C$ (concepts or events) are annotated by a rater such that $A < B$ and $B < C$, but another rater identifies the relation $A < C$: in such cases, traditional agreement metrics, such as Cohen's $k$ introduced in 3.2.2, will fail to identify $A < C$ as a shared relation, even if it is an implicit consequence of the other two [269][15]. This suggests that a better strategy to compute agreement might consist of considering transitive edges and path similarity in the two graphs, or at least employing metrics that do not penalise the absence, in the annotation, of relations that may be legitimately derived from the annotation graph. Our proposal and novel adaptation of traditional agreement metrics designed with the aim of overcoming the above limitations will be presented in Section 5.2.3.

[15] $<$ indicates both the temporal relation *before* and the prerequisite relation $\prec$

#### 4.3.2.2 Automatic PR Learning

Although several methods have been devised to extract prerequisite relations, they were mainly focused on educational materials already enriched with some sort of explicit relations, such as Wikipedia pages, MOOC materials or Learning Objects (see 3.5). Conversely, *a more challenging task is the identification of prerequisites when no such external relations are given*, and the textual content is therefore the only available source of information. This reflects our scenario: *how can we automatically extract prerequisites from educational texts?* We provide our answer to this question with our automatic PR learning model described in 8. As we will discuss, our model relies exclusively on information that can be acquired from the raw text of the annotated corpus as we designed the model to value the textual content referring to the description and contextualisation of a concept. The need to adopt a similar criterion of extraction arises from the observation that this would be: *(a)* suitable for prerequisite learning also when external sources of structured information are not available; *(b)* capable to infer prerequisite relations directly from the educational material where concepts are described.

## 4.4 Chapter Summary

In this Chapter we discussed the issues of tracing prerequisite relations in educational texts according to the pedagogical perspective. In particular, we explored the role of linguistic complexity, as represented by three textual varieties targeting different audiences, on the task of manually creating propaedeutically-motivated sequences of concepts. Thanks to our investigation we concluded that linguistic complexity plays an important role in the PR identification task: propaedeutic relations between concepts emerge more clearly from simple texts, as we expected, but also from complex texts addressing an audience of domain experts. On the other hand, texts designed to target a wide audience convey PR relations between concepts less clearly. Although the association between linguistic complexity comprehension was already investigated in previous studies, our crowd-based experiment is the first, to the best of our knowledge, to target specifically PR relations and offer quantitative evidence about the impact of linguistic complexity on the identification of PRs.

The results of our investigation on the one hand confirmed our intuition on the challenges (but also benefits) of uncovering PRs from textual instructional materials, but on the other hand they also shed light on some novel issues that we addressed in our methodology for dealing with PRs. In particular, for what concerns the challenges of text-bound PR identification, we confirmed our intuition that finding PRs within the content of instructional materials requires to take into account the overall content of the resource as some relations might remain implicit. These types of relations in particular constitute an interesting challenge also for automatic PR learning systems. However, a text-bound annotation approach might lead the way to novel analyses on prerequisite relations carried out on the content of the texts where these relations are identified.

From our perspective, this represents an exiting opportunity as it creates the conditions from multi–disciplinary investigations that combine the linguistic, educational and computer science perspectives.

Concerning the novel observations emerging from the results of our investigation, it was brought to our attention that not all instructional materials are equally suited to uncover PRs. Apart from the complexity level of the resource, we should also take into account that, in order to model a subject domain based on the content organisation proposed by an instructional resource, we should rely on homogeneous and coherent resources. Textbooks fit well in this requirement: they are curated materials, explicitly targeting students at a specific learning level, and designed to convey all the knowledge required by a learner. Furthermore, we observed that our text-bounded PR annotation approach demands to take into account and revisit some of the common approaches adopted in the literature, in particular for what concerns the evaluation of the agreement between manually obtained annotations. Specifically, we should take into account the overall annotation graph rather than individual pairs as some relations might be implied by other PRs.

We incorporated the above observations into our novel methodology for dealing with PR, as well as the observations emerging from the comparison with past works within the same line of research. The next chapters will describe in detail each component of the PR framework and their application for the construction of a PR-annotated dataset on computer science concepts that will be used also to display the use of the PR learning module. By describing our annotation protocol and basic principles for manual annotation and automatic extraction, we aim to clarify:
*(a)* how to carry out PR annotation on an educational text and how we encoded the information;
*(b)* how to compare and evaluate the reliability of such manual annotation while also taking into account the peculiarities of the task and of PRs;
*(c)* exemplify which type of novel analyses can be performed on such annotated resources;
*(d)* show how we can obtain state of the art performances on automatic PR learning systems without relying on external knowledge bases but also on the content of the annotated corpus.

CHAPTER 5

# PROTOCOL FOR ANNOTATING PREREQUISITE RELATIONS IN TEXTS

In this chapter we will describe the set of recommendations and instructions that we defined in order to produce textual corpora manually annotated with prerequisite relations. These instructions were systematised within the *annotation protocol (**PREAP**)*, which will be described along this chapter. The development of PREAP and the definition of its principles are the outcome of our efforts addressed towards a consensual definition of the PR annotation process aimed at overcoming the limitations of current PR-annotation strategies discussed in Section 3.4. The multiple interpretation and definition of 'PR relation' led to different datasets which are usually produced not relying on well-defined annotation protocols, which, conversely, could support dataset reuse, comparison of results, other than higher reliability of the datasets. Our goal, on the other hand, is to define a thorough protocol for obtaining PR-annotated datasets in order to reach a commonly agreed treatment of the relation.

To pursue our goal, a *research team* involving researchers with different backgrounds undertook multiple tests to discuss and evaluate possible annotation criteria in a joint collaborative and multidisciplinary effort. The different perspectives contributing to the research led to the definition of the PREAP protocol, whose main novelty consists of formalising the principles of the pedagogical view, as described in the previous chapter. The development of PREAP annotation protocol took into account the desiderata and good practices for designing annotation tasks that we outlined in 3.1. For this reason, in this chapter we will first present the iterative process that we undertook in order to reach the current final version of PREAP, then we will introduce PREAP protocol and discuss its principles and annotation guidelines in detail.

## 5.1 Design of the PR Annotation Protocol

An *annotation protocol* consists of guidelines and specifications aimed at indicating how to obtain corpora enriched with explicit information regarding a certain phenomenon and that can be reproduced on any unannotated texts at any time [230]. As described below and detailed in Sec. 3.1, defining a procedure for annotating textual resources implies a continuous process of testing and validating annotation choices while keeping an eye on the recommendations for obtaining good-quality annotated data. The next sub-sections will report the process we undertook to define PREAP and our efforts towards addressing the good practices for modelling the annotation task.

### 5.1.1 Iterative Process of Protocol Development

The design of our annotation protocol was inspired by the methodological framework of the MATTER development cycle described in 3.1.1.1, which defines a general model for obtaining annotated corpora to be used in Machine Learning experiments. In particular, we took into account the recommendations provided within the MAMA sub-cycle, concerning the phases of *model* and *annotation definition*. The MAMA model fits our needs as it is agnostic to the decisions made regarding corpus selection, annotation tools and representation formats, which we dealt with in PREAP. As recommended by the MAMA design process, we carried out multiple cycles of testing and revision of our PR annotation protocol, detailed below.

The iterative approach that brought us to the final version of the annotation protocol is displayed in Figure 5.1.

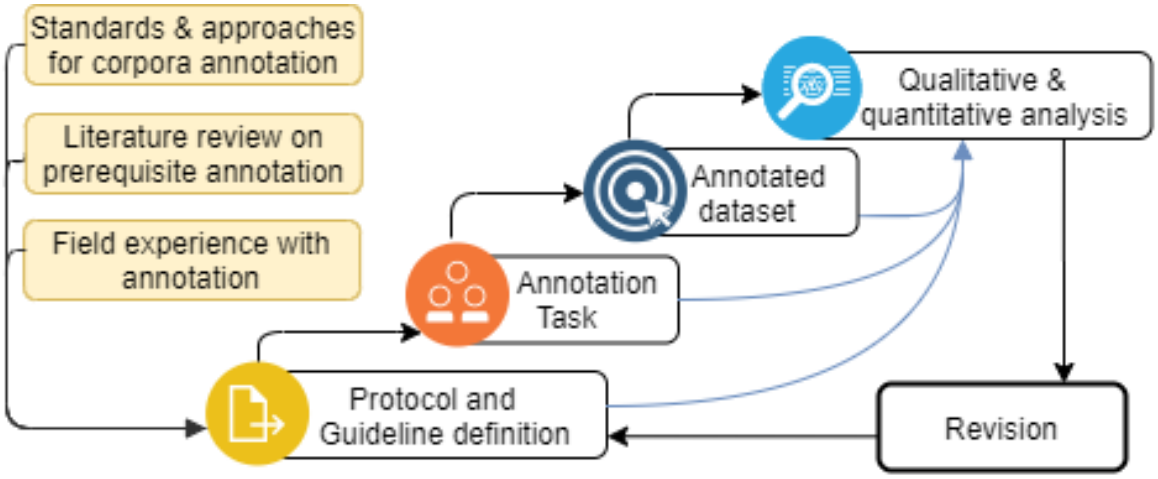



Figure 5.1: PREAP protocol: iterative design process.

As shown in the Figure, before dealing with the design of PREAP protocol and its principles, we performed some preliminary exploration in order to review existing literature related to standard approaches for corpora annotation and, specifically, prerequisite annotation. This allowed us to

*define the boundaries* of our problem, i.e., the interpretation of PR and the annotation model, which expresses how to perform annotation. The defined model was *tested* through annotation tasks that produced PR-annotated datasets, which we *analysed* to identify issues, addressed by *revising* the model.

Such 'definition-testing-evaluation-revision' cycle was performed three times, each resulting in a different version of the annotation protocol (namely, v1, v2, and v3). The revisions addressed the issues emerging from the application of the protocol to the annotation of textual data and were aimed at improving the richness and coherence of the annotations. Since developing annotation specifications and understanding the boundaries of a phenomenon mutually benefit each other [284], we believe that relying on such iterative design process, next to the literature review on corpora and PR annotation and field experience with annotation, eventually led to a shared vision among the members of the research team about how the PR annotation task should be defined and improved each time.

Currently, PREAP has reached its third version, which embodies the lesson learned from the iterative process of annotating and evaluating previous versions of the dataset. Before detailing the principles of PREAP for annotating PRs, we will discuss how we integrated the desiderata for annotation tasks into the protocol. Note that a systematic comparison of the three protocol versions, as well as the corresponding datasets, is carried out in Sec. 7.6.1.

### 5.1.2 Compliance with Annotation Tasks Desiderata

During the definition of the annotation task, we made an effort to comply with the existing desiderata for annotation tasks (see Sec. 3.1.2). Here we discuss the main issues that we addressed.

#### 5.1.2.1 Corpus Selection

As already discussed in 4.3.1.1, we decided to design the protocol for PR annotation having in mind a specific type of textual educational materials, *textbooks*. We made this choice for two reasons: (1) we assume that all concepts mentioned within the same textbook are related and discussed making reference to each other, and (2) they represent long and coherent explanations about the same subject domain. As a consequence of these, textbooks might represent the most challenging resource to be annotated: designing a protocol which is able to handle such type of text will make it potentially adjustable to all types of educational resources. For these reasons, we consider textbooks as a suitable resource to retrieve PRs and simultaneously put under stress our protocol.

A well-designed textbook is also supposed to have a good coverage of the topics in the subject domain. Considering the above as the general assumption that represent the foundations of our methodology, we recommend that each annotation project that wishes to perform PR annotation following our PREAP protocol should verify that the chosen text actually reflects such

requirements. This might seem trivial but is actually essential in order to guarantee the quality of the annotations. As discussed in 4.3.1, having an annotation strictly bound to the text has the advantage of making the annotation independent from any external resource but, at the same time, it reflects the author's organisation of the domain topics proposed in the resource being annotated. If a project relies on a poorly conceived instructional materials, the inappropriate concept organisation will be reflected in the annotation. Although this won't affect the possibility to carry out corpus explorations, it might affect possible applications of the dataset for, e.g., automatic curriculum planning. Hence, the selection of the text is possibly the most fundamental step of annotation projects relying on PREAP principles and it should be done carefully and considering the final goal of the annotation project.

Having a good quality corpus will also allow to reflect some formal requirements such as representativeness and balance. With respect to the latter, it should be noted that relying on a "balanced corpus" for PR annotation should simply mean having an amount of PRs in the corpus such that the goals of the project can be accomplished. Consider in fact that the number of non-PRs in a corpus (i.e., concepts not related by a prerequisite relation) is always higher than the number of PRs: they cover not only opposite relations (i.e., $multiplication \prec addition$, if $addition \prec multiplication$ is a PR) but also all pair-wise combinations of concepts not related by a prerequisite relation.

#### 5.1.2.2 Replicability

Replicability is an important aspect of NLP projects, and most notably of annotation protocols, since their purpose is to be applied to as many unannotated data as possible. The first step towards replicability consists of defining clear guidelines, i.e. the set of instructions to create the annotated corpus, aimed at describing the annotation task. In the effort to make our protocol as replicable as possible, we created an *annotation manual* comprising guidelines and questions to help annotators eliciting the knowledge they acquired from the content of the textbook. More details about the annotation manual are provided in Section 5.2.2.1 of this Chapter. Additionally, we recommend to include certain information about the project settings in any annotation report in order to improve the awareness of corpus explorers and to promote replicability. Specifically, we recommend to report the followings factors, inspired by those listed by [31]:
*(i) annotators*, i.e. number, recruitment process and criteria, training, expertise status with the text domain and annotation practice;
*(ii) corpus*, i.e. size, language, bibliographical reference, domain covered.

#### 5.1.2.3 Annotation Representation

As thoroughly discussed by Leech in [158], the annotation layer should be independent from the raw text of the corpus, which should always be recoverable. We achieved such multi-layered annotation thanks to a representation of the annotated information inspired by stand-off annotations

(see 3.1.2). Specifically, we separated the raw text, linguistic analysis and concept annotation levels from the prerequisite relations annotation level. This is done in order to preserve the integrity of the plain text but also to allow multiple annotations of PRs on the same text, possibly carried out at different times. More details on the annotation output will be provided along the discussion of the annotation protocol and tool (Chap. 6).

## 5.2 PREAP Prerequisite Annotation Protocol

The present Section deals with the basic principles of *PREAP* (PRerequisite Annotation Protocol), a protocol that formalises the process of manually annotating prerequisite relations on texts. In contrast with the definition provided by [94], which defines the annotation process as "any procedure or activity, at any level of granularity, involved in the production of annotation", we restrict the *'annotation process'* to the set of activities included in a specific annotation project, namely the definition of the project setting, the textual annotation and the corpus evaluation and analysis.

PREAP is thought to be applied in *PR annotation projects*, namely the set of activities aimed at building a dataset that includes explicit annotations about PR relations in educational texts. The protocol results from the effort of formalising our knowledge and considerations on prerequisite relations (see 2.2), acquired thanks the literature review, our own experiments and research backgrounds, into a set of recommendations for annotating PRs on educational texts. The ultimate goal of such process is the creation of *gold standard datasets*, i.e. manually labelled sets of items resulting from the annotation of a single expert or from the combination of all annotators' judgements that capture how the prerequisite relation is instantiated in the educational resources to be used as Ground Truth data [94, 286]. To this aim, PREAP relies on the most fundamental principle of the pedagogical perspective for uncovering PRs: rather than creating de-contextualised annotations, annotators must read a text and identify concept pairs connected by prerequisite relationship while reading, thus based on the content of the resource and not on their background knowledge of the domain. By doing so, we are able to simultaneously obtain *(a)* an annotation anchored to the content of the resource and *(b)* a text enriched with manually annotated relations, reflecting our intention of extracting knowledge from texts.

Considering our choice of adopting a text-bound annotation methodology, the resources we create are highly valuable for multiple purposes. For example, they can be used for training and/or testing the performances of automatic PR learning systems. Furthermore, they can be used to investigate how PR relations are realised within textual data. For these reasons, the protocol is meant to support the job of researchers working in the field of Computational Linguistics and NLP, who might want to use annotated corpora in their data-driven research, but also researchers working in the field of Education and Educational Technologies. The latter could use annotated corpora to carry out theoretical research about, e.g., how concepts are organised and

presented in different educational materials, or create their own PR-annotated corpora thanks to the lightweight annotation methodology which doesn't require extensive expertise in text annotation.

### 5.2.1 Before Annotating: Preliminary Decisions and Annotation Project Management

Annotating with PREAP principles requires to designate a *project manager*, who's in charge of taking decisions concerning the project goals and settings. Such decisions (which should be included in the project report) are preliminary to the actual annotation phase and concern:

a) *Annotation goal*: clearly define what the annotation is intended for and inform other people involved in the project about it;

b) *Textual corpus*: select the resource to annotate and perform any step of text pre-processing that may be necessary (e.g., perform and correct OCR, remove images, solve acronyms, etc.);

c) *Task settings*: determine how the annotation takes place and which annotation tool should be used, if any;

d) *Annotators recruitment*: define the ideal annotator's profile and recruit subjects if they comply with the requirements.

e) *Annotators training*: introduce selected annotators to the project and set up the pilot study to assess their understanding of the guidelines;

f) *Annotation revision*: decide whether and how to check and revise annotations;

g) *Agreement evaluation*: compute and evaluate inter-annotators agreement;

h) *Annotations combinations*: decide whether and how to combine multiple annotations in order to obtain a gold dataset (or *Gold-PR dataset* with reference to a Gold Standard annotated with PR relations).

### 5.2.2 Annotation Specifications

Here we present the annotation recommendations designed to help annotators carrying out the PR-annotation process. The first step when designing an annotation task is to define which phenomenon should be annotated and according to which rules and recommendations [230]. The research team designed the protocol pursuing the objective of supporting the creation of Gold-PR Datasets to be used as a ground truth for studying the linguistic realisation of PR relations in texts and for training and testing PR learning systems that exploit linguistic features. Considering these goals, our protocol proposes a PR annotation approach that anchors the annotation to linguistically-annotated texts. Having the text morpho-syntactically analysed allows to *i)* disambiguate concepts based on their grammatical category and normalised base form (i.e., lemma); *ii)* extract all contexts where concepts and their prerequisites occur; *iii)* identify the

syntactic structures underlying PR relations for further linguistic investigations. This approach is an innovative contribution with respect to current literature on PR annotation, which fosters analyses on PRs not otherwise possible, as we will discuss shortly.

PREAP high-level steps for the annotation process can be summarised in the following recommendations:

1. Find the relevant domain concepts from the educational resource;
2. Read the text and, if you encounter a concept that needs some prior knowledge to be understood, indicate its prerequisite concept(s) from those found at step 1;
3. Revise the pairs you detected reading again the portion of text where they were annotated.

Although simple in principle, these steps entail handling some methodological issues related to both the characteristics of the relation and the resource being annotated, that we formally addressed in the annotation manual.

#### 5.2.2.1 Annotation Manual

The annotation specifications are systematised within the *annotation manual*, fully reported in Appendix A and online[1]. The manual comprises two complementary resources: the *Annotation Guidelines* (AG), whose aim is to describe how the annotation process should be carried out in order to reduce inconsistencies in the annotations, and a list of *Knowledge Elicitation Questions* (KEQ), aimed at clarifying dubious cases through questions and examples. Both AG and KEQ are to be given to annotators prior to the annotation task in order to be discussed in the pilot study where the manager can test if annotators interpreted the instructions correctly. The manual remains at disposal of annotators throughout the entire annotation process.

AG in particular address potentially critical issues of PR annotation through 12 recommendations, concerning: concept identification (addressed by AG 1-3), text annotation (AG 3-6), PR features and properties (AG 7-9) and annotation revision (AG 10-12). We will now discuss them in detail and show how we propose to handle them in PREAP.

#### 5.2.2.2 Concept Identification

PRs are relations holding between two distinct concepts (namely, a *prerequisite* and a *target* concept). As discussed in Sec. 2.1, the term *concept* is very broad and its definition may slightly vary in different contexts. In PREAP, concepts correspond to domain–specific terms, either single or complex nominal structures, mentioned in the textbook (see 2.1.1). For our purposes, it is paramount to have a direct correspondence between a linguistic entity and the concept it represents since the annotation task consists of identifying PRs based on the content of educational texts.

[1]Available on the project's website `http://teldh.dibris.unige.it/pret/` and GitHub `https://github.com/Teldh/PRET`

As already mentioned for what concerns the sequence of concepts used in the crowd-sourcing experiment discussed in the previous chapter (see 4.2), concepts can have different granularity degrees, which may significantly affect the annotation in many ways. For example, it might be argued that the relation holding between the term used to denote a subject matter (e.g., *Algebra, Geometry, Physics*) and its topics (e.g., *addition, polygon, gravity*) is more a taxonomic relation rather than a PR. On the other hand, most taxonomic relations might as well correspond to a PR. For this reason, we account for such condition in PREAP by separating the concept identification and relations annotation steps. In this way, we create the conditions for letting the project manager define which is the desired level of concept granularity that should be preserved along the annotation. In fact, differently from [283], the identification of domain concepts in the text can be tackled in our protocol as an autonomous step of the annotation process and the project manager have to take preliminary decisions concerning the way concepts should be extracted and used. In particular, the manager can decide if *a)* letting annotators identify domain concepts while annotating the text with PRs, or *b)* providing a pre-selected list of validated concepts, i.e., a terminology, that the annotators have to use as–it–is, or *c)* providing a list that annotators can refine and update during the annotation. With option *a)* each annotator can define its own set of domain concepts, possibly obtaining a richer but less homogeneous annotation when compared with those produced by other experts on the same text. On the other hand, cases *b)* and *c)* require a preliminary step aimed at extracting the terminology from the text (a review of possible approaches was discussed in Sec. 2.1.2).

Note that the protocol does not impose any specific term extraction approach, however it recommends to take into account the text content since some strategies (e.g. supervised term extraction approaches) might require a certain awareness of the domain being analysed [191]. Defining in advance a list of relevant concepts, as for option *b)*, sets the desired granularity degree for concepts to be preserved along the annotation: depending on the final goal, the annotations should produce knowledge representations with, e.g., only high-level domain concepts (e.g., algebra, geometry, mathematics) as opposed to a rich fine-grained representation (e.g., radius, integer multiplication, fraction denominator). Case *c)* tries to balance pros and cons of *a)* and *b)*. Thanks to such approach, the project manager can decide what should be considered as a proper domain concept and what instead should be excluded from the annotation task.

#### 5.2.2.3 Text Annotation

The specific approach of our annotation protocol requires annotators to perform the annotation of prerequisite relations while reading the educational text, which implies identifying PRs through the explanations provided by the author to describe a new concept rather than relying on the annotator's background knowledge about the topic. In practice, reading the corpus and creating PR relations between an instance of the target concept in the text and its prerequisite concept should be done simultaneously. As a consequence, the annotation process doesn't produce

manually created non-PR pairs since this would make the annotation unfeasible: one would have to label at least $n(n-1)$ PRs, with $n$ being the number of concepts. Non-PRs remain implicit in the annotation, as well as non-annotated transitive relations, but they could be inferred through the asymmetric and transitive property respectively. Contrary to existing PR datasets that rely on external resources (see Sec. 3.4), our approach allows to capture the instructional design knowledge, namely the author's view on which concepts should be presented and how. This is particularly relevant if we consider that the content of a textbook is designed to guide students through a pre-arranged learning path designed to tackle relevant concepts and highlight their relations.

#### 5.2.2.4 PR Features and Properties

PRs, as intended in PREAP, are binary relations characterised by the properties of PRs outlined in 2.2.1, namely irreflexivity, asymmetry and transitivity. Such properties must be preserved in the annotation to avoid invalid relations from a structural and semantic point of view and, at the same time, to allow the acquisition of implicit relations from annotated pairs.

Among formal properties, we first mention irreflexivity: by definition a PR must involve two distinct concepts, thus self-prerequisites (e.g. *network* prerequisite of *network*) can't be allowed in the annotation. Moreover, the PR is an asymmetric relation: if concept $A$ is a prerequisite of concept $B$, the opposite cannot be true (e.g. if *network* is prerequisite of *internet*, then *internet* cannot be prerequisite of *network*). This rule also prevents the creation of loops in the annotation. Considering the semantic properties of the relation, PREAP accounts for different strengths of PR as a weight assigned by the annotator to each relation (s)he detects. The protocol suggests two categories: *strong*, to be assigned if the prerequisite is absolutely necessary to understand the target concept, and *weak*, to be assigned if the prerequisite is useful for a deeper comprehension of the concept but not strictly necessary. In order to identify PRs between concepts, it is also recommended to consider if the concepts are already related by some type of semantic relation (discussed in Sec. 2.2.3). KEQ (see Appendix A) specifically address this issue by offering examples of lexical taxonomic relations that might subtend PR, such as hyponyms, hypernyms and meronyms, or semantic relations like causal or temporal. These relations and their specific realisation in the text can sometimes cause divergent opinions among experts about the identification of PRs, thus the goal of KEQ is to provide examples involving commonly used terms in order to build a shared understanding about their interpretation.

#### 5.2.2.5 Annotation Revision

Annotation revision, to be performed after the annotation phase, allows to check if pairs created by an annotator comply with the formal and semantic requirements of prerequisite relations. Ideally, reconsidering PRs should be useful not only for identifying proper errors, introduced by the annotator by mistake, but mostly to think over hard cases. In fact, by double-checking the

annotations, a domain expert could easy understand if a relation was inserted intentionally or by mistake, but also reconsider annotation choices adopted at the beginning of the annotation process that changed as the annotator became more experienced (even unintentionally) .

In PREAP, we call the revision phase *"in-context revision"*. Revision here is aimed at identifying both proper errors and hard cases at once. We adopt an in-context approach since, in order to comply with the strategy adopted for creating pairs, the annotator is required to read again the portion of text where she/he found a PR relation before making the final decision of approving, excluding or modifying the relation. This choice is motivated also by the fact that previous work addressing automatic identification of annotation inconsistencies has found benefit to considering context when determining whether an ambiguous expression is inconsistently annotated [201, 205].

Indeed, manual revision is a very time-consuming process. For this reason, a consistent part of the body of literature on detecting annotation inconsistencies has focused on automatising the process of identifying similar instances that have been labeled differently [84, 120, 271]. For the time being, we didn't verify which one, among the methods available in the literature, is most effective to identify errors in PR annotation: we preferred to ask annotators to revise all their PRs since this approach is generally recommended when dealing with small datasets. Yet, a complete revision could be costly and tiring for the annotator, which could miss some pairs worth revising due to the long revision sessions. To address this issue, we defined an easy and simple way to balance the benefits and costs of revision. Relying on the intuition that variation in annotation can indicate annotation errors [84] and that the highest chance to find errors concerns phenomena rarely annotated [98], *we ask annotators to revise only those PRs identified by a low number of annotators*. Thanks to this approach, we significantly reduce revision time since we avoid revising those pairs that, being individually annotated by more than one annotator, might well be correct. Obviously, since we do not double check some of the pairs, few errors might still be present in the annotations, hence this revision approach should be used carefully only when the project goals allow it. This delicate decision should be taken by the project manager without involving annotators in order to avoid biased revisions.

### 5.2.3 Computing Agreement and Annotations Combination

If the project involves the revision process, once completed the project manager can use the annotations to produce the *Gold-PR dataset*, i.e. the Gold Standard Dataset of PR annotations. Such resource can result from a single annotation (relying on the judgement of a single trusted expert) or as a combination of multiple ones. In both cases, the goal is to create an error-free dataset, as coherent as possible, in order to obtain *(i)* informative analysis of its content, *(ii)* good performances of systems trained using it, *(iii)* reliable comparison with system outputs when testing their accuracies. When combining multiple annotations to create the Gold-PR, the manager should first account for annotations agreement and then choose the most appropriate

combination criterion.

#### 5.2.3.1 Agreement metrics

Inter-annotator agreement is computed in PREAP relying on the most prominent agreement metrics introduced in Section 3.2.2, namely Cohen's and Fleiss' *kappa* metrics. Cohen's $k$ is employed to compute pair-wise annotators agreement, while Fleiss' is used to compute agreement between more than two annotators. As discussed in 3.2.2 and 4.3.2.1, the classic implementation of *kappa* metrics hardly fit the characteristics of our PR annotation task. For instance, according to the metrics, two annotators agree on an item if the item received the same label by both annotators. This is limiting for our case as one annotator could leave a relation between two concepts implicit because (s)he created a path connecting the two concepts passing through other concepts (i.e., a transitive relation). To address this limits, we adapted the implementation of the metrics in order to account for relations that remain implicit in the manual annotation, namely transitive and negative PRs. Transitivity is specific to PR annotation, as we discussed in 2.2.1, so we propose to apply the $k$ metric to PR-annotations produced according to our protocol taking this property into account in order to obtain a more appropriate $k$ value. With respect to negative PR, which remain unexpressed in PREAP annotation approach, we would need to automatically acquire them by considering either all possible combinations of items or only inverse pairs (e.g., the pair $B \prec A$, if $A \prec B$ appears in the annotation).

In order to account for implicit relations, when computing *kappa* we assume that two annotators agree on the PR pair $A \prec C$ 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 $B \prec C$.

Consequently, the metric is 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 $T$, regardless the relation direction (i.e., $A \prec B$ and $B \prec A$ are both included in $P$). For each annotator, consider as positive PR each $p$ that is either manually created by the annotator or that can be derived using the transitive property. Consider $p$ as negative PR otherwise. Then, compute $k$ for each pair of annotators using the classical $k$ equation (fully displayed in eq. 3.1).

We do not provide a scale for qualitative evaluation of agreement scores in PR annotation, neither we recommend to rely on the traditional cut-off of 0.8 for distinguishing between reliable and unreliable annotations: PR annotation generally shows low agreement scores also when the datasets were successfully used to train machine learning algorithms for automatic PR extraction. However, one could rely on Landis et al. *kappa* interpretation scale [150] for qualitatively measure the results. High agreement is generally assumed as indicator of common understanding of the annotation specifications and the phenomena being annotated. In case of low agreement, the

annotation manager should, first of all, check the annotators understanding of the annotation specifications, and subsequently investigate possible issues with the annotation instructions [90]. During our protocol revision process, we were guided also by *kappa* scores obtained by our datasets in order to identify critical issues of our reviewed versions.

#### 5.2.3.2 Annotations Combination

Although combining individual manual annotation is the best option to capture the highest number of instances of the phenomenon, leveraging a commonly agreed set of annotated items from multiple judgements is not trivial. In fact, depending on the final goal, obtained agreement and level of expertise of the annotators, a consolidation procedure could be more appropriate than others. For example, considering as gold only the PR pairs inserted by all the users maximises the precision with respect to the ground truth and it's useful when the annotators are not expert since this way we have a higher degree of certainty for each pair. This approach is a sort of majority voting, a standard consolidation procedure, and it's optimal for creating datasets for training and testing automatic prerequisite learning systems since the result is a more homogeneous set of pairs. On the other hand, merging annotations into a single one means taking as valid PR each pair created by at least one user: this option maximises the recall since it allows to analyse every case where the experts believed to encounter a relation. For this reason, this choice is recommended when the annotators are experts and the boundaries of the phenomenon have some fuzziness that needs to be taken into account. In general, more inclusive combination approaches are to be preferred when the goal of the PR project is to analyse every case where the experts claimed to encounter a relation; e.g., this is the case when the goal is discovering linguistic patterns in PR's textual realisations, or when the annotators' judgements are highly reliable given a good domain expertise, provided that the revision of the annotation has been performed and the agreement score is not too low. Conversely, this approach is not recommended with low-experienced annotators and when the revision of annotations is not performed. Less inclusive combination approaches provide higher certainty and guarantee higher consensus about the relations included in the Gold-PR dataset, but provide more limited datasets, especially in case of lower agreement.

## 5.3 Chapter Summary

Along this chapter, we detailed the principles and motivations behind the instructions of PREAP annotation protocol. The main novelty of the protocol concern:

- The recommendation to annotate PRs while reading the text in order to create PRs motivated by the content of the annotated resource and to allow the exploration of the contexts where PRs are manually identified.

- The introduction of an in-context revision step into the PR annotation process.
- The adaptation of classical agreement metrics in order to account for properties peculiar to PRs, namely the transitive and irreflexive property.

Note that PREAP is language-agnostic, meaning that it could be applied to texts with no specific restrictions on the language. Given the novelty of the PR annotation methodology and in order to support the application of PREAP protocol during annotation projects, we designed an annotation tool, PRET, discussed in the next chapter.

CHAPTER 6

# ANNOTATION INTERFACE: PRET TOOL

In order to support the creation of annotated resources by following the principles of PREAP protocol, we developed an *annotation interface,* ***PRET*** *(PRerequisite Enriched Terminology)*. The interface was specifically designed to address the needs of an project annotating PRs with PREAP specification, thus putting into practice the basic principle of the protocol and guiding the annotator in their application. For this reason, PREAP and PRET annotation interface are deeply intertwined. In addition to textual annotation support, PRET offers also a set of functionalities aimed at quantitatively analysing and visualising the annotated datasets.

## 6.1 PRET Architecture and Functionalities

The literature reviewed on text annotation tool discussed in Section 3.3 revealed the lack of interfaces designed to address the needs of a PR-annotation project, and that also integrate fundamental functions of analysis. PRET (Prerequisite-Enriched Terminology) [1] annotation tool represents our attempt to fill this gap. PRET is an online annotation tool designed to support prerequisite relation annotation on educational materials, in particular it was designed to support PR annotation on educational texts in order to reflect the annotation principles of PREAP annotation protocol.

The tool is intended especially for researchers working in the field of NLP, AI and Education that want to study how prerequisite relations are established between concepts in educational materials, but it is potentially open to everyone, as the annotator only sees the raw text when annotating while the other annotation layers can't be perceived by the users. However, we observe

[1] Prototype available at `http://teldh.dibris.unige.it/pret/`

different levels of satisfaction depending on the user background, as we will discuss in Section 6.6.

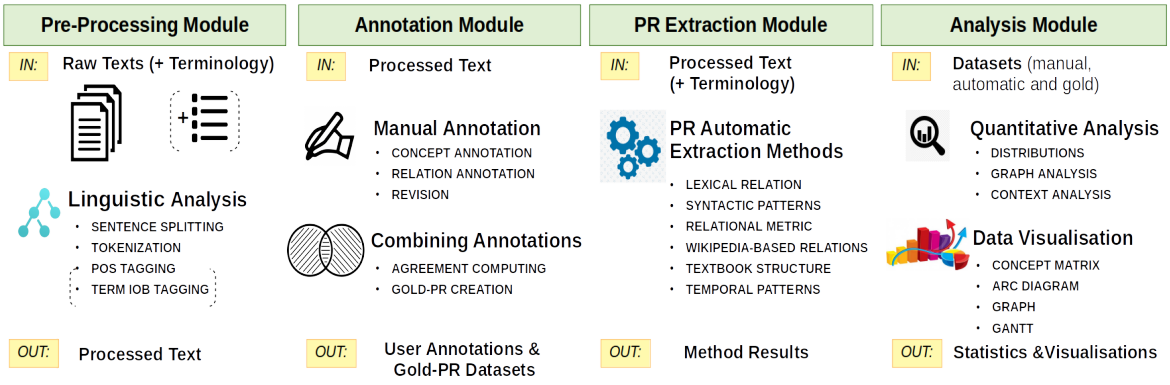



Figure 6.1: PRET tool architecture.

**Tool Architecture.** PRET offers different functionalities, each addressed by a specific module. The tool architecture shown in Figure 6.1 depicts how each module of PRET supports these functions. In particular, each module of PRET deals with a different phase of the annotation process, as detailed below.

- *Text upload and pre-processing:* the starting phase of each project, where the user uploads a textual corpus which is here prepared for the next phases;
- *Text Annotation:* where the pre-processed text of the previous step can be manually annotated by the user in order to obtain a gold dataset;
- *Automatic PR Extraction:* a module aimed at creating PR annotated datasets in a completely automatic way exploiting different PR extraction methods described in the literature and implemented in PRET;
- *Dataset Analysis:* a module aimed at performing quantitative analyses and visualising the annotations and gold dataset for research purposes.

In the next sections we will describe each module in detail.

## 6.2 Pre-Processing Module

The *Pre-Processing Module* is meant for project managers, who are in charge of setting up the annotation environment and prepare the data for the annotation step. Prior to the manual annotation, the manager of the project is in charge of selecting and uploading on PRET a textual corpus, i.e. the text (s)he wants to annotate with PRs, also including its metadata (i.e., corpus

name, authors, publication year, topic and language). Once the text is uploaded, the manager can see the text preview on PRET interface and insert information about the document structure (titles, sub-titles and paragraphs). Structure is annotated by ticking a checkbox associated to each paragraph to indicate if it corresponds to a title or sub-title in the text.

### 6.2.1 Linguistic Analysis

The tool automatically performs linguistic analysis on the uploaded text in order to acquire the morpho–syntactic structure underlying the content. PRET performs linguistic analysis exploiting UDPipe pipeline [259], a language-agnostic trainable pipeline for tokenization, tagging, lemmatization and dependency parsing. UDPipe produces a linguistically annotated text where the linguistic information is represented according to the UD formalism (see 3.1.2). Although any formalism could serve the purposes of the PR annotation, we chose to adopt UD for three main reasons: *(1)* UD is the most prominent formalism currently used for linguistic analysis; *(2)* being multilingual, it reflects our language-agnostic approach to PR annotation; and *(3)* the CoNLL-U format already includes extra fields to annotate miscellaneous information (such as concepts, as we will discuss below). Currently, in PRET tool UDPipe performs sentence splitting, tokenization and part-of-speech tagging using the pre-trained English model (version 2.5, the latest available at the time of the implementation). Although for now PRET supports linguistic analysis only for English texts, UDPipe can perform linguistic analysis on multiple languages, thus the tool could be easily expanded in the future to support other languages. As discussed, having a linguistically analysed text is essential when performing PR annotation according to PREAP principles: it allows to explore if there are any linguistic structures affecting the identification of PRs. The high attention on the interaction between text structure and content is possibly the main novelty of PREAP annotation protocol, which we carefully took into account while developing the functionalities of PRET. Note, however, that *the linguistic analysis is not displayed on PRET tool*, although it can be downloaded for local analysis. This choice is due to make the annotation easier to carry on also for 'annotation novices' (i.e., not familiar with corpus annotation practices or linguistic analysis formalisms).

### 6.2.2 Terminology Upload

The project manager can upload on PRET an optional terminology of domain terms extracted from the corpus. The terminology consists in a list of terms appearing in the text considered as particularly relevant for the text and/or the domain. Terms can be either single or multi-word noun phrases corresponding to domain concepts, either manually or automatically acquired. In case the terminology is uploaded on PRET, the lemma of each term is searched for in the text in order to find all its occurrences. During the annotation phase, each occurrence of the concepts will be highlighted for the annotator in order to indicate its presence and suggest that a domain concept is discussed in that text portion.

Formally, the presence of a concept is labelled on the 'miscellaneous' field on the CoNLL-U according to the "IOB" tagging scheme (i.e."Inside–Outside–Beginning"), originally presented in [232] and widely used in text chunking tasks such as Named Entity Recognition. This scheme has only three labels, which inspired its name: label "B-" is assigned to a token if it corresponds to the first token of an entity (here, a concept), "I-" if it is an internal or also the last token of an entity, "O-" if the token is not part of any entity. The "O-" label is actually deprecate and we frequently see texts annotated with "IOB" showing only the "I-" and "B-" labels. Consider as an example the sentence below, automatically parsed with UDPipe and annotated with "IOB" scheme to indicate the presence of concepts `algebra`, `mathematics`, `number theory`, `geometry` and `analysis`.

```
# text = Algebra is one of the broad parts of mathematics, together with
number theory, geometry and analysis.

1 Algebra algebra NOUN NN Number=Sing 3 nsubj _ B-
2 is be AUX VBZ Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 3 cop _ _
3 one one NUM CD NumType=Card 0 root _ _
4 of of ADP IN _ 7 case _ _
5 the the DET DT Definite=Def|PronType=Art 7 det _ _
6 broad broad ADJ JJ Degree=Pos 7 amod _ _
7 parts part NOUN NNS Number=Plur 3 nmod _ _
8 of of ADP IN _ 9 case _ _
9 mathematics mathematics NOUN NN Number=Sing 7 nmod _ B-
10 ,,PUNCT ,_ 3 punct _ _
11 together together ADV RB _ 3 advmod _ _
12 with with ADP IN _ 14 case _ _
13 number number NOUN NN Number=Sing 14 compound _ B-
14 theory theory NOUN NN Number=Sing 11 obl _ I-
15 ,,PUNCT ,_ 16 punct _ _
16 geometry geometry NOUN NN Number=Sing 14 conj _ B-
17 and and CCONJ CC _ 18 cc _ _
18 analysis analysis NOUN NN Number=Sing 14 conj _ B-
19 . . PUNCT . _ 3 punct _ _
```

Concepts represented by single-word terms are correctly marked by the "B-" label appearing in the miscellaneous filed of the CoNLL file. Multi-word concepts, such as `number theory` show both "B-" and "I-" labels, with the former marking the first token of the term, and the latter associated to the second token.

When a text and a terminology are uploaded on PRET tool, the corpus is stored on PRET as a CoNLL-U file where each sentence of the text is represented as in the example above. Currently,

PRET doesn't perform automatic term extraction. The role of the terminology is to support the annotator, for this reason it doesn't matter if the terminology is manually created, automatically acquired or not present at all: this decision depends on the project goal (e.g. the terminology can or can't be expanded during annotation) and it doesn't affect PRET functionalities.

## 6.3 Annotation Module

The *Annotation Module* supports the process of manually annotating the input corpus with PRs between pairs of concepts following the principles of PREAP annotation protocol. Moreover, this module handles annotation revision, agreement computing and the combination of multiple annotations in order to obtain a Gold-PR dataset. While the previous module was meant for the manager to set-up the annotation project, this module is mainly designed to support the work of annotators.

### 6.3.1 Concept and PR Annotation

The interface for manual text annotation is meant to facilitate concept pair creation while reading the text according to PREAP principles formalised in the Annotation Manual, accessible from the annotation page. Figure 6.2 shows the module interface: the left side of the image displays the annotation environment, while the right side shows the pop-up window for PR pair creation opening at double-click on a concept (example in the figure involving the concept *"number"*).

**Annotation Environment.** As can be noted from the figure, the annotation environment is divided into three sections, each handling a different functionality. On the left, the *"text annotation area"* displays the text to the annotator. Although the text was linguistically analysed in the pre-processing phase, annotators only see a plain text, which is easier to read. As anticipated in the previous Section, terms appearing in the uploaded terminology (if any) are highlighted in yellow along the text in order to be more visible, but they are also displayed in the *"concept sidebar"*, which is used to add new concepts as well. If the manager allows it, annotators can expand the terminology while annotating by selecting a concept occurrence in the text annotation area and drag the selected text fragment to the *"concept sidebar"*. Each occurrence of the new concept is then searched for in the text and highlighted in green so the annotator can easily recognise and distinguish them from terms appearing in the original terminology. If a concept was inserted by mistake, the bin icon in the concept sidebar can be used to delete it. Note that, while the terminology uploaded in the pre-processing step is shared by all annotators annotating the same text, the list of manually inserted concepts is personal for each annotator. In fact, adding a new concept to the terminology results from the annotator's reasoning about the content (s)he is reading and could, for instance, cause adding other concepts or creating PRs otherwise not possible.

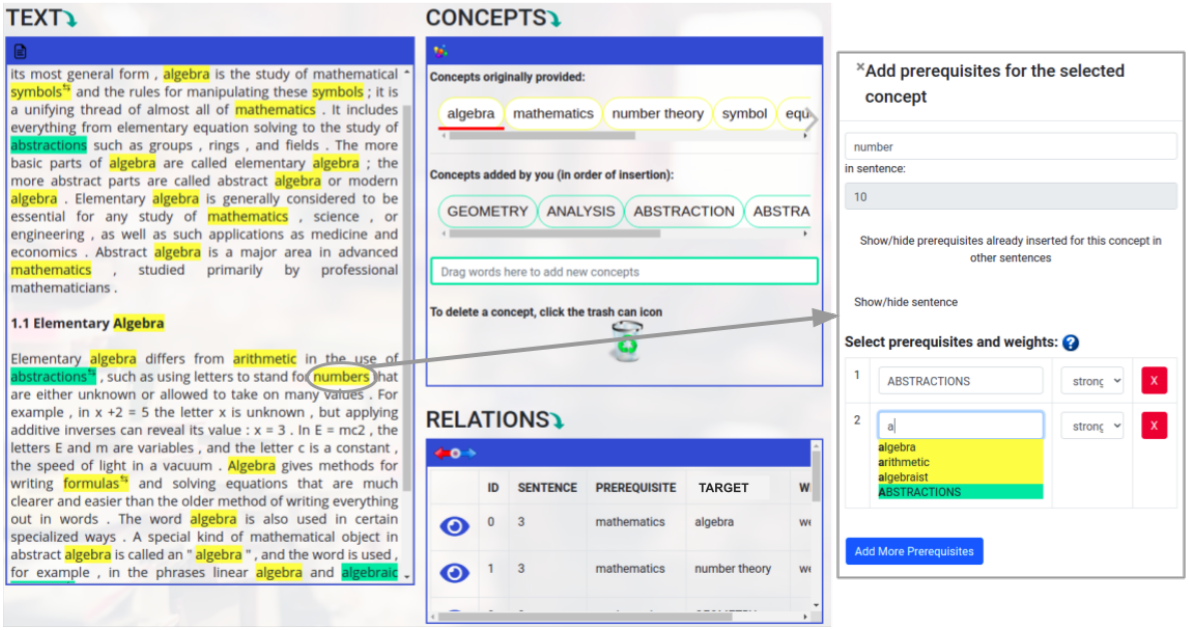



Figure 6.2: Annotation interface page of PRET tool (on the left), and window for PR pair creation of concept *"number"* as target (on the right).

**PR Creation Window.** In order to add a new PR, the annotator first selects a *target concept* from the "text annotation area" (a yellow or green highlighted term). Then, a pop-up window opens where the annotator can select the *prerequisite concept* from the terminology using a drop–down menu (see right side of Figure 6.2). For each pair, the annotator can also indicate the relation weight, choosing between 'weak' or 'strong', to make explicit if a prerequisite is necessary or only useful to understand the target concept as explained in the guidelines. The PRs are then saved and listed in the *"relations sidebar"*.

Considering that the same prerequisite pair can appear in different parts of the text with different strength, we assign to each annotated pair a unique context ID. The identifier also encodes the coordinates in the text (i.e. the sentence where it occurs) of the target concept, thus the same concept pair can be annotated multiple times in different part of the text and associated with different weights depending on the context where it appears. In summary, context ID, prerequisite concept, target concept and relation weight are the information encoded in a PR.

$$\text{PR} = (\text{ID}, \text{C}_1, \text{C}_2, w)$$

Following the principles of PREAP, creation of non-PRs isn't available on PRET since it would make the annotation process hard to carry on. On the other hand, PRET performs a cycle and transitive check for each newly created PR. When annotating, users have only a marginal view of their annotation. For instance, it is unfeasible to remember all previously inserted PRs, especially in the case of long annotation sessions. As a result, the annotator might accidentally

insert loops ($B \prec A$ when $A \prec B$ was already inserted) or transitive relations ($A \prec C$ when the annotation contains $A \prec B$ and $B \prec C$). PRET tool automatically detects these critical relations and supports validation check. The creation of a loop activate an error alert and can't be added to the annotation as they represent clear cases of error, whereas transitive relations, which might be actually correct, trigger a warning alert for annotators asking to check whether the relation has to be inserted or not.

**Annotation Output** Once the annotation is completed, the set of created PRs is saved on a separate file with respect to the linguistically annotated corpus. In summary, we have a sort of stand-off annotation output where, on one file, we store the information related to the text, linguistic analysis and concept occurrences (annotated on the CoNLL-U with IOB scheme); on a different file we store the PR triples. PR triples can always be traced back in the text as their context ID contains the coordinates to identify the occurrence of the target concept where the relation was entered. We store the text and PR information separately both for practical reasons (i.e., simplify data storage on the tool database) and to allow more flexibility of the annotations by supporting easy access to PR pairs in case a research team is not interested in perform linguistic analysis on the corpus and inclusion of multiple annotations of the same pair with different weights.

**Extra Functions.** As minor functionalities, the module also allows the user to insert bookmarks for labelling significant sentences and add free textual comments. Moreover, since the annotation usually requires a substantial amount of time, users can save their work so they can resume it later without loosing their annotations.

### 6.3.2 Annotation Revision

The tool offers revision support for users to check the correctness of their own annotation (*self-revision*). Revision is aimed at removing from the annotation relations produced by mistake or distraction. Performing this step is highly valuable for annotation tools since it allows to obtain more reliable and coherent datasets for performing further analysis [224, 230]. During self-revision, the tool presents to each annotator the pairs created during her/his annotation session for a pair-by-pair revision. As displayed in Figure 6.3, the interface is organised in two areas: the list of PRs is displayed on the left, while the area on the right side shows the raw text of the annotated corpus. Thanks to the support provided by the interface, revision can be carried out as an *"in-context revision"*: by clicking on a PR, the portion of text where the relation was entered is highlighted so the annotator can read it again to evaluate its correctness, namely check if her/his annotation was inserted intentionally, thus reflecting a real relation, or by mistake. The in-context revision is essential considering the annotation approach defined by PREAP protocol, which promotes the creation of pairs based on text content.

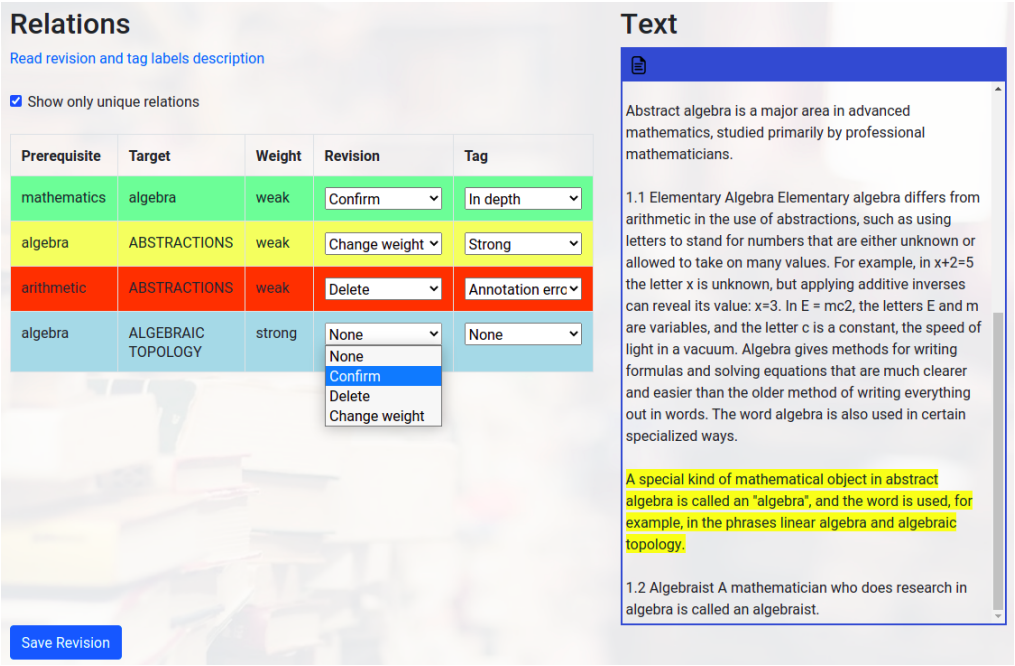



Figure 6.3: Revision interface in PRET tool. The text is taken from the usability test.

After reading the text, the annotator can decide whether to approve, exclude or modify the weight of the relation choosing the appropriate label (i.e. 'Confirm', 'Delete', 'Change weight') from those displayed in the *'Revision'* menu. The labels in the *'Tag'* menu allows to add further information about the revision decision. In particular, if a pair is confirmed, the user can indicate the type of prerequisite relation occurring between the target and prerequisite concept, or, if a pair is to be deleted, the motivation for removing it from the annotation. In case the user wants to change the relation weight, the 'Tag' label shows the newly assigned weight value (i.e. weak if it used to be strong and vice-versa). Revision labels are described in the next section (see 7.3.1).

In order to maximise the outcomes of revision while reducing its costs, we enabled in PRET the possibility to perform revision only on a motivated selection of PR pairs for each user. By ticking the box "*Show only unique pairs*" (see Fig.6.3), the tool shows to the user the PRs created only by himself after comparison with other annotations produced for the same text (if any). Indeed, it is known that rare phenomena in annotated corpora might be errors due to annotators' distraction or misunderstanding of the guidelines [98], so there are higher chances to find errors among them. Non–revised pairs are automatically confirmed.

### 6.3.3 Combining Annotations

Annotations can be used to obtain *Gold Standard datasets*, or Gold-PR in the case of a gold dataset annotated with PR relations.PRET supports annotation combination, but also agreement computing in order to evaluate the agreement between annotators before combining them and,

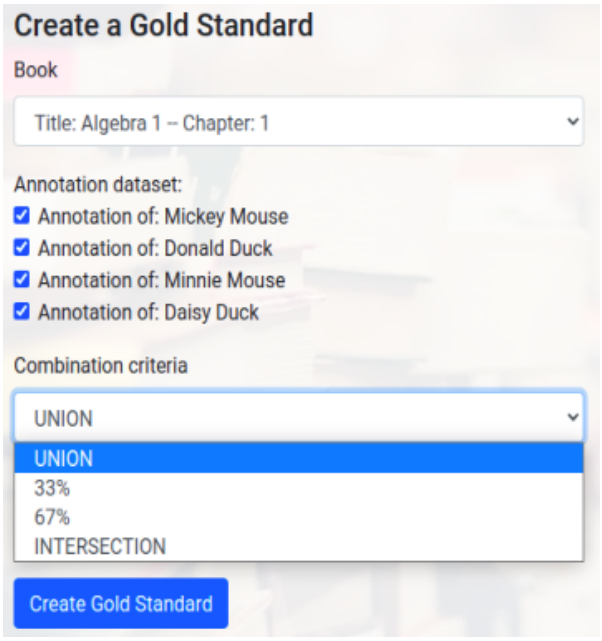



Figure 6.4: Interface for gold standard dataset creation.

eventually, remove unreliable annotations. Both can be carried out by the project manager after the annotation and revision processes are concluded.

#### 6.3.3.1 Agreement Computing

In line with PREAP principles, PRET tools implements two agreement metrics: *1)* Cohen's *kappa* [64] (to compute pair-wise agreement) and *2)* Fleiss' extension of *k* [95] (to compute group agreement). As discussed in 5.2.3, PRET implementations of *kappa* account for implicit transitive PRs and automatically acquires negative PR examples among concept pairs which are neither transitive PRs or positive relations manually included in the relation. This allows to obtain more accurate *kappa* values since we consider the annotations as wholes rather than considering individual sub-problems, as usually done in the literature. Currently, our agreement calculation doesn't put restrictions on the number of intermediate pairs involved in a transitive relations, but we plan to include it in further refinements of the tool.

#### 6.3.3.2 Gold Creation

The gold creation interface implemented in PRET, displayed in Figure 6.4, allows the project manager to select which individual annotations include in the gold dataset[2] and offers multiple combination criteria associated with tips for choosing the best option on the basis of each project

[2]At this stage, individual annotations should be already checked and revised by each annotator, as explained in the previous section.

needs. The most inclusive option available consists of taking the *Union* ∪ of annotations, namely including all PRs identified by at least one annotator. On the other side of the spectrum, the *Intersection* ∩ means including only shared PRs (i.e., PRs detected by all the annotators). In between of these two approaches, other options weight each relation based on the number of users that identified it: the manager can thus identify the most appropriate threshold for including or excluding PR pairs from the gold dataset. To implement this option, PRET dynamically defines inclusion thresholds based on the number of annotations that the user selects for building the final dataset: if, as in Figure 6.4, the user decided to create a gold dataset using four individual annotations, choosing '67%' means including PR annotated by non–less than half of the annotators. In other words, the PR is included if at least tree users included it in their annotations. Once the consolidation criterion is defined, the user can rename the gold dataset (changing the default name) and save the dataset that will appear in the list of gold datasets available for download.

## 6.4 PR Extraction Module

The *PR Extraction Module* offers a series of methods for performing automatic extraction of prerequisite relations from the texts uploaded on PRET. Those unsupervised methods do not need any gold data for training and they generate automatically annotated datasets that can be compared against the manual annotations or used to perform semi-automatic PR annotation.

Although we didn't use the functionalities of this module in the annotation project described in the next chapter, we list the automatic PR extraction methods implemented in PRET for a matter of completeness in the description of PRET tool.

1. *Lexical Relations*: this method indicate the presence of a prerequisite relation between any two concepts of the terminology showing a lexical relation, such as hypernyms-hyponyms and holonyms-meronyms, which frequently occur in PRs. The presence of the lexical relation is determined using those appearing in Wordnet lexical database[3].

2. *Lexico-Syntactic Pattern Match*: this method is based on the same principle as the one above, however it doesn't rely on any external knowledge. The presence of a lexical relation between two concepts is identified directly on the text using pattern matching. Specifically, the method relies on the linguistic patterns defined in [283], specifically conceived to detect PRs rather than other relations.

3. *Relational Metric*: this method exploits the Ref-D metric [165] for prerequisite relation identification. The metric computes the associative strength between two concepts by considering the number of hyperlinks existing between the Wikipedia pages associated with each of the concepts. This method generally performs well and it's not affected by corpus size.

[3] `https://wordnet.princeton.edu/`

4. *Wikipedia-based relations*: the method based on Wikipedia pages, described in [283], extracts domain key concepts from educational resources and then identifies prerequisite relationships between them. The key concept extraction exploits Wikipedia to construct a domain dictionary in which each term of the resource is promoted as domain concept if there is a Wikipedia page having the term as title. Then, prerequisite relation identification occurs through a combination of three metrics:

   - *Usage in definition.* Definitions often convey prerequisite relations: if concept $A$ is used in $B$'s definition, then $A$ is likely to be $B$'s prerequisite. For extracting definitions from Wikipedia pages, [282] assumes that the first sentence in each page is a definitional sentence for the concept described in the page.
   - *Content Similarity.* If two pages cover similar topics, it is likely that the two represented concepts have some learning dependencies, i.e. either $A \prec B$ or $B \prec A$. Lexical similarity between Wikipedia pages can be measured with cosine similarity, as done in [282]. This assumption, however, is critical for two reasons: content similarity i) is a necessary but non sufficient condition for PR (not all pairs of pages with similar content have a prerequisite relation), and ii) does not tell us the direction of the relation (if it exists). For these reasons, [282] proposes to identify pages that may be covering similar topics but are not at the same level of learning, as explained below.
   - *Learning Level.* Concepts with a lower learning level should be indeed learned first. According to [281], such level can be inferred with two features:
     - Range of topic coverage: the more topics that a concept covers, the more basic the concept is. To do so, [282] runs a topic model on the collection of Wikipedia pages representing concepts to generate topic distributions for each concept.
     - Number of in-links and out-links: considering the graph nature of Wikipedia, cross-page links between its pages can be useful for detecting concept learning levels. According to [282], If $A$ receives numerous in-links from other concepts, it is likely that $A$ is a fundamental domain concept and thus should be learned first (a similar conclusion can be drawn when counting the number of out-links of a page).

5. *Text Structure*: TOC (Table of Content) distance between two concepts, described in [283], computes the relation strength between the concepts as the distance between the sub-chapter numbers where they appear. As evident, this method can be applied only if the corpus shows a TOC or, at least, a hierarchical structure corresponding to chapters and sub-chapters.

6. *Temporal Patterns*: this method correspond to one of our approaches for PR extraction described in [1]. The method exploits Burst Analysis to identify relevant portions of texts

(burst intervals) for each concept and then exploits a set of heuristics based on Allen's temporal patterns between burst intervals to find concept pairs showing a prerequisite relation. This method is described in more detail in Section 8.2.2.1.

Methods 1, 2 and 4 return a categorical value (binary output) for each concept pair, thus naturally produce a binary annotated set of concept pairs as with manual annotations. On the other hand, methods 3, 5, and 6 return a continuous variable for each pair: in order to eventually produce a binary labelled set of concepts, these methods require to manually define a threshold to discriminate between pairs showing a PRs (obtaining a value above the threshold) and non-PR pairs (if the value is below the threshold).

## 6.5 Analysis Module

Taking as input the annotations resulting from the previous modules, namely gold standard datasets, the annotations of each user and the results of automatic extraction methods, the *Analysis Module* provides an overview of the annotation characteristics and, if multiple annotations are available, similarities and differences between pairs of annotations in the *'Comparison'* page. This module provides analysis at various levels, designed to address different goals: visualisation methods are useful for those who aim at building training dataset for automatic systems, while linguistic and context analysis are more valuable for studying how PR are instantiated in the texts. In what follows we will describe each feature implemented in PRET to support analysis and visualisation.

### 6.5.1 Quantitative Analysis

Quantitative analysis allows to take a deeper look into the annotations and explore what affects annotation similarity or discrepancy. The quantitative analyses are distinguished between *Data Summary* and *Linguistic Analysis*.

**Data Summary.** *Data Summary* reports descriptive information about the annotation. Applied to individual datasets, it is organised into three sections: *a)* text information, showing the number of sentences and tokens of the annotated text, *b)* concepts and relations distributions and *c)* concept graph information. Concerning *b)*, the tool details how many concepts are involved in the annotation, either overall, belonging to the original terminology or entered by the user. With respect to PRs, PRET reports the amount of entered relations, the number of PRs showing a weak or strong weight, and the amount of unique pairs[4]. As to the concept graph information, they are acquired from the graph generated from the annotation (representing the annotation information as explained in 2.2.4). Specifically, considering the main properties of the prerequisite relation,

[4]Considering that the same concept pair can be inserted multiple times in different parts of the text, a pair is unique if it is entered only once.

PRET performs analysis on transitivity, connectivity, loops, average in-degree and out-degree, disconnected nodes, diameter of the network, longest path and number of roots and leaves.

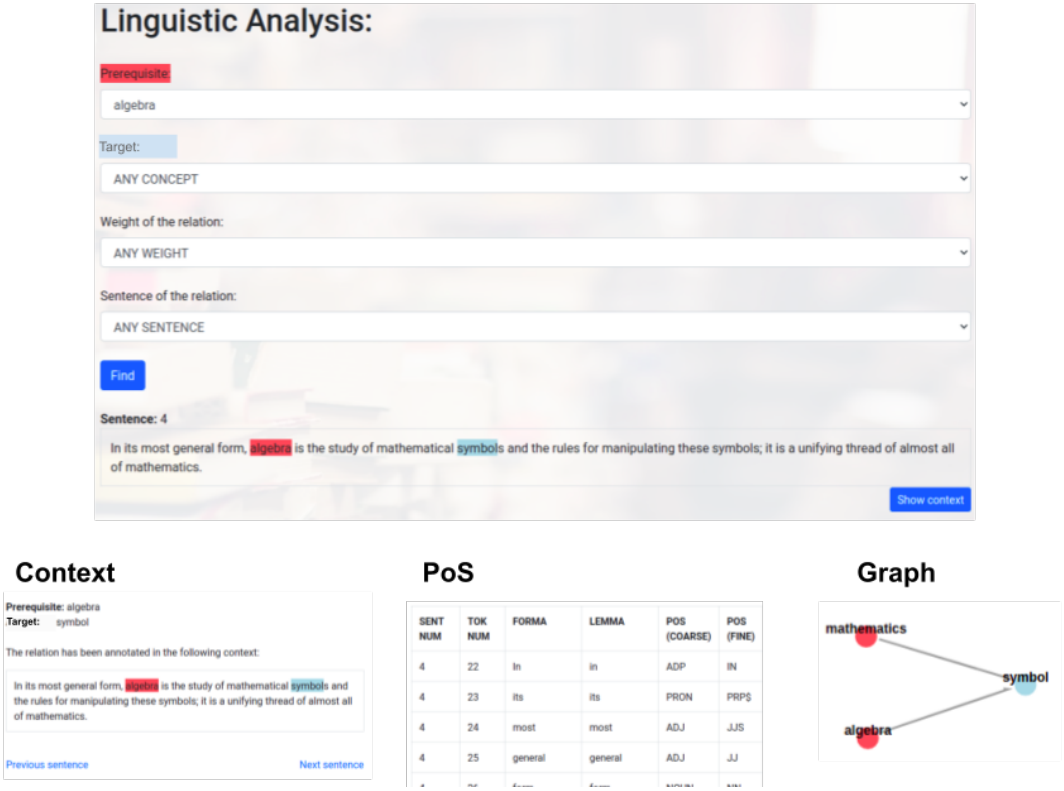



Figure 6.5: Linguistic Analysis (on top) and detailed analysis windows (bottom).

**Linguistic Analysis.** Next to the quantitative distributions, *Linguistic Analysis* offers the contextual analysis of each PR in a dataset. Context analysis allows to retrieve the textual context of an inserted PR and to investigate its linguistic features. The latter is possible thanks to the morpho-syntactic analysis performed in the pre-processing phase, while the former results from the act of anchoring the PRs to the sentence where they were identified. The context analysis comes with an interface called *Prerequisite In Context* (PIC) (see Figure 6.5), inspired by Key Word in Context (KWIC) analysis [178], that allows the user to query the annotated corpus and retrieve all the relations that match the querying criteria. The user can create filters using drop-down menus selecting the prerequisite and target concepts, the relation weight and the sentence where the PR should be searched. In the example of Fig.6.5, the query searches for PRs created anywhere in the text having 'algebra' as prerequisite, and it retrieves the relation between 'algebra' and 'symbol' annotated in sentence 4. The *"Show Context"* button, available for each relation, shows more information about the retrieved relation. Specifically, the user can *i)* read the sentence where the queried relation occurs and its surrounding context (Context window, at the bottom of Fig.6.5), *ii)* see the morpho–syntactic analysis of the sentence (PoS window) and

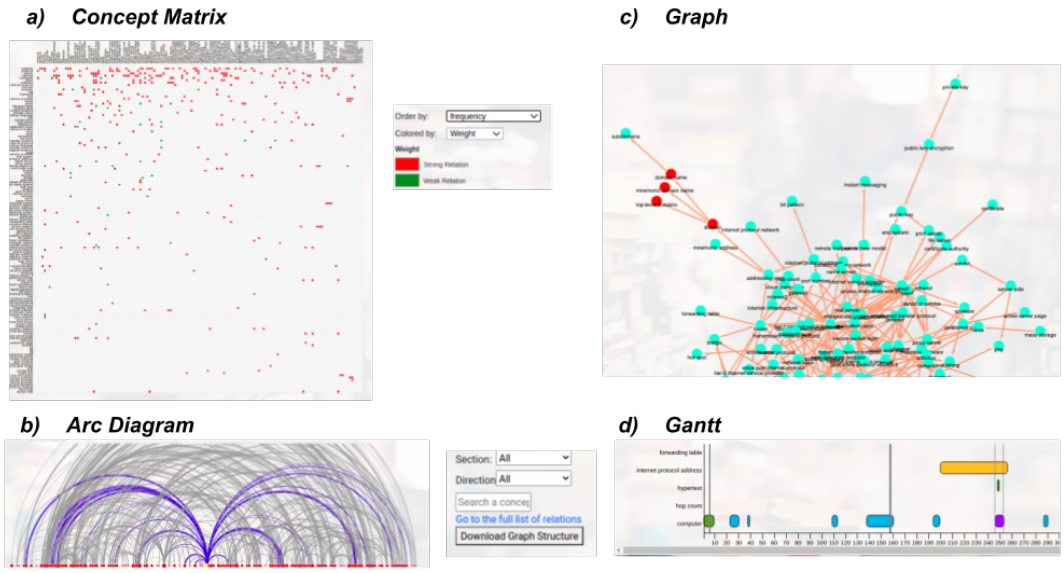



Figure 6.6: Visualisation methods implemented in PRET.

*iii)* explore the sub-graph showing incoming and outgoing edges from the node representing the the target concept in the relation (Graph window).

### 6.5.2 Data Visualisation

Visual representation of PR relation is mainly thought for researchers aiming to explore the annotations by means of several dynamic and interactive graphic representations. For the purposes of the analysis presented in [218], we tested a range of visualisation techniques. Based on the results obtained and lesson-learned, we implemented in PRET the visualisation techniques that we considered as the most appropriate to visualise educational relations in texts and to guide the research involving PR-annotated datasets. The four chosen techniques are displayed in Figure 6.6 and described below.

**a) Concept Matrix.** A dynamic and interactive representation of a $|T| \times |T|$ asymmetric adjacency matrix $M$, where each coloured cell $M_{i,j}$ represents a prerequisite relation between concepts $i$ and $j$. The matrix arrangement is dynamic: the concepts along the axes can be sorted based on *i)* alphabetical order, *ii)* occurrence in the text, *iii)* frequency of mentions in the text (from higher to lower, as in the case depicted by the Fig.6.6) and *iv)* cluster membership[5]. Red and green coloured cells distinguish strong and weak PRs. If $M$ is used to visually depict a gold standard, different colours represent the degree of raw agreement (i.e., how many users inserted the relation).

[5]As detected using the Infomap algorithm described in [239]

**a) Arc Diagram.** A flat representation of a matrix: concepts are displayed along the horizontal axis (red dots in Fig.6.6) and arcs represent prerequisite relations connecting them (changing colour when mouse-over). Through filters, the user can visualise only pairs annotated in specific sections of the book or having a certain direction, namely 'forward' if the prerequisite concept is mentioned in the text before the target concept, 'backward' otherwise.

**a) Graph.** A network-like representation of the dataset where concepts are represented as nodes and edges represent prerequisite relations. This visualisation is the most closely resembling a concept map. It is useful to detect explicit transitive edges. Additionally, it allows to explore sub-graphs belonging to individual annotators, helping the analyst to investigate differences in annotations produced by annotators with a different profile. However, as the dataset becomes larger, graph visualisations become harder to explore, thus we plan to add filtering functions in the next release of PRET to support graph exploration.

**a) Gantt Chart.** A Gantt diagram showing the progression of concept relevance, as detected by the Temporal Patterns method (see Sec.6.4), through time. Each concept of the terminology is associated by the Temporal Patterns method to one or more intervals of text where they are particularly 'relevant'. These intervals are represented in the diagram as coloured blocks (see Fig.6.6). The horizontal axis represents the progression of time, measured in terms of sentences (meaning that sentence 1 represents time $t_1$, sentence 2 represents $t_2$, sentence $n$ represents $t_n$ and so on). Concepts are arranged along the vertical axis, according to their temporal order (i.e., by first appearance in the text). The Gantt visualisation is aimed at supporting the comparison between pairs of relevance blocks referring to different concepts, thus facilitating the analysis of temporal patterns occurring between concepts along the text flow.

## 6.6 PRET User Evaluation

To measure the quality of the proposed annotation approach and its implementation in PRET tool, we carried out evaluation tests aimed at answering the following questions:

Q1) Are PREAP methodology and its implementation in PRET tool good enough to support annotation and to satisfy the requirements of PR relations annotation in educational texts?
Q2) Are users able to use PRET tool and understand its intended purpose?
Q3) What is the users' perception of the system (PRET) usability?
Q4) Is there an effect of users' expertise (either with annotation tasks or IT tools) in the use of the tool?

The first question (Q1) is addressed in the annotation project described in chapter 7, involving four participants for crating a PR-annotated dataset. The other three questions address the

matter of PRET *usability*. ISO standard 9241-11 defines usability as "the extent to which a product can be used by specified users to achieve specified goals with *effectiveness, efficiency and satisfaction* in a specified context of use" [34]. In order to test PRET usability, we recruited 12 participants to perform some guided tasks on the tool and answering standard usability questionnaires. In particular, we conducted *formative testing* (i.e. aimed at diagnosing and fixing problems) on the demo version of PRET currently available online in order to identify critical issues and verify the use of the tool on different populations of users. The latter issue, namely the use of PRET by users with different backgrounds, is far from trivial since the tool groups several functions for different goals.

Note that PRET usability evaluation reported here is part of an ongoing study aimed at investigating the above issues. For the purposes of this dissertation, we briefly report the main results emerging from the tests. However, note that a more thorough analysis will be reported in the future as part of a dedicated study.

### 6.6.1 Methodology

PRET quality and usability was tested using standard usability tests, which we employed to answer the questions outlined above. Specifically, questions Q2 and Q3 address the standard definition of usability provided by ISO 9241-11 cited above, based on effectiveness, efficiency and satisfaction. Q2, in particular, is related to **effectiveness** and **efficiency**. Effectiveness measures the accuracy and completeness with which users achieve specified goals; efficiency measures the accuracy and completeness of goals with respect to resources (i.e., time) expended. Q3, on the other hand, concerns the users' perception, thus it is more related to the overall **satisfaction** of the user. In order to investigate Q4, which deals with the effect of **expertise**, we relied on users background, discussed in 6.6.1.2, to investigate whether different backgrounds are associated with different results on the tests.

#### 6.6.1.1 Tests Setting

Each usability test involves two subjects, namely a user and an experimenter who supervises the test. All tests were performed in remote and moderated mode. In particular, experimenter and users were involved in a recorded video call where the user is asked to share the computer screen and, if comfortable, also comment his/her actions while performing the test (*think-aloud protocol*), while the experimenter speaks only if the user requires a guide to proceed in the test.

Tests were organised into the following 3 phases.

1) The first phase starts one day before the actual test, when users are given the PRET Quickstart guide[6] (no link to PRET is included) and the Test Document, i.e. the list of tasks;

[6] https://github.com/Teldh/PRET/blob/master/docs/PRET_Quickstart_Guide.pdf

2) In the second phase, first the experimenter and the user review together the test documentation (i.e., Quickstart guide and Test Document) in order to address possible user's doubts. Then, the test begins and users are asked to solve a set of tasks on PRET during the video call with the experimenter;

3) After taking the test, users answer two questionnaires to capture their qualitative evaluation of the tool and the overall experience of using it.

PRET Quickstart guide is a sort of user manual of the tool, where instructions for using PRET functionalities are completed with screenshots of the interface. The Test Document (fully reported in C) contains a scenario and 5 tasks to perform on PRET, organised in 21 sub-tasks (at least one task for each tool module). The scenario describes a plausible use case to users, thus we ask them to keep the scenario in mind while solving the tasks. While the user performs the test, the experimenter assigns a *success score* to each sub-task based on user performances. The success score ranges between 0 and 2: 0 is assigned to a task when the user gives up the task without completing it; 2 is used when the user is able to solve the task easily without any help and without making mistakes; 1 is used when the user can solve the task only with the help of the experimenter. In-between scores are used to mark the gradient of errors that can be encountered. From our point of view, a task is considered successfully completed when the user obtains a score higher than 1.5.

After completing the tasks, users were asked to take two standard usability questionnaires, SUS (System Usability Questionnaire) [41] and PSSUQ (Post-Study System Usability Questionnaire) [160], measuring satisfaction with using the tool. We included both questionnaires in our study since they capture usability from two complementary perspectives: the former allows to investigate learnability with the system, while the latter better captures information quality. We didn't modify the questions of SUS and PSSUQ, although we added few non-mandatory open questions to allow users share their thoughts on the system, if they want.

#### 6.6.1.2 Participants Sample

The evaluation tests involved 12 participants, 8 males and 4 females, ranging between 25 and 40 years of age. We conducted a preliminary interview in order to acquire information about users' profiles. Among them, we selected 6 users as "experts" and 6 as "non-experts". The distinction is mostly based on their background: the 6 users belonging to the *expert group* have a computer science or engineering degree, thus we assume that they all have a certain confidence with data analysis and IT tools. 4 of the 6 experts claimed to be also familiar with tasks related to prerequisite annotation and/or extraction. Non-experts were selected as humanities graduates with little or no expertise on tasks related to prerequisites annotation or even textual annotation in general. Apart from 1 participant (user n.12), non-experts claimed themselves as having little

interest in computers and IT tool and reported that, in their everyday professional life, they use a limited number of computer facilities.

### 6.6.2 Results of the Usability Tests

The fist step of our analysis is aimed at answering Q2 above: are users able to use the system and understand its purpose? As said, this question captures aspects of system effectiveness and efficiency. We measure effectiveness in terms of *completion rate* obtained by users on each sub-task. The completion rate is a standard usability metric which represents the ratio of successes with respect to failures on a given task. Namely, given the number of tasks undertaken, compute how many users (%) achieved a success score higher than 1.5. Results of the completion rate analysis (i.e., average completion rate for task and their standard deviations) are reported for each of the 21 sub-tasks in Figure 6.7.

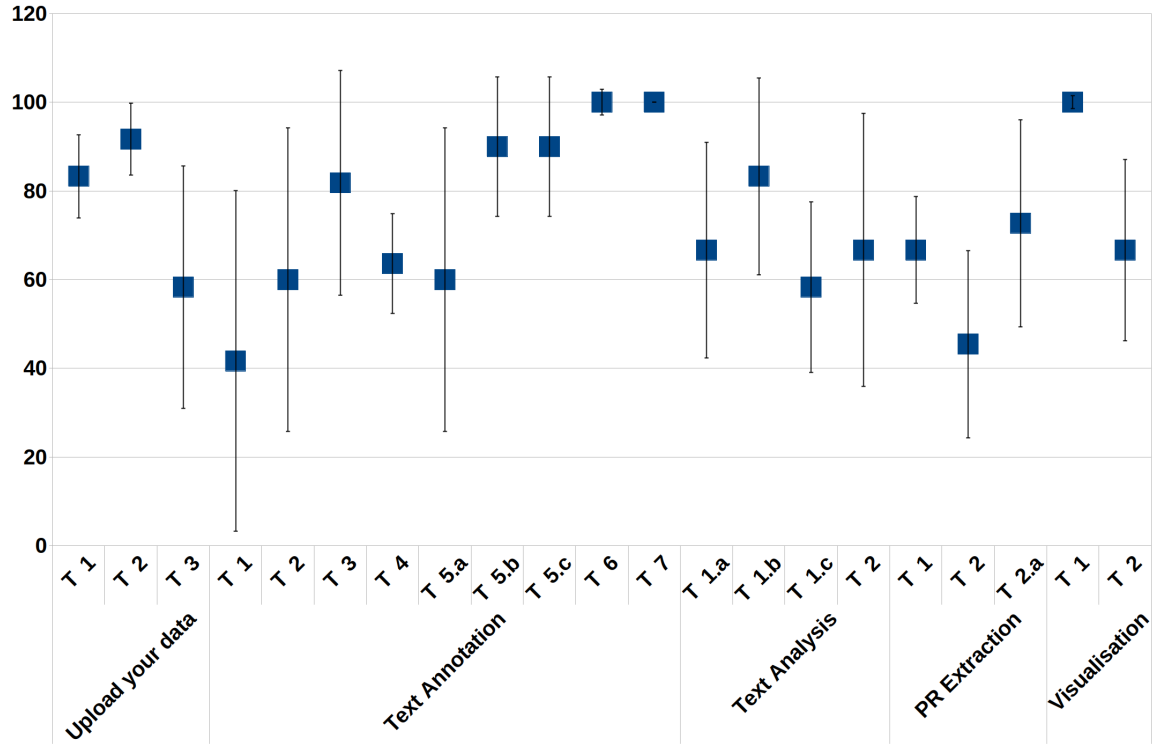



Figure 6.7: Effectiveness analysis: completion rate for each sub-task of the usability test (i.e. each dot represents a task). Error bars represent standard deviation values.

The picture depicted by the plot in the Figure suggests that, overall, the tool is effective for its purposes. Indeed, the completion rate are quite high in all cases: in three cases they reach the maximum value of 100% (meaning that all users successfully completed the task) and only in two cases the score goes below 50%. As further proof, if we compute the average success rate (i.e., the score assigned by the experimenter to measure the user performance on the task) for the overall

test by considering the success rates obtained by all users for each sub-task, we obtain 1.74[7] (normalised as 86.57%), meaning that in almost all cases users were able to successfully complete a task. However, the graph highlights an uneven difficulty of the sub-tasks. Text annotation tasks, for example, obtain the lowest scores, in particular for what concerns those tasks addressing the creation of PR pairs ('T1' and 'T2'). This could be either due to a difficulty inherent to the task or to a misplacement of functionalities on the interface and should be further investigated in the future.

Efficiency, on the other hand, is computed in terms of time (in seconds) spent on a task by a participant. If we look at the graph in Figure 6.8 we notice that, expectedly, the tasks achieving lower completion rates also take much longer than the others to be completed. Note that here we don't distinguish between successfully completing of failing the task. Such results is partially expected, as the novelty of the annotation task might require a certain amount of learning time, however it also suggests that the issue with the annotation interface is possibly related to the arrangement of the tool functionalities in the corresponding annotation interface module, which we plan to revise in the near future.

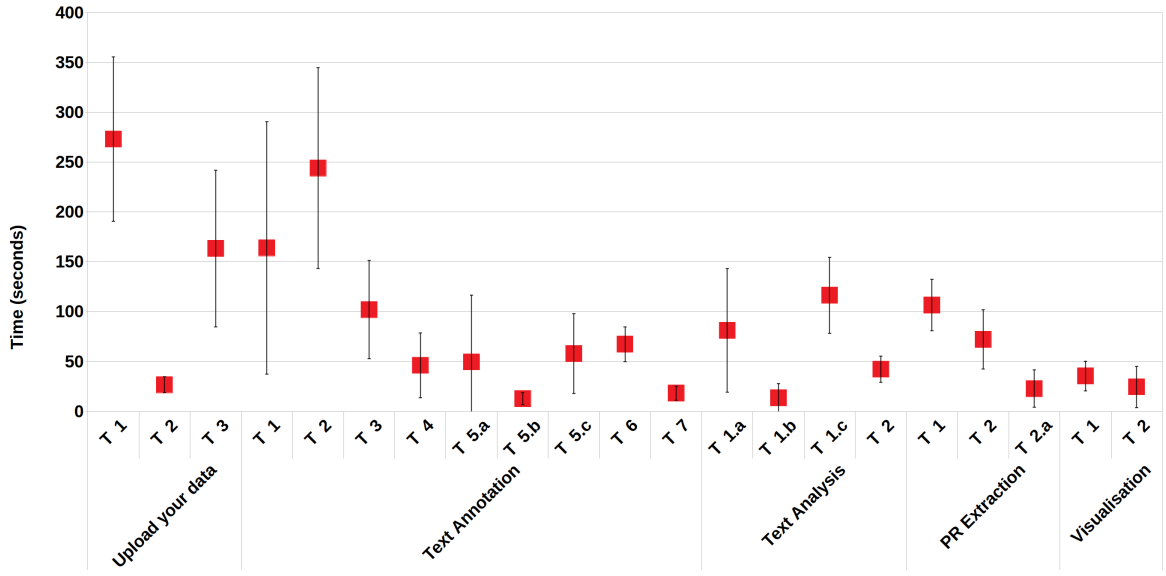



Figure 6.8: Efficiency analysis: time (in seconds) spent on a task, regardless if the user successfully completes a task or if (s)he fails. Error bars represent standard deviation values.

In order to have a better understanding of these results, considering that we have 12 different participants involved in the test, we should take into account the error bars reporting each task standard deviations, both for completion rates and task times. Considering that error bars indicate the variability of the data, it shows clearly from the graphs that participants results are extremely diverse, especially for some of the tasks. For what concerns completion rates,

[7]Note that the maximum value for success rate is 2.

most of the tasks show wide error bars. This fact reveals something that we can't notice from average completions rates, i.e. that we have two main groups in terms of performances: one group obtaining high scores while another one obtaining low scores. This suggests a possible influence of users background, which we investigate below. With respect to task times, wider bars are reported in the first tasks in the test. In other words, regardless if they eventually managed to complete the task, users spent more time on the tasks at the beginning of the test, perhaps to get acquainted with the test.

We investigated the influence of users background on the interaction with the interface by exploring the results obtained by the users divided into two groups based on their background knowledge, as described in the Section 6.6.1.2: experts and non-experts. Graphs reporting the results of each group in terms of completion rates and time taken to complete the tasks are reported in Appendix C. The results seem to provide an affirmative answer to Q4 above: we do observe an impact of participants background on their performances during the test. The results obtained by the expert group are better overall: not only they achieve higher completion rates in all tasks in less time, but their error bars are also narrower. We will further corroborate this evidence by performing analyses aimed at investigating the statistical significance of the group differences.

We concluded our analysis by investigating user satisfaction with PRET after taking the test. Satisfaction measures users' comfort and positive attitude towards the system, and it can be measured through standard psychometric questionnaires. As anticipated, in our test we employed SUS and PSSUQ, both widely used to measure users' perceived satisfaction. Results of the questionnaires are reported, both overall an for each group, in Table 6.1.

| | **SUS** | **PSSUQ** | **SYSUSE** | **INFOQUAL** | **INTERQUAL** |
|---|---|---|---|---|---|
| *Average* | 68.96 | 2.90 | 2.90 | 2.89 | 2.87 |
| *Experts* | 82.92 | 1.98 | 1.83 | 2.03 | 2.14 |
| *Non-Experts* | 55.00 | 3.82 | 3.97 | 3.75 | 3.60 |
| *Reference cut-off* | 68.00 | 2.82 | 2.80 | 3.02 | 2.49 |

Table 6.1: Satisfaction analysis: results of the usability questionnaires computed considering all 12 participants ('Average') and each of the group of experts and non-experts individually. 'Reference cut-off' reports the values widely accepted as cut-offs for good quality when interpreting the questionnaires results (i.e., values should be not lower than the cut-off in SUS and possibly lower than the cut-off for PSSUQ.

When taking a SUS questionnaire, users are asked to rate from 1 (strongly disagree) to 5 (strongly agree) ten statements about the system they just tested. The final SUS score ranges between 0 and 100 points. Although the higher the SUS score the better, 68 is generally accepted as the cut–off for claiming system good quality. As can be noted from the results reported in the table, PRET overall achieves a score of 68.96 ('good', on SUS qualitative scale). However, also this time the two groups show very different results: while the expert group reports a

SUS score reported on the qualitative scale as 'excellent', the non-expert group reported a 'poor' satisfaction level. Comparable results are observed with respect to PSSUQ. Similarly to SUS, PSSUQ requires to rate 16 statements using a 7-point Likert Scale (plus a N/A option), where 1 equals strongly agree and 7 strongly disagree. In general, the lower the score, the better the performance and satisfaction. Unlike SUS, as there are more questions in PSSUQ, it also has 3 sub-scales, namely system usefulness (SYSUSE), information quality (INFOQUAL), and interface quality (INTERQUAL). The sub-scales provide a more detailed breakdown of different factors affecting the system. PSSUQ results are usually evaluated relying on the scale proposed by [243], developed on the basis of the results obtained in 21 different studies. The values proposed by [243] are reported as 'Reference cut-off' in table 6.1. As can be noted from the table, the group of experts shows scores which are systematically lower than the reference score, while, conversely, non experts show consistently higher scores. Again, we confirm the assumption that background affects the use of PRET, as well as the user satisfaction with using the tool. However, we will verify in the future the statistical significance of such systematic differences in other to understand whether our results can be generalised.

#### 6.6.2.1 Discussion

The results reported above, although still preliminary and limited with respect to the complete analysis we are still carrying on, highlight the strengths as well as the limits of PRET tool. Concerning its strengths, in spite of the differences occurring between the experts and non-experts groups, PRET ultimately allowed different populations of users to accomplish the proposed tasks. Familiarity with IT facilities, even more than the experience with annotation practices, seems to have a positive effect on the use of PRET. Possibly, users which are not familiar with the text annotation practice and that aren't comfortable with IT facilities might need more time and supports to get started with the tool. Indeed, the answers to SUS questions specifically addressing learnability show that, while experts don't feel the need of a long training or support, non-experts generally claimed to require longer training sessions before being able to use the tool autonomously. With respect to PRET limits, some modules seem to require a rearrangement of the interface. Specifically, users complained about the functionalities related to the inclusion and removal of concepts from the terminology while annotating: the user interface is reported as non-intuitive when activating these functions. On the other hand, all users claimed to be satisfied with the Quickstart guide, although participants first tried to solve the tasks based on their intuition and they only relied on the guide when in doubt.

Concerning the questions outlined at the beginning of the section, we can now provide the following answers to the questions Q2-Q4 (Q1 is addressed by the annotation project described in the next chapter):

Q2) PRET tool effectively allows to accomplish the annotation and analysis tasks.

Q3) Users are overall satisfied with the tool, although some modules might need an adjustment of the interface, as recommended in the answers to open questions added to SUS and PSSUQ.

Q4) The tool is designed for a wide audience, and indeed users with different backgrounds can achieve the goals of PRET. However, users not experienced with IT tools need more guidance and they require more time to become familiar with the tool.

Although we were able to achieve most of our goals, we are aware that our study design might be affected by some limitations. First of all, as we mentioned already, the study is still ongoing thus the results discussed here report only part of the collected data and our findings should be further supported with additional analyses. We plan to do that in the near future and discuss the results in a dedicated report. Concerning the overall setting, the study might be affected by an *experimenter bias*. Such bias effect can have multiple effects. For example, participants might be influenced, albeit unintentionally, by the person conducting the study, or the experimenter might tend to give better or lower grades out of sympathy. A future study carried out excluding altogether human contact, for example using a crowd-sourcing platform, could help getting rid of this effect. Additionally, it should be mentioned that the test is not designed to evaluate the accuracy of the automatic extraction methods implemented in PRET tool. However, their reported performances are available in the respective cited papers.

## 6.7 Chapter Summary

In this chapter we showed how we put into practice the recommendations and instructions of PREAP protocol into an annotation interface which also supports annotation analysis. Developing an annotation interface was actually an essential part of the work addressing the definition of PREAP and for reaching its current refined version. As we already mentioned, defining an annotation protocol is an iterative process of constant testing, validation and revision. Implementing in PRET the principles of PREAP forced us to evaluate the feasibility of certain ideas and, eventually, make adjustments to meet practical requirements. After all, an annotation protocol that can't be applied in practice is useless, so PRET represents a key element of our work, in particular to perform fast testings. Furthermore, combining PREAP guidelines with an interface to support its application is useful also for what concerns protocol dissemination: PRET makes easier to put into practice PREAP recommendations also for those not familiar with PR annotation or text annotation in general. The result of the usability test carried out with different populations of users seem to corroborate this idea, although they also highlight that users with small or no experience wit IT tools and/or annotation practices might need longer training sessions before reaching a good level of satisfaction and confidence with the tool.

Note that the usability test wasn't meant to evaluate the principles and effectiveness of PREAP protocol for creating PR-annotated datasets. In the next chapter we will report an

annotation project where we applied the principles of PREAP and exploited PRET to obtain a gold dataset annotated with prerequisite relations. For the purposes of such project, we recruited a pool of domain experts as our goal was not only to test the feasibility of PREAP protocol, but mostly to create a novel resource annotated with prerequisite relations.

CHAPTER 7

# Annotation Project for Building a Gold PR-Annotated Dataset

Previous chapters described the novel PREAP annotation protocol for prerequisite annotation on textual educational materials and PRET annotation tool, designed to support the annotation process performed according to the principles of PREAP. As mentioned, both PREAP and PRET underwent multiple revisions and improvements before reaching their current version (v3). In order to validate the annotation protocol and interface, we carried out multiple tests with the purpose of addressing critical issues and improving the quality of the data produced according to our methodology. Among these tests, one annotation project was carried on throughout all three version of PREAP annotation protocol. The project relied on the same text and was aimed at verifying the variations brought by protocol revisions. Each run of the project resulted in a different PR-annotated dataset, which we refer to as Gold-PR v1, v2 and v3. Each dataset version reflects the annotation principles defined by PREAP at different revision phases and, consequently, it was build exploiting the functionalities implemented at different times in the annotation interface to match the needs of each protocol version.

In this chapter we focus on the phase of the annotation project that produced Gold-PR dataset version 3, which was obtained following the principles of the last version of PREAP protocol described in Chapter 5. The project was carried out on PRET tool, in particular we relied on the pre-processing, annotation and analysis modules described in Chapter 6. Before concluding the chapter, we also briefly report in Section 7.6.1 the main differences between PREAP v1, v2 and v3 and which issues were addressed on the basis of the results obtained at different steps of the annotation project.

## 7.1 Project Set-up and Management

As recommended by PREAP protocol, the first step of the annotation process consists of naming a *project manager*. The manager is in charge of supervising the project and of defining the setting and goals of the annotation tasks that will be carried out. Consequently, the manager should be highly confident with protocol recommendations and well aware of the project goal in order to take decision reflecting the desired output. In the present project, the role of the manager is played by a team of five scholars who took part in the definition of PREAP protocol. All five are thus experts in the task of text annotation and of PR annotation in particular. The project manager is in charge of supervising the project in all of its phases. With respect to the preliminary decisions detailed in 5.2.1, the manager set-up the current project as detailed below.

**Annotation Goal.** The project described in the present Chapter is aimed at building a Gold-PR dataset suitable to serve a dual purpose. On the one hand, the dataset should allow the analysis of the textual realisation of PR instances, on the other hand it should be suitable for training and evaluating an automatic PR learning system using features extracted from the text. The latter goal is addressed in chapter 8. The dataset analysis was carried our relying on the analysis and visualisation module implemented in PRET annotation tool (see 6.5) and presented in section 7.5 of this Chapter.

**Textual Corpus Selection and Preparation.** The corpus chosen for carrying out the annotation is an introductory textbook on computer science. In particular, we relied on the chapter about networking[1] According to the book description provided by its authors, this textbook is an introductory book on computer science adopting a bottom-up, concrete-to-abstract explanatory approach. The textbook is meant to cover the needs of computer science students, but also of learners from other disciplines as it is designed to have a broad coverage and a clear exposition. Individual chapters are designed for being independent from one another, allowing us to focus on a single chapter without missing out relevant information for understanding the chapter content. The text underwent a pre-processing step consisting of linguistic analysis performed at morpho-syntactic level by UDPipe pipeline [259] and semi-automatic terminology extraction. The text preparation and the terminology extraction steps are detailed in Section 7.2 of this Chapter.

**Annotation Task Setting.** For the purposes of the current project, annotation was supported by the PRET tool (Prerequisite-Enriched Terminology), the annotation tool described in Chapter 6 that we developed to carry out the process of annotation according to our guidelines and also to analyse the realisation of PR in educational texts.

[1] Glenn Brookshear and Dennis Brylow, 2015. *Computer Science: An Overview*, Global Edition, chapter 4 "Networking and the Internet". Pearson Education Limited.

**Annotators Recruitment and Training.** The project manager recruited four annotators (hereafter, referred to as *Annotator 1, 2, 3* and *4*) among graduated master students in Computer Science. Although they were junior domain experts with respect to the book content, none of them was familiar with annotation procedures or the PREAP annotation protocol. Hence, a preliminary training phase, on two days, was conducted individually for each annotator before starting the actual annotation task, aimed to explain and try the recommendations of the *annotation manual*.

**Annotation Revision.** In-context revision, aimed at checking if pairs created by an annotator comply with the formal and semantic requirements of prerequisite relations, is carried out in this project by each annotator adopting the "*Show only unique pairs*" option. More detail about this are provided in Section 7.3.1.

**Agreement Evaluation and Annotations Combinations.** Once the annotations are done and revised, the project manager is in charge of evaluating and combining them. Annotation evaluation is aimed at quantifying the homogeneity degree between annotations, while annotation combination is the last step of the annotation project, which produces the Gold-PR dataset. Details about agreement evaluation and annotation combination are discussed in Sections 7.4.1 and 7.4.2 respectively.

## 7.2 Text Preparation

As a first step of the annotation project, the raw text of the chapter to be annotated is prepared for the manual annotation phase. Exploiting the functionalities available on PRET tool pre-processing module, the text underwent a linguistic analysis step. Specifically, we exploited the UDPipe [259] pipeline available on PRET. As discussed in 6.2.1, having a linguistically analysed text upon which carrying out the manual annotation is essential here for a dual purpose. On a practical note, the pre-processing step allows to identify domain terms relying on term extraction tools, as we describe below, and to handle different word forms of the same concept used in the text. Next to that, and most importantly, thanks to PREAP annotation approach which requires to create PRs reflecting the actual text content rather than de-contextualised relations, we can collect data encoding the linguistic structures underlying PRs. Such data enable analysis aimed at comparing how different textbook authors present related concepts, but also if there are linguistic cues that could be exploited to automatically find PRs in texts.

The analysis performed on the raw text returned a morpho-syntactically analysed text composed of 20,964 tokens distributed along 780 sentences. The linguistic analysis reflecting the Universal Dependencies formalism is represented in CoNLL-U format, as detailed in the description of PRET pre-processing module. Keep in mind that annotators do not perceive the linguistic analysis: PRET interface displays only the surface representation of the text in order

to make the annotation more feasible for those less familiar with the linguistic annotation formalism.

### 7.2.1 Terminology Extraction

Although the creation of a pre-defined list of relevant terms mentioned in the text to be considered as the text terminology is optional in PREAP, in the present project we extracted a terminology from the textbook chapter adopting a semi-automatic strategy.

As a first step the terminology extraction phase, we extracted a list of candidate terms using Text-To-Knowledge (T2K$^2$) [81], a software platform exploiting NLP, statistical text analysis and Machine Learning to extract domain terms from a linguistically annotated text. Specifically, the multi–word term extraction methodology based on T2K$^2$ hinges on a combination of "termhood" measures, assessing the likelihood of being a valid technical term, and contrastive methods [38]. In particular, term extraction is carried out on the automatically POS–tagged and lemmatised text by searching for candidate domain–specific terms, expressed by either single nominal terms or complex nominal structures with modifiers (typically, adjectival and prepositional modifiers). The latter are retrieved on the basis of a set of POS patterns encoding morpho–syntactic templates for multi–word terms. Domain relevance of single- and multi- word terms corresponding to the patterns is then weighted based on their C-NC value [99]. The ranking of identified terms is refined on the basis of a contrastive score calculated for the same set of terms with respect to corpora testifying general language usage. The main peculiarity of the T2K$^2$ approach lies in the fact that, differently from other strategies, the contrastive analysis is applied to a previously identified terminology with the aim of further filtering it: domain relevant term extraction is based on the concrete frequency of terms occurrence in corpora. Such an approach becomes particularly useful when the domain text collection also includes particularly frequent commonly-used words which make the final result noisy or, more crucially, when the resulting terminology is highly heterogeneous as in the case of educational texts. The final step consists of terminology revision. Three domain experts manually-revised the list of terms extracted using T2K$^2$ in order to remove erroneously extracted non-concepts and add missing ones. Eventually, the revised terminology included 140 concepts. Whenever a concept of the terminology occurs in the text, the instance is annotated in the CoNLL-U file according to the "IOB" tagging scheme, as described in 6.2.2. Again, the annotator can't perceive the term annotation formalism or the term extraction phase, which is performed by the project manager.

## 7.3 Annotating with PREAP Specifications

The manual annotation phase consists of creating individual annotations for each of the four annotators involved in the project. As said, before annotating all annotators underwent a two-days training phase aimed at gaining familiarity and confidence with the PR-annotation task. During

training, first each annotator was introduced by the project manager to PREAP principles and annotation manual. Since the goal and purpose of an annotation depends on the project, we didn't include into PREAP manual many specific indication about this aspect. However, annotators should have a clear vision of the application in order to justify the methodological choices and understand the underlying logic of the annotation [97], so we discussed those with annotators during training. The tutoring phase also included a short session of PR annotation using PRET mentored by the manager. Then, each annotator had one whole day to train alone on a sample text. In the last part of training phase, annotator and manager meet again to address potential questions. The four annotators all reported to feel confident with the annotation interface after practicing with it and that the annotation manual was supporting them adequately in cases of doubts and uncertainty.

After the training phase, each expert performed the textbook chapter annotation individually without consulting the other experts and using the PRET tool annotation module interface as annotation environment. Annotators were given around one week to complete the annotation and return it to the project managers. None of the annotators asked for an extension although they didn't work on the annotation task full-time. Note that for the purposes of the current project the experts were asked to create PR pairs only among the concepts of the previously revised terminology $T$. This choice is aimed at improving the comparability of their annotations, which we then investigated in terms of inter-annotators agreement. Since we framed PR annotation as a concept pairing task on text, the result of the annotation process is a list of concept pairs, where one concept is the prerequisite while the other is the target. Each PR is anchored to the portion of text where the relation was entered: this corresponds to the sentence where the target concept is mentioned. It should be recalled here that annotators are not asked to create negative examples of prerequisite pairs (i.e., non-PR pairs), consequently the annotations don't contain manually created non-PRs: for now, they remain implicit in the annotation.

Table 7.1 in the next sub-section reports, in the 'PAIRS' columns, the number of PR pairs entered by each annotator and the percentage of strong PRs among them. As can be noted, although each expert produced different amounts of pairs, the distribution of weight labels is consistent. The role of annotators in the project could end at this point, as in most annotation projects: they internalised annotation specifications during training and carried out the annotation process according to PREAP principles as requested. An informal post-annotation interview didn't reveal any particular problem encountered by annotators during annotation. However, our experience on PR-annotation and the cases discussed in the literature review (3.2.1) made us aware of a well-known issue of manual annotations: self-inconsistency. The revision approach adopted in this project and the resulting dataset is discussed here below.

### 7.3.1 Annotation Revision

After completing the annotation, experts were asked to perform the in-context self-revision of the annotation as defined in PREAP, which consists of checking the correctness of their own created pairs. As mentioned above, manually produced annotations might contain items that do not reflect the principles of the annotation guidelines. Such erroneous PRs could be introduced either because of annotators distraction of inconsistent application of the guidelines. The pre-annotation training phase was introduced in the last version of PREAP to mitigate the inconsistency effect, at least with respect to inconsistencies due to the misinterpretation of the guidelines: by discussing the annotation manual with the project manager, annotators can address doubts concerning the way in which they should perform annotation. However, even in cases when the guidelines are equally internalised by all annotators, the holistic nature of the PR annotation approach defined by PREAP makes errors due to distraction even more impactful. Performing an annotation aimed at modelling the content of a whole document forces the annotator to always keep in mind previously made choices and already created pairs, thus distraction is very dangerous. In order to remove these cases from the annotations, the project manager set up the revision phase asking each expert to revise only PRs identified solely by her/himself. To do so, the manager used the 'Show only unique pairs' option of the revision interface of PRET. This choices not only maximises the effect of the revision while balancing its cost in terms of time and effort, but it is also the most effective way for removing rare pairs from the annotations.

During revision, PRs could be *i)* deleted, *ii)* confirmed or *iii)* modified. Modifying a PR involves changing its weight, from *strong* to *weak* or vice versa. In order to understand which are the main causes of error, experts were asked to select the error type from a list of 'Error Types' on PRET revision interface when deleting a pair. The following list represent Error Type labels loaded on PRET for the current project and their corresponding descriptions.

a) *Background knowledge*: the PR is not expressed in the text (it comes from the expert's own knowledge);

b) *Too far*: the relation is too weak (in the path, there are too many concepts between the target and prerequisite);

c) *Annotation error*: mistake due to distraction;

d) *Wrong direction*: target and prerequisite concepts should be reversed in the pair;

e) *Co-requisites*: no PR between the two concepts, even though they are both related to another concept.

Those error types were first used in a preliminary study described in [13], where annotation revision was performed on the annotations produced according to the principles of PREAP v2. Note that the first two error types represent cases where the expert wanted to revise what should be considered as a PR, while the three others are proper mistakes. In [13], the inventory of

| | PAIRS | | REVISION | | |
|---|---|---|---|---|---|
| | **PRs** | **Strong** | **Revised** | **Del** | **Mod** |
| **A1** | 141 | 96.45% | 39 (27.7%) | 11 | 4 |
| **A2** | 257 | 84.82% | 85 (33.1%) | 21 | 25 |
| **A3** | 199 | 89.45% | 50 (25.1%) | 15 | 10 |
| **A4** | 163 | 90.18% | 46 (28.2%) | 20 | 9 |

Table 7.1: Annotation and revision summary: for each expert we report the number of created pairs, the proportion of Strong PRs, and the number of pairs that underwent through revision. We also detail the amount of Del[eted] and Mod[ified] pairs.

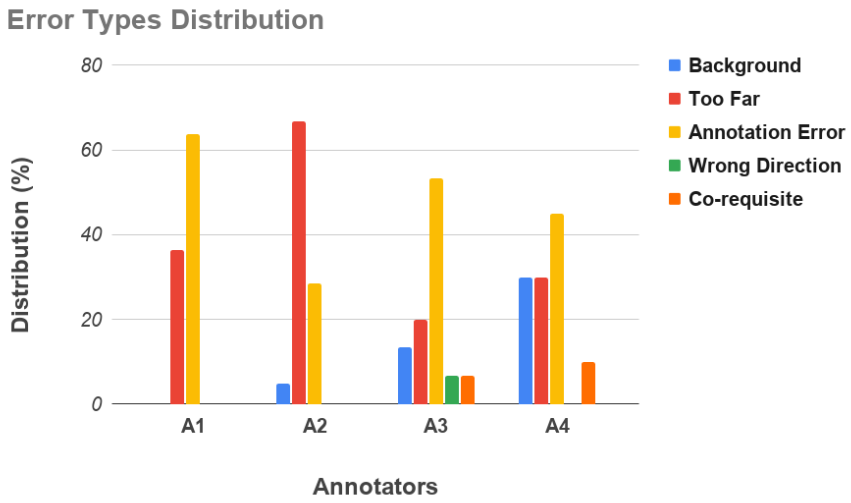



Figure 7.1: Error types identified by annotators for deleted pairs.

labels included also “Not a Concept” error type, meaning that at least one concept of the pair is not a domain term, since experts were allowed to modify the terminology adding new concepts. This error type turned out to be the most recurrent problem for all annotators: common-usage terms like *channel* and *system* were added as domain concepts, but then revised. The problem might arise as an interpretation error commonly observed for educational concepts [280], whose boundaries verge on subjectivity [17]. These results supported the decision of removing the option of adding novel concepts to the terminology in the current project.

Table 7.1 and Figure 7.1 report the summary of annotation and revision. The revised annotation, for each expert, consists of originally created PRs excluding deleted pairs, resulting in 130 pairs for annotator A1, 236 for A2, 184 for A3 and 143 for A4. With respect to the overall number of created PRs, revision involved a comparable amount of pairs among annotators (between 25% and 33%). Considering the number of modified and deleted pairs out of those involved in the revision, we obtain the following percentages of changed PRs: 38,46%, 54,12%, 50,00%, 63,04% for A1–A4 respectively. This means that, on average, more than half of the checked PRs

have been corrected in the revision phase, which shows the importance of this process to obtain reliable datasets. We also notice some interesting commonalities in annotators' revisions. As reported in the graph of Figure 7.1, distraction was identified as the main cause of error for three fourth annotators: generic mistake is consistently the most frequent type, while structural errors, such as wrong direction and creation of co-requisites, are rarer. As discussed in the literature, distraction is inevitable while annotating, thus some kind of revision is a good practice to obtain a more reliable annotation [224, 230]. Only Annotator 2 identified *Too Far* as main cause of error, although this label is quite frequent overall, meaning that the annotation check is useful to remove pairs between concepts having a too weak relationship, thus leaving the final annotations more solid. It should be noted also that the low frequency of *Background Knowledge* label suggests that prior experts' knowledge didn't play a role in the annotation, suggesting that annotators carefully followed the annotation protocol which requests to exploit only the information of the text to create pairs rather than their own knowledge.

## 7.4 Agreement Evaluation and Gold Dataset Creation

### 7.4.1 Inter-Annotator Agreement

Annotations' homogeneity was computed pre- and post-revision using the inter-rater agreement metrics adapted for PR, namely the PRET implementation of Cohen's [64] and Fleiss' $k$ [95] accounting for the transitive and inverse properties of prerequisite relations. We calculated both pairwise Cohen's $k$ and group Fleiss' $k$ because the annotation has been performed by four annotators. As discussed in 4.3.2, the limits of $k$ on PR annotation tasks concern the fact that, in its classical use, $k$ does not allow to account for the overall similarity of the annotations, overcoming individual PR pairs. For this reason, we exploited its variation described in Section 5.2.3 which takes into account not only explicit PRs, manually created during annotation, but also transitive PRs emerging from the exploration of the annotation concept graph.

Instantiating $k$ computation description outlined in 5.2.3, the metric is computed on the annotations produces in the current project as follows. Given the list of 140 concepts of $T$, we create the list $P$ of all possible concept pairs obtained by automatically combining each two distinct concepts of $T$, regardless of the relation direction (i.e. relations *computer*, *network* and *network*,*computer* are both considered). $P$ eventually contains 19,460 concept pairs. For each annotator, we considered PRs of $P$ as positive pairs if they were either manually created by an annotator or if there is a path between the two concepts in the annotator's concept map. Then, we computed Cohen's $k$ pairwise agreement for each pair of experts for both their original and revised annotations. Table 7.2 reports Cohen's $k$ agreement value obtained by each pair of annotators. According to the widely-adopted interpretation of $k$ provided by [150], we observe an average *moderate agreement* (0.60) among the original annotations when considering pairwise agreement, which improves to 0.62 considering the revised annotations. In fact, although small,

| | PRE-REVISION | | | | | POST-REVISION | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **A1** | **A2** | **A3** | **A4** | | **A1** | **A2** | **A3** | **A4** |
| **A1** | 1 | | | | **A1** | 1 | | | |
| **A2** | 0.56 | 1 | | | **A2** | 0.58 | 1 | | |
| **A3** | 0.51 | 0.59 | 1 | | **A3** | 0.54 | 0.60 | 1 | |
| **A4** | 0.59 | 0.71 | 0.65 | 1 | **A4** | 0.63 | 0.73 | 0.66 | 1 |

Table 7.2: Pair-wise agreement, computed in terms of Cohen's $k$, pre- and post-revision.

we observe a consistent improvement among all pairs of experts confirming that the revision allowed to obtain more coherent and homogeneous annotations. For what concerns Fleiss' $k$, the value obtained when considering the original annotations is 0.43, while it raises to 0.45 when computed on revised annotations.

Overall, these results are higher than those observed for other prerequisite annotated datasets in the literature. [56] and [91], for example, report an average pairwise agreement among positive pairs in their annotation tasks of around $k = 0.30$. The agreement problems that arise from the high subjectivity of the task seem to have been mitigated by our annotation protocol. Indeed, also raw agreement [20], computed as a ratio of agreed pairs (i.e., receiving the same label by both annotators) and the overall number of items in the task, confirms the similarity between annotations: having 1 as maximum value indicating perfect agreement, we obtain an average pair-wise raw agreement of 0.97 (0.96 before revision). Note however that this metric should be used carefully: it doesn't reflect a real inter-annotation consistency, but it simply captures a surface similarity, highly influenced by the distribution of labels in the data (see the prevalence problem of $k$). Nevertheless, when raw agreement and $k$ show highly different values, as in this case, it is useful to carry out a qualitative evaluation of the causes of disagreement [62].

#### 7.4.1.1 Causes of Disagreement

As we already observed when analysing the output of past PR annotation tests [13], the cases where we observe disagreement are mostly caused by the order in which concepts are presented. This happens especially when a concept is mentioned for describing another one, for example, to provide details. Consider the case where, e.g., concept $A$ uses concept $B$ to enrich the explanation. Disagreement appears as mainly caused by different understandings of the text portion being red: some annotators are more prone to create the relation $A \prec B$, meaning that you need to know $B$ in order to understand the novel properties of $A$, while other annotators might create the relation $B \prec A$ to indicate that the basic principles of $A$ must be acquired before reading about $B$. It is practically impossible to define an absolute rule to handle such cases, since their interpretation mostly depend on the textual context and the individual processing of the sentences meaning and structure, if the construction is particularly ambiguous. What is interesting from our perspective is that, although such cases are unavoidable when dealing with text written in natural language,

disagreement didn't seem to arise from the interpretation of the annotation approach.

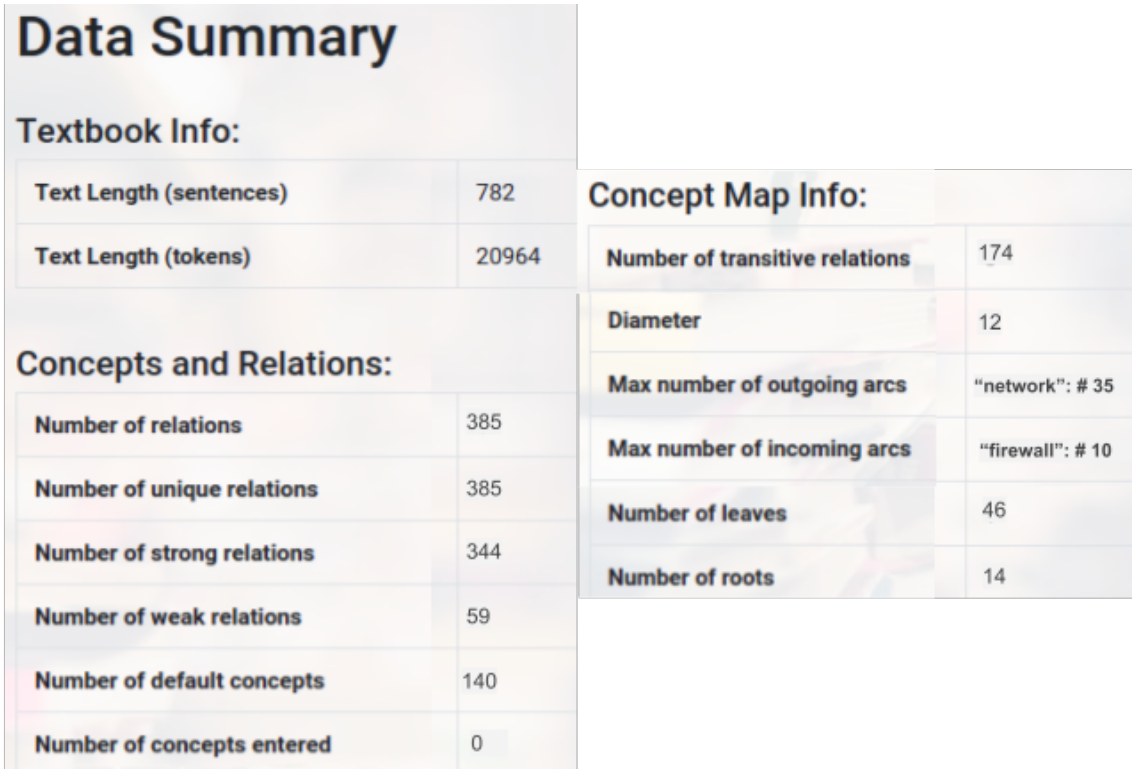
### 7.4.2 Gold-PR Dataset

The combination method of annotations for Gold-PR Dataset creation was chosen by considering the goal of the project and by checking the applicability conditions recommended by PREAP protocol. For the reader's convenience, we remind that the annotation protocol provides a set of recommendations about which combination criterion should be used depending on the agreement obtained by comparing the annotations and the project goal. Indeed, more or less inclusive combination methods, all implemented in PRET tool, can be adopted. In general, inclusive approaches are better suited when the Gold-PR is to be used to analyse, e.g., how concepts involved in PRs are mentioned in the text: more examples to analyse imply a higher variability of cases, thus richer analyses. Conversely, less inclusive combination approaches provide higher certainty and guarantee higher consensus about the relations included in the Gold-PR dataset.

In the current project, the Gold-PR Dataset was built by merging all the four annotations after revision (*Union* option among those implemented in PRET). The 385 pairs annotated as PR by at least one expert appear in the Gold-PR as *positive PRs*, i.e. showing a prerequisite relation. We didn't take into account relation weights in this project.

The Union option fits the goal of the project of creating a Gold-PR suitable for analysis of the textual realisation of PR relation, discussed below, and for training a PR learning system using features extracted from the raw text, as we will show in the next Chapter. The conditions for the applicability of the Union method are also satisfied. In fact, it should be noted that multiple factors contributed in supporting our choice. First of all, the expertise level of annotators made their judgements reliable, and the average agreement computed on the annotations assured a common understanding of the annotation task. Furthermore, the revision step, performed on all manually annotated datasets, provided that as many possible checks were performed to exclude wrong annotations. Annotation revision not only slightly improved the agreement (thus homogeneity), but above all augmented improved correctness and thus reliability. If the project didn't include a revision step, a more strict approach would have been more appropriate regardless of the project goal.

Eventually, the obtained Gold-PR comprises 385 PRs, with 30 pairs annotated by all experts, 47 by three, 89 by two and 219 by one annotator. Note that, for now, the Gold-PR includes only positive pairs, manually created by experts. While the analysis of PR instances could be performed considering only such cases, in order to exploit the Gold-PR dataset for, e.g., training a PR extraction system, a further step aimed at enlarging the dataset with negative examples is needed. We describe the process of dataset augmentation in Section 8.3.1.



Data Summary

Textbook Info:

| | |
|---|---|
| Text Length (sentences) | 782 |
| Text Length (tokens) | 20964 |

Concepts and Relations:

| | |
|---|---|
| Number of relations | 385 |
| Number of unique relations | 385 |
| Number of strong relations | 344 |
| Number of weak relations | 59 |
| Number of default concepts | 140 |
| Number of concepts entered | 0 |

Concept Map Info:

| | |
|---|---|
| Number of transitive relations | 174 |
| Diameter | 12 |
| Max number of outgoing arcs | "network": # 35 |
| Max number of incoming arcs | "firewall": # 10 |
| Number of leaves | 46 |
| Number of roots | 14 |

Figure 7.2: Data Summary, as reported by PRET tool, of the Gold-PR dataset.

#### 7.4.2.1 Dataset Statistics

While agreement metrics are useful to measure how much homogeneity or consensus exists in the annotations produced by multiple annotators, quantitative dataset analysis is aimed at exploring the distribution of specific phenomena within annotations. First of all, we explored the characteristics of the Gold-PR dataset obtained in the present project. The final number of PRs included in the dataset, as said, is 385, with no new concepts entered since the project setting excluded that option. All relations are unique, meaning that we included in the dataset each relation only once (its first insertion). This choice might seem in contrast with our goal of obtaining an annotation aligned with the realisations of PR in texts, but in the reality it is motivated by the applications planned for this dataset. As mentioned, our goal is to have a gold dataset suitable for both conducting investigations over the realisations of PRs in texts and training/testing PR learning models. In this case, using only unique relations enhances data diversity, which is good for training datasets, without loosing information: non-unique relations represent less than 10 PRs in each annotation. Reflecting the distribution of relation weight in individual annotations, 85.35% of PRs are strong also in the Gold.

By analysing the dataset as a concept graph, we are able to acquire information concerning the map structure which reveals also some properties of the final output of the annotation. First, the concept map can be exploited to navigate paths between concepts and identify explicitly

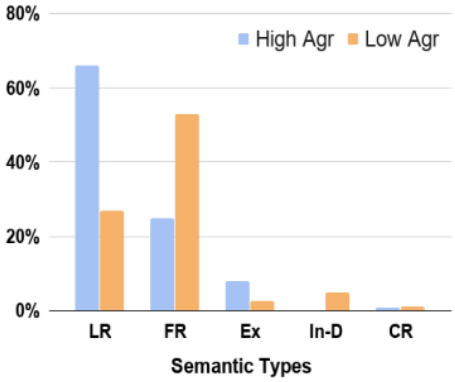
annotated transitive relations. Around 45% (174) PRs in the gold are transitive. This value is much higher than the one observed for individual relations (A1 entered only 3.55% transitive PRs, A2 22.57%, A3 17.59% and A4 14.72%), however it is an expected effect of annotation combination. Map navigation also reveals the graph diameter, i.e. the longest shortest path between any two edges, which could be used to compare the connectivity of multiple golds obtained using different criteria.

What is more interesting from an educational point of view is the information related to edges and nodes. Leaf nodes, for example, correspond to the learning outcomes, i.e. the result of learners' learning process; roots in the map, on the other hand, correspond to the most fundamental concepts (primary notions) presented in the text, which should be either known in advance or that represent the basic concepts of the domain. As reported in the table in Figure 7.2, we observe a high number of leaves as opposed to few roots: this confirms that the text used in this project is an introductory textbook, which assumes little prior knowledge about the domain. Specifically, the node corresponding to the concept *'network'* is the one with the highest number of outgoing arcs (in other words, *'network'* is prerequisite of 35 other concepts), which is not surprising considering the topic of the textbook. On the other side of the spectrum, *'firewall'* is the most "advanced" concept in the dataset as a student is required to know other 10 concepts (including *'network'*) in order to understand *'firewall'*.

## 7.5 Dataset Exploration

As said, manually annotated datasets can serve a variety of purposes, from model training and/or testing as well as theory development. In this Section, we deal with the latter use, i.e. use manual annotations to acquire information about the phenomenon at hand through the analysis of its instances in texts. In fact, by taking advantage of our annotation protocol strategy that allows to obtain an annotation anchored to the text, we can analyse the contexts in which PRs take place and explore their properties.

We first tackled this research direction during a preliminary study, described in [13], aimed at investigating which type of relation is generally established between two concepts showing a PR. The analysis was performed while developing the annotation protocol with the purpose of finding possible improvements of the guidelines with regard to PR definition. To pursue our goal, we relied on the Gold-PR dataset v2, which is the largest of the three gold datasets (see 7.6.1), thus the most diverse with respect to annotators' judgments. The PR properties analysis was carried out during the revision phase of the annotation process: annotators were asked not only to revise their pairs to find errors, but also to classify all confirmed PRs describing which kind of relation could be observed in that segment of text according to their opinions. We provided a pre-defined set of "Confirmation labels" capturing different properties we expected to find among PRs. Labels and their description are reported in the right side of Fig.7.3. By analysing the distribution of




| LABEL | DESCRIPTION |
|---|---|
| *Lexical Relation (LR)* | Hypernyms (is-a), meronyms (part-of), taxonomic relations |
| *Functional Relation (FR)* | One of the two concepts uses the other, or is needed by it. |
| *Example (Ex)* | One concept is explained as an example of the other concept. |
| *In-Depth (In-D)* | The concept is used as in-depth description of another one. |
| *Causal Relation (CR)* | Cause-effect relation between two concepts. |

Figure 7.3: Semantic Relationship Type distribution and description for PRs annotated by 3 or more experts (*High Agr*) or 1 expert only (*Low Agr*).

"Confirmation labels", we investigated if there are any semantic relations and linguistic patterns that can be identified as frequently occurring in PRs, either entered by half or more annotators (High Agreement PR group) or by less than half experts (Low agreement PR group).

As Fig. 7.3 shows, lexical relations are the most frequent semantic type among PRs within the high agreement group, covering more than 60% of cases. This relation type can indeed occur in multiple cases: when the two concepts involved in the PR are described as a taxonomic relation (which, by definition, exploits dependency relations to classify elements), when the text deals with the details of a topic, or also when one of the two concepts is a multi-word term and the other is its lexical head. Functional Relations are more common within the low agreement PR group. Our intuition is that this relation is highly affected by the presentation of concepts in the text: since there is no expressed linguistic cue revealing the PR, the relation between the two concepts is inherently logic and depends on the way in which the author described them. The interpretation of the text content may vary between annotators, and subjectivity has low agreement as a consequence. Consider, as an example, the sentence below, extracted from the textbook annotated in this project (prerequisite concept underlined, target in italic):

> In a bus network based on the Ethernet standards, the right to transmit messages is controlled by the protocol known as *carrier sense, multiple access with collision detection* (CSMA/CD).

The example shows the final agreed PR, with the bus network being the prerequisite of *CSMA/CD*, however one might argue that knowing the principles of the *CSMA/CD* could be helpful to better understand how bus networks work, thus creating the inverse relation.

Other labels were not significantly represented in our dataset. This was surprising in particular for what concerns causal relations (e.g. concept $A$ is caused by, or is an effect of, concept $B$). This could be a peculiarity caused by the domain of the textbook: other domain, such as physics,

history and medicine, could show a larger use of causal relations. Cross-domain investigations could be conducted to confirm similar intuitions in the future.

The results of the above study suggested that providing annotation guidelines with clear examples of PRs not involved in lexical relations is crucial in PR annotation as those types might be harder to identify along a text flow and could give raise to disagreement. This recommendation was included in the last revision of PREAP protocol (specifically, in KEQs) and it turned out to be beneficial, as we discussed for what concerns the inter-annotator agreement. These improvements allowed us to exploit the latest version of the dataset to address two less investigated issues:

1. Are PRs affected by the temporal order of appearance of concepts along the text?

2. Are there any differences in the way fundamental and advanced concepts (representing prerequisite and target concepts of the PR pair respectively) are presented to learners?

By exploiting the annotation analysis module of PRET tool on Gold-PR v3, we investigates the two questions sketched above. In what follows we present the result of our analyses: although results are still preliminary as we exploited a limited number of examples, the evidence acquired is promising and it shows the potentiality of using PRET for deeper PR explorations as well as the effectiveness of PREAP annotation principles for achieving such goal.

### 7.5.1 Temporal Effect on PRs

Some PR learning methods rely on the assumption that PRs are affected by time [2, 257]: explaining a new concept should be done by presenting the fundamental prerequisite concepts first and then including in the discussion the more advanced knowledge (represented by means of target concepts in the PR pairs). We also underlined the similarity between temporal and prerequisite relations in 2.2.3. Our annotation protocol and PR dataset allow us to verify which relationship occurs between time and prerequisite relations. This is possible, in particular, thanks to the text-bound annotation approach which allows to anchor PRs to the portion of text here the expert believed to encounter the relation. To validate the temporal hypothesis on our data, we used the DAG visualisation to observe how many times the prerequisite concept is introduced before (*backward* $\prec$ relations) or after (*forward* $\succ$ relations) the target concept in the text. The ordering of concepts mentions in the text entails many interesting information regarding the textbook and the PR properties. On the one hand, the order in which concepts are presented could reflect the choice of the textbook author to adopt a top-down versus a bottom-up presentation approach: the former tends to explain a topic starting from broad concepts and definitions, while the latter starts from specific cases or examples. On the other hand, this choice could also be guided by the peculiar nature of PRs: intuitively, we expect fundamental concepts to be introduced before the advanced ones, as we generally experience when telling a story (having events in place of concepts) where the background of the events is usually reported first.

The results of the DAG visualisation exploration seem to confirm the above intuition: with respect to relation direction, only 31 (8.05%) PRs are forward relations in the Gold-PR, meaning that the first mention of the prerequisite concept rarely occurs after the mention of the target concept in the text or even after the sentence where the relation was entered. See the sentences below to see two examples of backward relations belonging to the former group.

a) Another early application of the client/server model was used to reduce the cost of *magnetic disk* [target] storage while also removing the need for duplicate copies of records. Here one machine in a network was equipped with a high-capacity mass storage system [prerequisite] (usually a magnetic disk) that contained all of an organization's records.

b) In fact, *hypertext markup language* [target] is the markup language based on the extensible markup language standard [prerequisite] that was developed for representing webpages.

By combining the result of DAG visualisation and Linguistic Analysis (PIC), we can take a step further in the analysis and examine *PR link length*, i.e. the absolute distance in terms of sentences occurring between the mention of the target concept and the closest occurrence of its prerequisite concept. In practice, we say a PR has link distance equal to, e.g., 1 if the prerequisite concept is mentioned either in the preceding or following sentence with respect to the one where the relation was entered. If the target and prerequisite concept co-occur in the same sentence, as in the examples a) and b) above, their link length is equal to 0. On average, our gold dataset shows a PR link length of 7.88 sentences including co-occurring concepts, which becomes 10.16 excluding them. This values are higher than we expected: indeed, we imagined concepts to be generally closely mentioned. To better understand such results, we compared the distribution of six link length intervals in the annotation of the experts, whose results are reported in Figure 7.4.

As we can observe from the bar graph, in each annotation the majority of PRs involve co-occurring, or at least closely mentioned, concepts, as we expected. PRs with a medium link length (ranging from 2 to 4) are rarely created, as well as very long relations (10 or more sentences between the prerequisite and target concept). However, we notice a consistent amount of PRs falling in the 5-10 length interval. This unexpected result is quite interesting: those PRs occur when the two concepts are presented through an example, which formally increases the sentence distance between their mentions, but at the same time it makes more clear the relationship between them. In fact, even when those relations were revised by the expert, they were confirmed as valid PRs in all cases. Interestingly, if we analyse the interplay between time and distance, e.g. by observing concept ordering and link length at once, we notice that forward relations tend to be shorter. Specifically, if we consider the most common concept ordering (i.e., backward: first introduce prerequisite concepts, then present advanced knowledge) the average link length observed on the gold dataset is 10.66, while it lowers to 6.51 for forward PRs. Such result suggests that the textbook author tends to minimise the distance between prerequisite and target concept

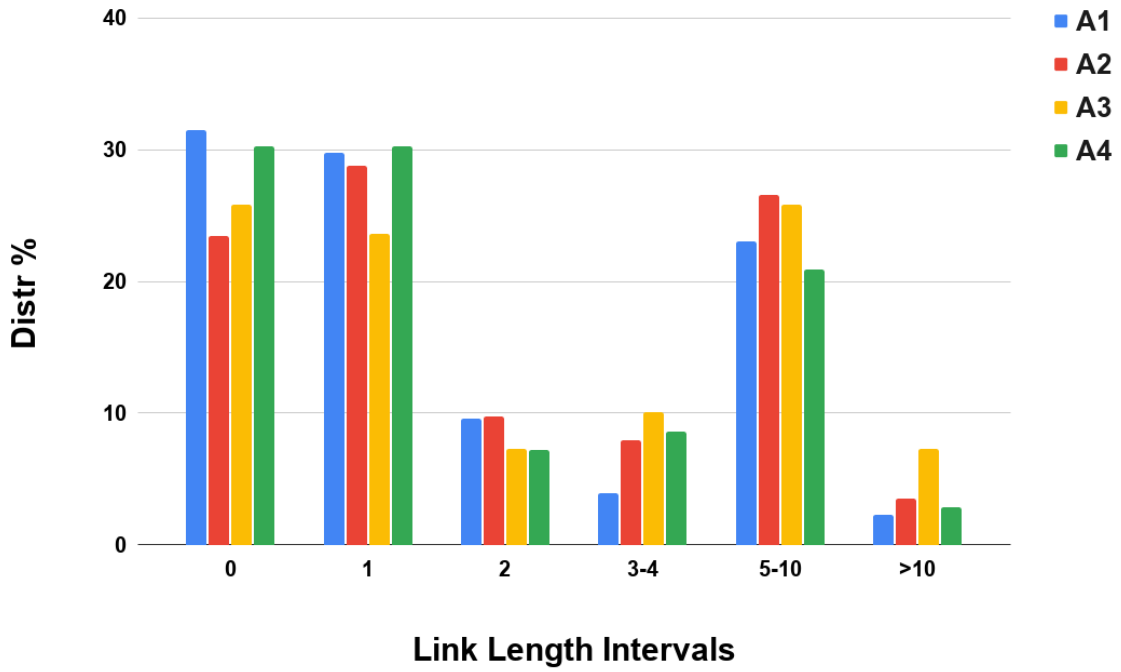



Figure 7.4: Distribution of PR link length intervals for each annotator.

when an alternative (with respect to the dominant, i.e. most frequent one) order is used. This could be a conscious decision of authors and teachers when presenting a topic to avoid confusion and, possibly, reduce the cognitive overload of students. Such **link minimisation effect** should be further explored in larger datasets: given the limited number of forward relations, so far we can only provide a preliminary, although promising, evidence. If confirmed also on larger dataset, this information could be use to provide writing suggestions to authors highlighting particularly complex segments of texts characterised by too difficult relations.

### 7.5.2 Primary Notions and Learning Outcomes Comparison

Our second investigation concerns the way in which fundamental and advanced knowledge is presented in the text. Our intuition, supported by the results of the crowd-based concept ordering task described in 4.2.4, is that the most basic concepts of a domain, representing its fundamental notions and that we refer to as *primary notions* (PNs), should be presented in a simple and clear way, in order to gently introduce the students into the subject. Vice-versa, *learning outcomes* (LOs), i.e./ the end points of the learning process, can be involved in more complex presentations since, at that point, the student should have already acquired all necessary notions to understand them. We explored such idea by investigating whether it is true that text complexity is influenced by the learning level of the concepts described within that text portion. Complexity here refers

both to linguistic complexity, measured in terms of morpho-syntactic properties of the text, and to content complexity, reflecting a higher or lower density of concepts in the text portions.

As a first step, we identify the PNs and LOs in our data. Similarly to [282], we distinguish the concept level (intended as referring to basic as opposed to specific knowledge) exploiting the number of incoming and outgoing links from each node in the concept graph. By relying on the graph analysis provided by PRET tool, we extracted PNs as the top-5 nodes (concepts) having the highest number of outgoing links and no incoming links; conversely, LOs where selected as the top-10 nodes having high incoming links and no outgoing links. Then, we computed concept frequency, both in our textbook corpus and on TenTen English corpus [131], a large corpus of around 22 billion words crawled from the Web used as reference. Its large size makes the TenTen corpus a good enough approximation to observe the distribution of phenomena into the English language. Our idea is verify whether frequency can be used as a reliable feature to distinguish fundamental and advanced concepts. Indeed, word frequency is known to be related to lexical complexity [234], with low frequency terms also generally associate with a higher complexity and vice-versa, thus we expect that fundamental (easier) concepts are characterised by a higher frequency in the language then advanced, thus more complex, concepts.

Table 7.3 reports the absolute frequencies of concepts playing the role of primary notions and of learning outcomes, both in the textbook and in the TenTen English corpus. Results show that **PN concepts are more frequent than LOs in both corpora**, confirming our hypothesis. Indeed, concepts acting as PNs are commonly used single-word domain terms: *'computer'*, *'network'* and *'internet'*, for example, can be mentioned in a variety of contexts, and not exclusively in educational settings or in specialised texts. Conversely, LOs here are mostly multi-word domain terms, referring to highly specific concepts, rarely used outside of the computer science domain.

Next to their frequency, we now want to explore whether there is a difference in the way PNs and LOs are presented in texts. To pursue this goal, we exploit the PIC analysis in PRET tool to extract and compare the contexts where PNs and LOs appear. Our first investigation is aimed at understanding whether the two groups of concepts share similar contexts. The dispersion plot in Figure 7.5 shows how many concepts are mentioned along the text flow, distinguishing three groups of concepts: PNs, LOs and intermediate concepts (i.e. nor PNs or LOs). Specifically, we created intervals of 10 consecutive sentences and computed how many times each group of concepts is mentioned within the interval. The plot reveals that **there is no particularly concept-dense text area**: the distribution of concepts across the text is more or less homogeneous, with few dense portions toward the end, specifically between sentences 500 and 600. The distribution of primary notions is quite predictable: they appear more frequently in the first half of the chapter, when introduced. However, they never really disappear. LOs on the other hand are missing from the very last part of the chapter. This could be easily explained if we take the book content into account: the last part of the textbook chapter deals with network security and closes with an overview about legal aspects related to network communication, thus it doesn't

| | Concept | TenTen Frequency | Textbook Frequency |
|---|---|---|---|
| **PNs** | network | 5,084,100 | 104 |
| | internet | 3,086,197 | 108 |
| | protocol | 1,046,810 | 26 |
| | software | 3,872,136 | 33 |
| | server | 2,256,705 | 46 |
| | computer | 3,990,153 | 75 |
| **LOs** | firewall | 164,951 | 9 |
| | application layer | 8,129 | 16 |
| | webserver | 11,168 | 18 |
| | gateway | 331,978 | 9 |
| | denial of service | 15,317 | 7 |
| | internet protocol address | 1,070 | 10 |
| | port number | 12,402 | 5 |
| | router | 239,121 | 24 |
| | transmission control protocol | 120,735 | 18 |
| | uniform resource locator | 473,575 | 16 |

Table 7.3: Absolute frequencies of primary notions (PNs) and learning outcomes (LOs) both in the TenTen corpus and in the textbook used to build Gold-PR v3.

introduces new computer science concepts but just recalls previously mentioned notions. The distribution of advanced concepts is extremely interesting: they concentrate in certain portions of the text, while being completely absent in others. Such areas roughly correspond to the end of sub-chapters or of paragraphs. This distribution, although intuitive, is extremely valuable since it validates our annotation protocol as a way to reflect the textbook structure in annotations.

As a last step of our analysis, as in 4.2.4, we performed the linguistic profiling analysis using Profiling-UD on the sentences containing mentions of PN and LO concepts. Our goal here is to check whether there are significant differences between the linguistic structures used to present fundamental and advances concepts.

To address our goal, we first collected two groups of sentences, one containing sentences mentioning a PN and one containing sentences mentioning a LO; then, we used the Mann–Whitney U test to check if the feature values computed for all sentences of each group showed significant differences. The full list of features and their values is reported in Appendix B (in B). Significant differences ($p < 0.05$) were observed again for what concerns features capturing the inflectional morphology of verbs and the verbal predicate structure of the sentence, the most predicting with respect to sentence readability [43]. The values obtained by those features on our data suggest that **PNs are presented in generally easier-to-read, thus simpler, sentences and, on the contrary, LOs appear in slightly more complex (difficult-to-read) linguistic constructions**. This observation should be better investigated on larger data, although we are confident that these results would be confirmed considering that they are consistent with those observed on the texts used in the crowd-based experiment (see 4.2.4). Indeed, our scenario is particularly

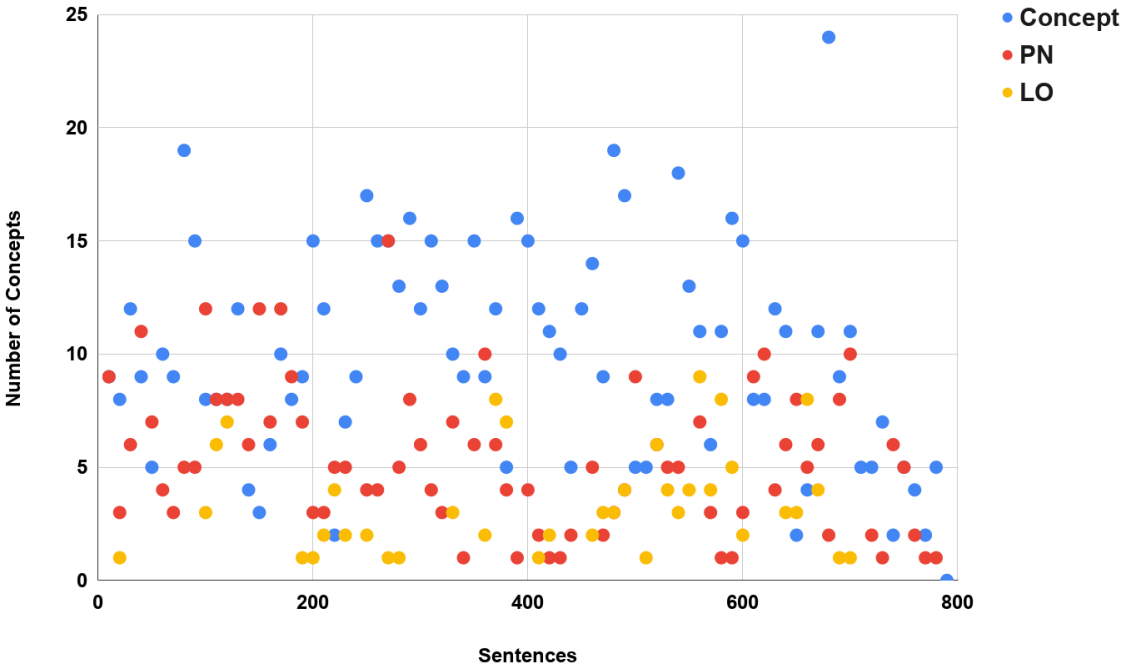



Figure 7.5: Number of primary notion (PN), learning outcomes (LO) and intermediate concepts (Concept) mentioned in different portions of the textbook.

challenging: considering that we are comparing the linguistic features extracted from sentences belonging to the same text, differing only on the concept mentioned, it would have been plausible to expect that differences were concealed by the shared writing style. We expect to observe even stronger differences when comparing description of fundamental versus advanced notions extracted from different instructional materials.

### 7.5.3 Results Summary and Discussion

This Section showed some potential uses of the PR-annotated datasets for studying how PRs are instantiated in texts and explore their properties relying on the tools implemented in the Analysis module of PRET. Regarding our initial research questions, namely *(1)* is there a relationship between time progression and prerequisite structure of a resource, and *(2)* do fundamental and advanced concepts appear in different contexts with respect to text complexity, the analyses were provided insightful evidence which allowed us to draw some conclusions and to gain higher awareness regarding the nature of PRs in educational texts.

With respect to question *(1)*, our analysis showed that temporal sequence of concepts has a strong impact on the relations they establish with each other: prerequisite concepts tend to appear earlier in the text than their target concepts. Even more interestingly, we observed that

the ordering of concepts in the text plays a crucial role in determining the distance (in terms of sentences) that can occur between two related concepts. In particular, we observed that concepts involved in a PR can occur also distant in the text, especially when they are presented by means of an example. However, when concepts are presented in an awkward, non-standard order (i.e. target concept first), their distance tends to minimize. Concerning question *(2)*, we observed that PNs and LOs are actually characterised by different properties and contexts. For example, we noticed that primary notions of the textbook are characterised by a higher frequency both in the annotated text and in every-day language and, conversely, learning outcomes are generally less frequently mentioned. Furthermore, the result of the correlation analysis performed on the feature values extracted using Profiling-Ud show that the context of occurrence of Primary Notions generally corresponds to a less complex text than the context of occurrence of Learning Outcomes.

Although our results should be corroborated with further, possibly multilingual, investigations conducted on multiple educational materials, they provide preliminary evidence about the interplay between textual representations and PRs. These findings, if confirmed, could be integrated in automatic PR learning models to improve their accuracy or, more interestingly, to develop educational materials assessment models able to validate the completeness and clarity of the materials and suggest to writers which text areas could be improved.

## 7.6 Discussion

The experience of defining a new annotation protocol for prerequisite relations, a tool to support dataset creation and the annotation project brought us to tackle relevant issues related to the annotation of educational relations in textual data. Thanks to such experience, we were able to elaborate a set of recommendations and good practices for guiding other researchers in the process of manually annotating PRs to create their own annotated resource. Few of our solutions and tips for applying PREAP to unlabelled data result from common sense, thus they might seem trivial for those familiar with textual annotation tasks. However, part of our intended audience might not have large experience with textual annotation, so it is worth pointing out for them also aspects that might seem obvious. Other recommendations result more directly from our experience and the iterative process of testing and evaluating the different versions of the annotation protocol. Evaluation was carried out relying on the datasets produced at each revision step of the protocol. Dataset versions differ from each other as we used them to experiment with different possibilities of the annotation protocol and make adjustments whenever necessary based on the obtained results.

In the remainder of this section we will outline the most significant differences between each version of the protocol and how they are reflected in the dataset produced accordingly with their principles. On the basis of the observed results, we will provide our general recommendations

and list of good practices for carry our PR annotation according to PREAP principles.

### 7.6.1 Protocol Adjustments and Corresponding Datasets

As mentioned in the previous Chapter, PREAP protocol underwent three cycles of evaluation and revision. An annotation project carried out throughout all protocol revisions on the same text allowed to produce three PR-annotated datasets, which we refer to as PR datasets v1, v2, v3, accordingly with the version of PREAP that was adopted to produce them. Dataset versions are directly comparable as they result from the annotation of the same textbook. The current version of PREAP (v3), as well as the corresponding dataset, were described in Chapters 5 and 7 respectively. Here, we compare the current versions of PREAP and PR dataset with their previous versions, outlining what was improved in each version revision and how the changes impacted the annotation.

It should be noted that modeling the content of an instructional resource, thus creating PR pairs reflecting the content of the textbook, was our goal since the first version of PREAP. Hence, protocol revisions were aimed at refining instructions and projects settings in order to improve the correspondence between the final dataset and the textbook content. Although we eventually experimented with all variations allowed by the protocol, the main differences of each revision, summarised in Fig. 7.6, involve *(i)* the tool used to support the annotation; *(ii)* the selection of domain concepts; *(iii)* the number and expertise of annotators; *(iv)* the revision step and *(v)* the annotation manual.

Our first attempts in PR annotation resulted in PREAP and dataset v1, a first exploratory test with only few formal instructions and no revision of either annotations or domain concepts. The latter were automatically extracted as a list of domain relevant terms exploiting $T^2K$ and used as-it-is. This caused the inclusion of a certain amount of non-concept, while others, regarded as relevant by the experts, were missing. Since PRET tool was introduced in version 2, the first version of PREAP didn't rely on any annotation interface: the annotation process to obtain dataset v1 was carried out on a $n \times n$ matrix of concepts, where $n$ equaled the number of concepts involved in the annotation. A pool of six annotators was asked to enter a binary value in the intersection between two concepts to indicate the presence of a PR without any preliminary training or presentation of the guidelines, which took the form of informal recommendation. At that moment, we didn't consider domain expertise as a necessary requirement for annotators so, among the six annotators, we included four domain experts and two domain novices. Eventually, the domain novices withdrawn from the project because the text was perceived as too difficult, which made us aware that only domain experts could successfully complete the annotation task. We noticed a certain negative impact of the matrix affecting multiple phases of the annotation process. First of all, the matrix was difficult to use for annotators: the quite large matrix structure was likely to cause distraction errors. Second, we weren't able to add any additional attribute to the pair, such as the relation weight that we subsequently included. Furthermore, the matrix

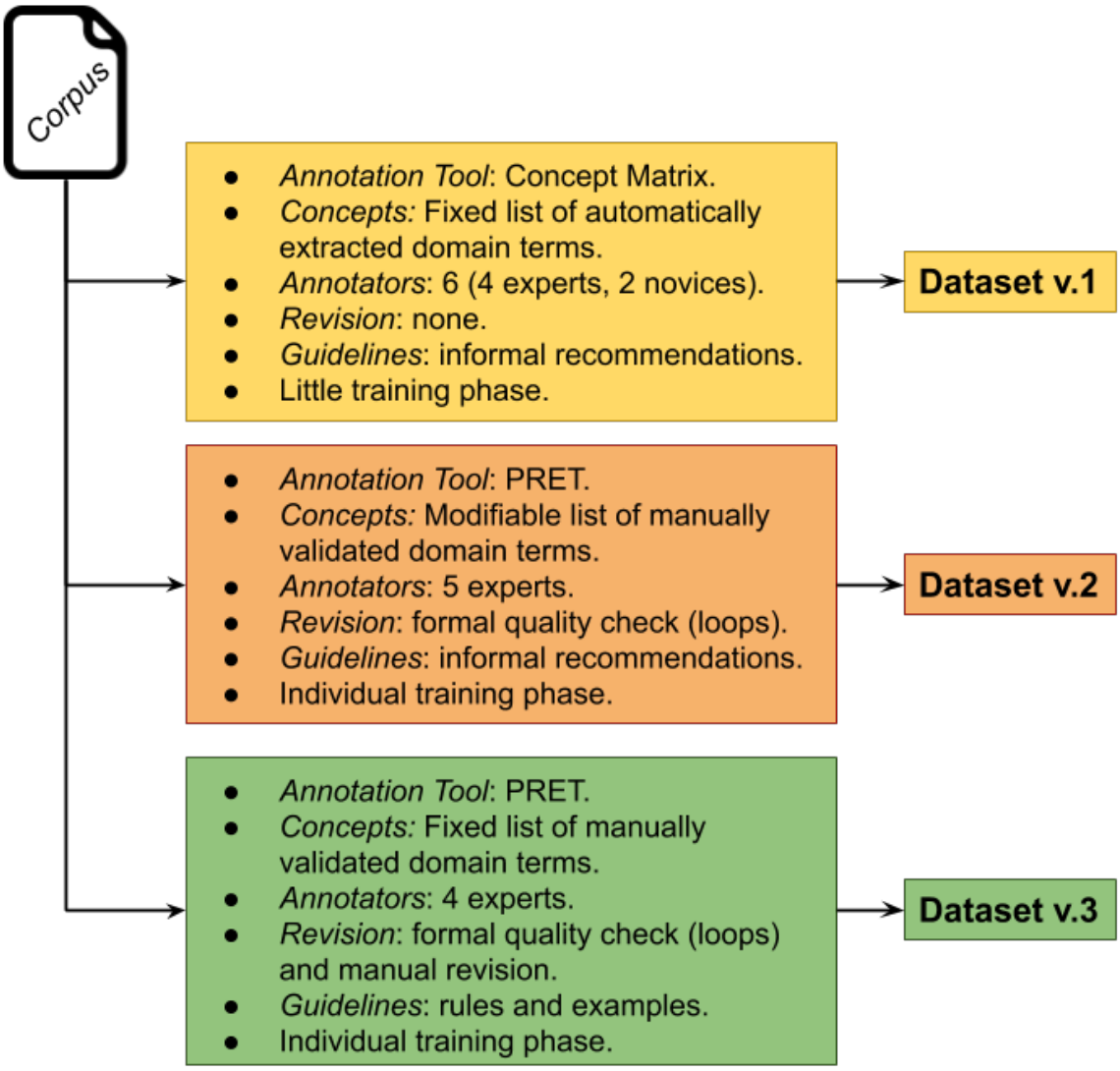



Figure 7.6: Dataset versions comparison.

structure misses the anchoring between PRs and the text: we could always go back to the text and search for every occurrence of the annotated concepts and imagine that the relation was found while reading one of those text portions, however we weren't able to say which one.

To address the issues emerging from our first PR-annotation test, dataset v2 was created by a different pool of five annotators, all domain experts, relying on PRET annotation tool. The tool allowed us to include a formal revision step for each annotation, excluding the insertion, e.g., of involuntary loops while annotating. Although the annotation manual still wasn't final and formalised as annotation documentation, version 2 included an individual training phase for each annotator before approaching the annotation where the project manager explained the basic principles of the annotation to domain experts. Training was aimed to clarify what was

intended for PR annotation to domain experts by means of examples and tests performed by the annotator next to the project manager. In this novel version we experimented with domain concepts: instead of using the automatically extracted list of domain terms as-it-is, we asked three domain experts (not involved in the annotation) to manually validate the automatically extracted (using $T^2K$) list and identify a commonly agreed set of concepts. Besides these terms, each expert could independently add new concepts to the terminology while annotating the text if he/she regards them as relevant. Consequently, experts produced different sets of concept pairs annotated with prerequisite relations.

The annotation approach used to build dataset v3 tried to find a good balance between letting annotators freely express their opinion and converging their annotations toward a common treatment of the PR relation. For this reason we introduced a manual revision step, during which we stimulate annotators to reconsider their pairs at the end of the annotation process, and defined in advance the fixed set of concepts to use during annotation.

### 7.6.2 Good Practices and Recommendations for PR Annotation

During the protocol development process, the different versions of the PR annotated dataset, discussed above, were evaluated both using measures aimed at capturing inter-annotators agreement and using the datasets to train and test the performances of the automatic PR learning model PREL described in the next chapter. The results of the evaluation comparison between dataset versions are reported in Section 8.4.2. Here, we conclude the part of this dissertation dealing with manual annotation by presenting what we learned in form of advice for other researchers that want to apply our protocol in their projects.

What follows is a set of good practices for performing PR annotation to which we came as the result of multiple iterations and testings over PRET tool and PREAP protocol.

**Textual Educational Material Preparation.** We recommend to perform annotation on plain texts (removing, e.g. images and other graphical elements) in order to simplify annotation analysis and dataset creation. Another optional, but highly valuable, pre-processing operation would be to solve acronyms: depending on the domain, acronyms could be highly frequent, but they might make the text more dense and cryptic, thus we suggest to solve them before annotation.

**Annotators Recruiting.** The annotation projects that we carried out to obtain Gold-PR datasets as well as the PRET tool user testings revealed that best results are obtained when domain experts are employed in the annotation. Although PREAP protocol is designed to leverage the knowledge contained in texts, explicitly asking to leave aside any prior knowledge about the domain, domain novices tend to drop the task claiming that their lack of background knowledge in the domain impairs their ability to distinguish PRs from any other type of concept relationship

(even simple co-occurrence) [14]. This might particularly affect projects involving texts belonging to the hard sciences, where the number concepts mentioned in the text is more dense. The background and degree of confidence with IT tools also showed an impact in the use of PRET interface, however even a small practice seems to mitigate this effect making the tool suitable also by humanists with low experience with annotation tools.

**Annotation Training.** A training phase is essential to address unclear aspects of the annotation before proceeding with the ultimate annotation process. On the other hand, clear guidelines with examples and knowledge elicitation questions will support the annotator after the training phase. These aspects are usually neglected in favour of a faster and larger annotation (i.e. on larger texts or performed by more untrained annotators), but they are actually fundamental for annotation projects since they guarantee that the task will be conducted according to the rules defined by the project manager on the basis of the final goal.

**Concepts Selection.** Allowing annotators to express their intuitions as freely as possible, independently adding new concepts to the Terminology while annotating, could result in sparse annotations and low agreement values. Consequently, such approach is discouraged if the final goal consists in obtaining a gold standard dataset, whether it is recommended if the goal is to perform corpora analysis where richer annotations better represent the different realisations of a phenomenon.

**Annotation Revision and Combination.** PRET annotation protocol is designed to create resources valuable for analysing phenomena and/or to train machine learning systems. The tool offers different methods for building a gold standard dataset. As discussed, the best-suited criterion depends on the final use of the dataset. For example, including in the gold dataset only items annotated by most or all annotators might be a good option if your goal is to maximise coherence and you don't plan to carry out a revision step on your annotations. On the other hand, including all PR pairs created by at least one expert might improve dataset coverage, but a revision step is highly recommended in such case, although this practice is costly and not feasible in case of large annotations. The latter approach is a valuable option if the annotation project goal is to explore PRs realisation within educational text, as we did in 7.5. However, we believe that revision is actually extremely effective in improving the overall quality of the annotations. For this reason, in the next chapter, *we will show the same dataset used to carry out the analyses presented in 7.5 can be as well used to perform automatic PR learning using a machine learning approach.*

CHAPTER 8

# Automatic Prerequisite Relation Learning: PREL Approach and Experiments

In this Section we present PREL (Prerequisite RElation Learning), our approach for automatically acquiring prerequisite relations between educational concepts based on the content of educational materials. PREL is based on a machine learning model which exploits raw text features extracted from pieces of text describing concepts for training and no structured information acquired from knowledge structures. As described in Chapter 3.5, automatic prerequisite relation learning consists of identifying prerequisite relations between educational items. Multiple approaches can be adopted, as well as different types of educational resources. PREL model reflects the principles of PREAP annotation protocol as it addresses the task of classifying pairs of concepts as PR or non-PRs relying solely on the information that can be acquired from the educational materials. This is an original approach with respect to existing strategies, and also more challenging as no structured information is provided as feature to the model.

The concept of PREL model was first presented in [193], a work accepted at the 2019 edition of the ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2019). In that occasion, we developed a deep learning system for sequencing Learning Objects based on their prerequisite ordering. Learning Objects were mimicked by pages of Italian Wikipedia adopting the page-as-a-concept paradigm already discussed in 4.2.1 and common to many works addressing concept prerequisite learning (see 3.5). PREL model results from our previous experience on the task of automatic prerequisite relation learning and our considerations regarding the importance of exploiting the information contained in educational materials for acquiring PRs rather then relying on external knowledge bases. As a result, the current version of PREL, presented hereinafter in the chapter, results from the adaptation of the deep learning method introduced above, designed for acquiring prerequisite relation from Wikipedia pages, to

non-encyclopedic textual instructional materials (e.g., textbooks or lecture notes). More precisely, the process of adjusting the model discussed in [193] to the novel scenario primarily consisted in developing an automatic strategy to recreate learning units for each concept mentioned within the material. To this aim, we tested multiple strategies in order to verify which one allows to obtain the best results in terms of system performances. Among the strategies for automatic learning unit creation, we tested and compared a simple strategy based on concept occurrence with a strategy based on Burst analysis.

In this chapter we report and expand the research on PREL by carrying out two experiments. First, we present the experiments described in [15]: we detail the prerequisite relation learning workflow, the variations in the model for what concerns learning unit creation and the results obtained by the models on PRET v2, which was the version of PRET dataset available at the time of model design and development. In the second experiment, we employed the configuration of the model obtaining the best results in the first experiment to test which of the three versions of the PR dataset allows to obtain the best results. This second experiment allowed us to *(i)* study performance variations of the classifier given different input data and *(ii)*, most importantly, compare the impact of different annotation principles on prerequisite learning models.

## 8.1 Prerequisite Relation Learning from Texts

Prerequisite relations learning is proposed here as a task of binary classification of concept pairs. In principle, given a pair of target-prerequisite concepts ($A$,$B$), the goal of the classifier is to predict whether or not concept $B$ is a prerequisite of concept $A$ by relying on the information about the two concepts provided by an educational resource. As a final result, we would obtain a set of concept pairs that can be used to build a knowledge graph where edges represent the automatically retrieved prerequisite relations. Before describing the approach workflow, the models that were employed and the data we relied on, we will discuss the basic principles of our method and the scenario where we tested our strategy.

First of all, we should clarify what we mean in this context by *concept* and *prerequisite relation*. In line with the PR Framework and, in particular, with the principles of PREAP annotation protocol, a concept represents what a student should understand in order to acquire new knowledge and it is represented in educational texts as a lexical entity constituted by a single or multi-word term. Concerning prerequisite relations, they are defined, both in PREAP and in the current work, as dependency relations that naturally occur between educational concepts and that are aimed at determining learning precedence of concepts. PREL method relies exclusively on the content of educational texts for acquiring the features aimed at capturing the presence of a prerequisite relation. Our assumption is that the way concepts are described in the resource should hint at the presence (or absence) of a prerequisite relation between them. For example, if two concepts are frequently mentioned together, intuitively they should be somehow related,

possibly by a prerequisite relation. We use this sort of information acquired from the raw textual content of the resource for training our prerequisite learning model. This approach is in contrast with most strategies for prerequisite learning described in chapter 3.5, which generally rely on external resources, such as ontologies or Wikipedia, but it reflects the annotation principle proposed by the PREAP annotation protocol. As in PREAP, we assign a high informative value to the textual content referring to a concept and we use only that to extract the information for recognising a prerequisite relation. Moreover, since multiple sequences can be legitimately proposed for building up the knowledge referring to the same target concept (as clear from the multiple textbooks concerning the same topic), our approach allows to obtain multiple knowledge graphs for the same domain, each representing knowledge based on the content of the resource used as source of information. According to this principle, we could say that our prerequisite learning task, tackled as a binary classification problem, is at the same time also a textual resource modelling task. However, it should be noted that acquiring information for prerequisite learning relying only on raw text, as in this case, is particularly challenging, as we will discuss shortly.

Considering the above description, our prerequisite relation learning task can be summarised as follows: *given a target and a prerequisite concept mentioned in a text, analyse the content of the resource focusing on the portions of text where each of the two concepts is mentioned in order to identify whether there is a prerequisite relation connecting the prerequisite to the target concept.*

## 8.2 Approach Workflow

As shown in Figure 8.1, the workflow of our approach is organised into different phases. First, given an *educational material* and the *list of concepts* discussed within its content, a *Learning Unit Extraction Module* creates a learning unit for each concept of the list. Then, a set of *features* is acquired from each pair of concepts appearing in the *training set* and passed to the *Classification Module* for training a deep learning classifier. The classifiers returns a set of *pairs binary labelled with prerequisite relations*, which are evaluated against a Gold-PR dataset used as test set in the *Evaluation Module*.

In what follows, we will detail each step of the workflow. In particular, we will describe the different strategies that we tested for creating learning units and detail the features used for training the classification model.

### 8.2.1 Input Data

Our prerequisite relation learning approach PREL requires certain information to be provided beforehand as input data. Input data, encoded by the light blue colour in the workflow in Figure 8.1, are the following:

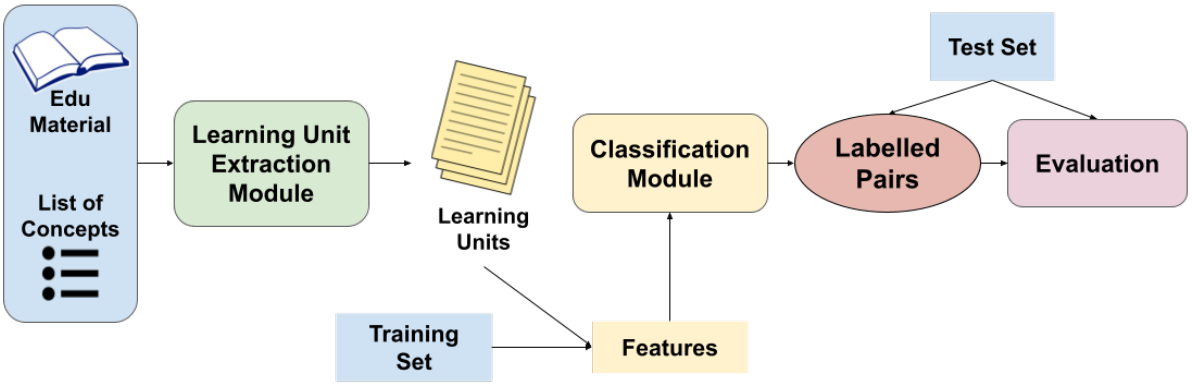



Figure 8.1: PREL Workflow.

- *Educational Material:* the chosen resource (e.g., a textbook) in form of raw text. Ideally, the text should be as clean as possible, meaning that if formulas, figures or acronyms are contained in it, they should be removed or normalised. Although this cleaning step is not necessary, it is recommended in order to obtain better performances (note that it also recommended by our good practices for manual PR annotation).
- *Gold-PR dataset:* a ground truth annotated dataset, created according to the principles of PREAP annotation protocol in order to reflect, in the annotation, the content of the resource as discussed in 5.2.2. The gold dataset, containing concept pairs annotated with labels indicating the presence of a prerequisite relation between the concepts, is expanded with negative and transitive relations as described in Section 8.3.1 in order to obtain a balanced dataset of positive and negative examples. The expanded balanced dataset is then split into a portion of training and a portion of test data.
- *List of concepts:* as for PREAP annotation protocol, a list of the concepts mentioned in the resource is required. In this case, the concepts of the list must correspond to the concepts appearing in the Gold-PR dataset used for training and testing the classification model.

### 8.2.2 Learning Unit Extraction Module

The Learning Unit Extraction Module is aimed at retrieving from the textual resource the portions of text including the learning content associated to each concept appearing in the list of concepts. In other words, since the neural architecture of the classification module was originally designed to receive Learning Objects as input, this module recreates, for each concept, a simulated learning unit by automatically acquiring it from the whole resource. Eventually, the automatically produced learning unit corresponds, for each concept, to a set of sentences extracted from the text. In order to perform the automatic learning unit extraction, we tested multiple criteria, which are detailed below. Note that, in learning design, a 'learning unit' or unit

of learning, is a general term referring to any instructional or learning event of any granularity, e.g. a course, a workshop, a lesson or an informal learning event, that can be instantiated and reused many times for different persons and settings in an online environment [144]. While reading this Chapter, it should be kept in mind that here the term *learning unit* is simply meant as a learning content extracted from resource, with no reference to the notion of learning units in learning design.

In order to verify the impact of different input data on the performances of PREL, we tested different strategies for the creation of learning units. In particular, we defined the following three models:

a) Occurrence Model;

b) Burst Intervals Model;

c) Most Relevant Burst Interval Model.

The **Occurrence Model** is the simpler of the three models: we create a learning unit for each concept by including in the unit all sentences where the concept occurs, namely all sentences where the concept is mentioned. Burst Intervals and Most Relevant Burst Interval models on the other hand are both based on *Burst analysis* [140].

Burst analysis is used to identify the intervals of relevance for an event unfolding over time and it is based on the principle of temporal evolution of phenomena. The general idea behind burst analysis is that, when observing a phenomenon along a time series, the event might become particularly relevant in a certain period of time, most likely because its occurrence rises above a certain threshold [140]. Given its nature, burst analysis was highly employed for detecting the relevant events happened within a time window from news data produced in that period of time [100, 141, 263]. When applied to textual data – e.g., for text clustering [117], summarization [260] or relation extraction [156, 291] – the time series was mimicked by the linear progression of the text. As described in [1], we already employed burst analysis on instructional materials with encouraging results. It was exploited to detect, for each concept discussed within a textbook chapter, the portions of text where the concept was particularly relevant: time was represented by the progressive succession of sentences along the textbook, while the investigated event corresponded to a concept, or better, to its mentions in the text. The burst algorithm (detailed in Section 8.2.2.1) selected the most relevant textual content related to a concept from the textual material and returned one or more *Burst Intervals* for each concept, i.e. intervals of sentences extracted from the textbook. Temporal reasoning [5] between burst intervals was then employed to find prerequisite relations between concepts. The idea behind the approach is that concepts in educational texts may appear with different scopes along the text flow: first they might be just mentioned or introduced, then used inside their definition and later recalled to explain some new information. Therefore, by viewing the textbook as a stream of sentences, one could analyze these changes and better understand how the relation between two concepts evolves in the document.

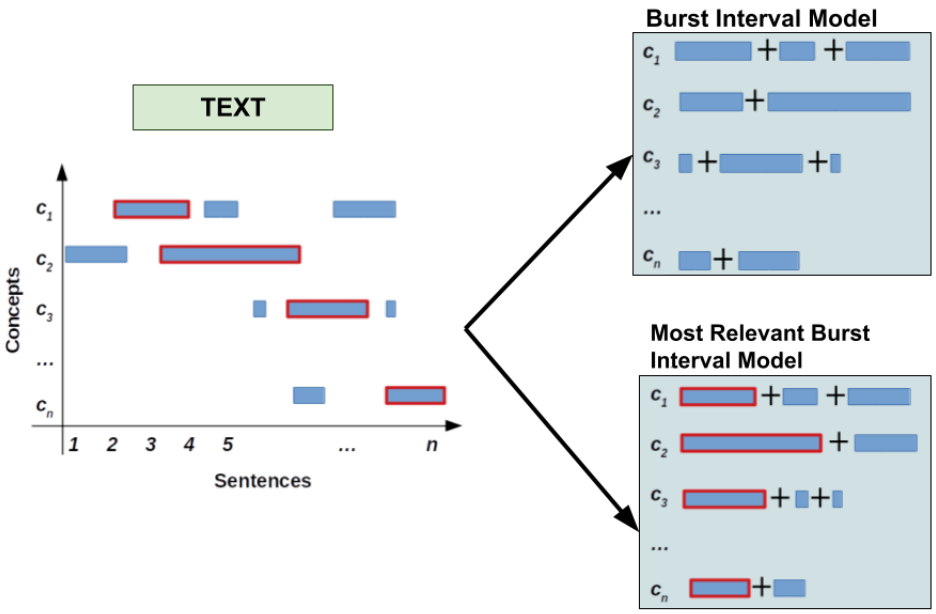



Figure 8.2: Learning Unit creation approaches based on the Burst Interval and Most Relevant Burst Interval Models.

For what concerns PREL, we propose to use the burst intervals retrieved as described in [1] to select relevant content of the educational resource for each concept and obtain the learning units in model *b)* and *c)*. In particular, as represented in the sketchy example depicted in Figure 8.2, the **Burst Intervals model** creates the learning unit of each concept by including its burst intervals (blue blocks in the 'Text' area of the image). Our intuition is that, by exploiting this model for learning unit creation, we should be able to capture not only sentences mentioning the concept, but also surrounding information that, being mentioned close to the target concept, must be somehow related to it. The **Most Relevant Burst Interval model** operates similarly, however in this case burst intervals are reordered, having the most relevant burst interval as first[1]. The most relevant burst interval is defined as the first burst interval that exceeds the average length of all the bursts of that concept, as in [1, 218]. They are depicted as blue blocks with red edges in Figure 8.2. In this case we want to investigate if we can boost the performances of our approach by feeding the model with the most relevant portions of text associated to concepts first. Our intuition is that burst intervals should capture the most informative portions of text for each concept from the whole resource content. Contrary to the Occurrence model, sentences are selected not simply because they contain a mention to the concept being searched, but because the Burst algorithm regarded them as being somehow relevant for the concept. Note that for this experiment we only used the bursts detected with the first phase of the algorithm described in

[1]We also experimented with different placements of the most relevant burst interval within the learning unit, but the solution presented here was empirically evaluated as the most effective solution.

[1]: the temporal reasoning is not employed here.

It should be noted that, in order to capture all concept mentions, all three models require a preliminary step aimed at splitting the text into sentences and normalising the tokens appearing in each sentence to their base form, namely the lemma. This step allows us to account for each concept mention without missing out occurrences that are, e.g., written in plural form.

#### 8.2.2.1 Burst Analysis and Algorithm Implementation

Kleinberg formally defines and models the periods of an event along a time series as a two state automaton in which the event is in the first state if it has a low occurrence, but then it moves to the second state if its occurrence rises above a certain threshold, and eventually it goes back to the first state if its occurrence goes below the threshold [140]. These transitions are repeated along the entire duration of a time series. Burst intervals correspond to the periods in which the event remains in the second state. If applied to a single document rather than a set, Kleinberg's algorithm can be used to detect the bursting intervals of keywords [157, 291], that we can see as roughly corresponding to the concepts of our scenario. Intuitively, a rising of *bursting activity* associated with a concept signals its appearance or re-appearance in the flow of the discourse, revealing that certain features, mainly the frequency of the concept in that interval, are sharply rising [140] and suggesting that the concept has become more prominent.

In principle, the burst extraction algorithm takes as inputs a document $D$ containing the full text to analyze, a terminology $T$ consisting in a list of terms appearing in $D$ and a set of parameters for constructing the Markov's chain according to Kleinberg's description, namely the base $s$ of the exponential distribution used for modeling the event frequencies, the coefficient $\gamma$ for the transition costs between states, and the desired level $l$ within the hierarchy of the extracted intervals. In our scenario,$D$ is the textbook while $T$ corresponds to the list of concepts. First, $D$ is transformed into an ordered list of sentences by means of sentence splitting, and the result is $\mathcal{Q}_D = \{q_1, q_2, ...., q_i\}$, where $q_i$ is the $i$-th sentence of $D$. For each concept in $T$, the burst intervals of $t_u$ are identified by an infinite hidden Markov model among the sentences where concept $t$ occurs as $B_{t_u} = \{[b_{starts_1} - b_{ends_1}], [b_{starts_2} - b_{ends_2}], ..., [b_{starts_i} - b_{ends_i}]\}$. In addition, two parameters, $s$ and $\gamma$, need to be set in advance: the former controls the exponential distribution from which an event is assumed to be drawn (i.e., how frequent an event must be in order to trigger the detection of a burst); the latter modifies the transition cost to a higher state. Higher values of $s$ increase the strictness of the algorithm's criterion for how dramatic an increase of activity has to be in order to be considered as a burst; higher values of $\gamma$ mean that a burst must be sustained over longer periods of time in order to be recognized [37].

For the current experiment, we relied on the same implementation developed for [1]. Specifically, the infinite hidden Markov model was developed as described in [140] relying on an implementation of Kleinberg's algorithm available for Python[2]. Parameters value were set to

[2]Library $pybursts$, https://pypi.org/project/pybursts/0.1.1/

$s = 1.05$, $\gamma = 0.0001$. These represent the minimal parameter values and they were set in this way with the aim of maximizing the extraction of bursting intervals.

### 8.2.3 Classification Module

The Classification Module relies on the neural network developed of *Learning Object* ordering described in [193]. The classifier was part of a novel method based on deep learning applied to the task of automatic prerequisite relations learning between Learning Objects (LOs). When trained on the AL-CPL dataset [165], we report F1=92.21% for the task of sequencing Wikipedia pages. The main peculiarity of the method is that it relied exclusively on linguistic feature extracted from textual resources. Considering only textual content, without relying on structured information, is possibly the most challenging setting to infer relationships between educational concepts, but at the same time it is also the closest condition to a real world scenario: when facing a new topic, students do not cope with ontological relations, but only with those appearing in the learning materials they are using. Thus, one of the most challenging goals of the approach proposed by [193] was to demonstrate that textual content can be sufficient to infer a pedagogically motivated ordering of LO pairs.

As frequently done by work addressing automatic prerequisite learning [16, 103, 168], learning objects were represented in [193] by *Wikipedia page*. In a broad sense, Wikipedia entries can be considered LOs [202]: a Wikipedia page consists of textual content pertaining to a single unit of learning referring to a topic, or concept, represented by the Wikipedia page title. Since the model was designed to find prerequisite relations between Wikipedia pages, the classification algorithm needed to be adapted to a novel scenario: extracting the prerequisite structure of an instructional material. The main differences between the two tasks concern content organisation within the resource. For example, LOs are designed to cover the content of a unique concept, and the same happens in Wikipedia where, being a work of encyclopedic scope, each entry of the encyclopedia discusses a different concept. Multiple concepts might be as well mentioned within the page of a certain concept, if they are somehow related, but the same information will be repeated also in the page dedicated to the content of the other concepts. On the other hand, in instructional materials such as textbooks, lecture notes, slides, etc., the distinction between the content associated to two distinct concepts is not equally clear: the discussion is possibly carried out as a continuum with frequent reference to other portions of text, hence a strategy to extract the content referring to each concept must be defined. To this aim, we developed the strategies for automatically creating learning units described previously (Sec. 8.2.2).

The model proposed in [193] comprises three different neural network models, represented in Figure 8.3, each tested individually in order to investigate which one performs best for the task.

- Model 1 (**M1**) is composed of two identical LSTM-based sub-networks with 32 units, whose outputs are concatenated and classified by the outer Dense Layer. Each sub-network received as input only pre-trained word embeddings (WE) of the first 400 words of the

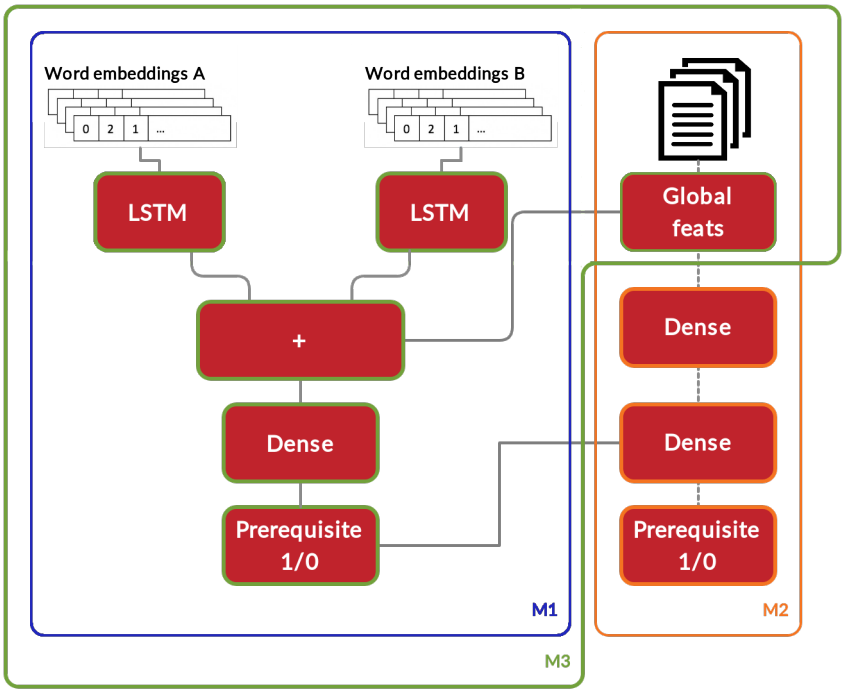



Figure 8.3: Classifier architecture.

corresponding Wikipedia page of each concept appearing in a concept pair of the training set ($A$, $B$). The two LSTM outputs are then concatenated ($V_A \oplus V_B$) and passed to a last Dense Layer.

- Model 2 (**M2**) is based on a feedforward neural network that takes as input the a set of hand-crafted features extracted from the concept pair ($A$,$B$) and passes them to a multilayer perceptron neural network (3 layers with ReLU activation).

- Model 3 (**M3**) combines the previous two, joining the two sub-networks of M1 with the input of M2.

Each output layer of the three models consists of a single dense unit with sigmoid activation function. The models are trained maximising the F-Score on the validation set, which corresponds to the 30% of the training data. The training stops after a certain number of epochs without improvement.

The performances of the three neural networks were tested in [193] on the AL-CPL dataset [168], an English dataset manually annotated with prerequisite relations between educational concepts, and on ITA-PREREQ[3], the first Italian dataset, to the best of our knowledge, annotated

[3]https://sites.google.com/view/prelearn20/data?

with prerequisite relations between pairs of concepts, that we built completely automatically. Considering that results obtained by **M3 model** on both datasets were in line with the state-of-the-art, we decided to further challenge that system and adapt it to the novel scenario of predicting prerequisite relations between concepts extracted from a textual instructional material.

We will now detail the features exploited by the classification module. Considering that *the classification model chosen for our experiment is M3*, we exploited both lexical and global features.

#### 8.2.3.1 Features

For each concept pair appearing in the training set, we extracted two different sets of linguistic features based on those used in [193]: (*i*) *lexical features*, i.e. features that pertain to a single concept, and (*ii*) *global features*, i.e. features referring to the interaction between the two concepts appearing in a pair. Originally, features were meant to refer to the Wikipedia pages corresponding to the LOs of the dataset. In this case and in [15] are acquired from the learning units automatically created for each concept of the dataset. Along the following description, $A$ will denote the content of the learning unit $A$, while $A_t$ will refer to the concept name assigned as title to the learning unit; $(A,B)$ is how we refer to the concept pair.

**Lexical features.** The lexical features correspond to pre-trained word embeddings (WE). Specifically, we relied on a WE lexicon with 128 dimensions built with word2vec [194] generated using the ukWac corpus, a collection of approximately 2 billion words constructed from web pages published under .uk domains [28, 93]. WE can be computed for the entire set of sentences in the learning unit or only for a subset of $n$ sentences, where $n$ is defined at configuration time.

**Global features.** Global feature were devised to extract linguistic information from each two learning units appearing in a pair $(A,B)$ of the training set. Contrary to lexical features, computed considering only $n$ sentences of the learning units, global features take into account the entire content of the units. Specifically, for each pair $(A,B)$, we extracted the following set of 16 text-based features, partially inspired by [166]:

- *In text* (#1, #2): two binary features assuming the value of *1* if $B_t$ ($A_t$) appears in $A$ ($B$), *0* otherwise.
- *Count* (#3, #4): two features whose value corresponds to the number of mentions of $B_t$ ($A_t$) appearing in $A$ ($B$).
- *In first line* (#5, #6): two binary features indicating if $B_t$ ($A_t$) appears in $A$ ($B$)'s first line, which we assume corresponding to $A_t$ ($B_t$) definition.
- *In title* (#7): a binary feature capturing if $B_t$ appears in $A_t$.
- *Length* (#8, #9): two features whose value corresponds to the number of words of $A$ ($B$).

- *Jaccard Similarity* (#10): this feature reports the Jaccard similarity between $A$ and $B$.
- *Jaccard Similarity of Nouns* (#11): the Jaccard similarity between nouns appearing in $A$ and $B$.
- *RefD* (#12): the feature value corresponds to the value obtained by the RefD metric [165] computed between $A$ and $B$. See below for a detailed description of our implementation of the RefD metric.
- *LDA Entropy* (#13, #14): two features corresponding to the Shannon entropy of the LDA vector [80] of $A$ ($B$).
- *LDA Cross Entropy* (#15, #16): two features corresponding to the cross entropy between the LDA vector [80] of $A$ ($B$) and $B(A)$.

We included the RefD metric among the set of linguistic features, although it is usually considered a graph-based feature since, in its original version, it was computed taking into account the incoming and outgoing hyperlinks from Wikipedia Pages [165]. We adapted the RefD metric in order to apply it also to those contexts where no structured information (i.e. hyperlinks) is provided, e.g. in textbooks. In fact, contrary to [165], we computed the metric considering simply the presence of an occurrence of concept $B_t$ ($A_t$) in the page content of $A$ ($B$).
Specifically, we implemented RefD as follows:

$$RefD(A,B) = \frac{\sum_{i=1}^{N} r(t_i,B)\cdot w(t_i,A)}{\sum_{i=1}^{N} w(t_i,A)} - \frac{\sum_{i=1}^{N} r(t_i,A)\cdot w(t_i,B)}{\sum_{i=1}^{N} w(t_i,B)} \tag{8.1}$$

where $t_i$ is a concept from our concept space $T$ (the concepts appearing in the resource); $r(t_i,B)$ is a binary indicator showing whether $t_i$ is mentioned in the content of page $B$; $w(t_i,A)$ is a weight indicator of the importance of $t_i$ to page $A$. [165] proposes two different method for computing the relevance of a concept with respect to a page: *(i) EQUAL*, a binary metric assuming value of *1* is the concept $t_i$ is mentioned in $A$, and *(ii) TFIDF*, where the relevance of a concept is measured in terms of tf-idf if the concept $t_i$ is mentioned in $A$. Regardless of the employed method, if concept $t_i$ is not mentioned in $A$, $w(t_i,A) = 0$.
In our case, we computed $w(t_i,A)$ using the TFIDF method where tf-idf value is computed as follows:

$$w(t,A) = tf(t,A)\cdot \log \frac{N}{df(t)} \tag{8.2}$$

where $tf(t,A)$ is the number of times $t$ being mentioned in $A$; $N$ is the total number of learning units and $df(t)$ is the number of learning units where $t$ appears.

In [193], we conducted an analysis aimed at understanding the relevance of the *global features* on the classifier when sequencing Wikipedia pages. Following [166], we computed the feature

importance by "mean decrease impurity" using an Extra-Trees Classifier, an implementation of a decision tree classifier. The feature analysis was performed for both the Italian and English datasets used for the experiments. The results of the analysis showed that *RefD* (#12), *LDA entropy and cross-entropy* (#13 – #16), *Length of B* (#9) and $B_t$ *in first line of A* (#5) were the most relevant features, regardless of the language of the experiment. In particular, *RefD* seemed the most relevant feature in almost all evaluated settings. Indeed, it was originally proposed as a metric able to identify prerequisite relations based solely on its value, discriminating between prerequisite, non–prerequisite and inverse prerequisite pairs. Although the metric is quite effective also when used alone, our experiments showed that it highly benefits from the combination with linguistic information. With respect to differences, we observed that *Length of A/B* (#8, #9) features were more relevant for Italian than English. Such observation possibly reflects a different structure and content of Italian and English Wikipedia pages that were considered for the experiment, more than a property of the model.

### 8.2.4 Output and Evaluation

The output of the classification module is a set of concept pairs labelled with binary labels indicating the presence or absence of a prerequisite relation. In other words, given an unlabelled pair ($A$,$B$), the classifiers assigns the label *1* to the pair if concept $B$ is identified as being a prerequisite of $A$; if the pair ($A$,$B$) is labelled as *0*, it means that the classifier didn't find any prerequisite relation between concepts $A$and $B$. The set of unlabelled concept pairs of the test set is a portion of the Gold-PR dataset used also for training the model. Labelled pairs are compared against their gold counterparts in order to evaluate the model performances. For evaluation we used standard metrics, such as Accuracy ($A$), Precision ($P$), Recall ($R$) and $F_1$-score ($F_1$).

## 8.3 Dataset

For performing the experiments described below we relied on Gold-PR datasets manually annotated following the principles of PREAP annotation protocol. In particular, we exploited Dataset v1, v2, v3 built by the PR annotation project carried out on the textbook [42]. The creation of Dataset v3 is specifically addressed by Chapter 7. In the present section we describe how the Gold-PR datasets were expanded in order to obtain a balanced set of positive and negative examples.

### 8.3.1 Dataset Augmentation

The Gold-PR dataset manually annotated by experts during the annotation project described in Chapter 7 comprises only positive examples of prerequisite relations, namely only pairs labelled as *1* are available in the dataset. In order to train the classification model based on a neural network, we need to expand the dataset to include also negative examples. Table 8.1 below

compares the datasets in terms of concepts and pairs. The distinct methodologies adopted to build the datasets, guided by the principles of PREAP at different revision stages as described in 7.6.1, produced varying amounts of concepts and annotated pairs. Dataset v2, where annotators where allowed to expand the terminology, is naturally the larger in size, whereas v1 and v3 are quite comparable.

| Dataset | Concepts | Pairs | PRs | Negative PRs | Transitive PRs | Agreement ($k$) |
|---|---|---|---|---|---|---|
| *v1* | 185 | 2,252 | 526 | 1,126 | 600 | 0.40 |
| *v2* | 353 | 6,768 | 1,035 | 3,384 | 2,349 | 0.25 |
| *v3* | 132 | 1,974 | 385 | 1,278 | 311 | 0.60 |

Table 8.1: Number of concepts, pairs, pairs showing a positive, negative and transitive prerequisite relation in each dataset version. Agreement is computed as average pair-wise Cohen's $k$ between all pairs of annotators.

Our dataset augmentation process, performed as a preliminary step of PREL, consists of inferring transitive and negative pairs, already employed for agreement computing, from positive PRs appearing in the gold dataset. *Transitive relations* are added between concepts $A$ and $C$ if the dataset contains $A \prec B$ and $B \prec C$. For example, *computer* $\prec$ *network* and *network* $\prec$ *internet*, then *computer* $\prec$ *internet*. *Negative pairs* are obtained by reversing positive PRs (e.g., add $B \prec A$ as negative if $A \prec B$ is a positive pair), and by adding a random negative example for each automatically created transitive pair. The high reliability of the pair direction observed in 7.5.1 justifies adding negative PRs as inverse PRs. Random negative pairs were added in order to avoid biases when training the classification model: if negative pair corresponded only to reverse positive pairs, the model would learn that a pair is automatically positive if its negative correspondent was encountered in training.

## 8.4 Experiments and Results

The workflow described in the previous section is employed in two different experiments:

- **Experiment 1** is aimed at evaluating the performances obtained by PREL in different configurations, varying with respect to the employed learning unit creation model. We refer to this experiment as 'Model Configuration Analysis'.
- In **Experiment 2** we employed the best performing configuration of experiment 1 to verify whether PREAP annotation protocol revisions, and accordingly dataset variations, contribute to improve the performances of PREL system. We refer to this experiment as 'Impact of Dataset Versions' analysis.

Note that, since we rely on our Gold-PR dataset described in the previous chapter as training and testing data, we will test our approach employing the annotated textbook as learning material.

As a consequence, the experiments will be carried out on English text referring to the computer science domain.

### 8.4.1 Model Configuration Analysis

The model configuration analysis is aimed at investigating which is the best model for automatically creating Learning Units. We evaluate the three models proposed in Section 8.2.2 in terms of performances obtained by the PREL classification workflow. In other words, we test the three models and verify which one resulted in higher accuracy and F_1 scores.

#### 8.4.1.1 Experimental Setting

The model configuration analysis relies on M3 classification model, which uses pre-trained word embeddings (WE) and global features automatically extracted from the dataset, and experiments with variations concerning the Learning Unit Extraction module and WE dimensions.

In particular, concerning Learning Unit Extraction, we test all three approaches described in Section 8.2.2, namely the Occurrence, Burst Interval and First Burst Interval models. Our goal is to verify whether our hypothesis, i.e., burst analysis is more effective in identifying meaningful textbook content related to a concept than simple occurrences, is confirmed. The Occurrence and Burst-based models produced the same number of learning units (corresponding to the number of concepts in the terminology) but with different sizes: burst-based models produced longer learning units, showing an average size in terms of tokens of 534; on the other hand, the Occurrence Model produced smaller LUs, having 250 tokens on average.

Learning units extracted according to each of the three approaches are then used to extract the global and lexical features to train M3. While global features are extracted from the whole learning unit content, lexical features (i.e., WE) are computed only for the first $n$ sentences of the units in the pair. We tried different length of $n$, namely 5, 10, 15 and 30 in order to test if we could reduce computational effort and at the same time provide only meaningful information to the classification model.

The output of the classification is evaluated using a 5-fold cross validation. In practice, we random shuffle pairs contained in the Gold-PR dataset and split the dataset in five portions: the training set corresponds to 4/5 of the Gold-PR dataset, while the test set corresponds to the remaining 1/5. We repeat that five times so that all portions are used as test at least once, then average the results obtained on each portion. Results in terms of F-Score and accuracy are compared against a Zero Rule algorithm baseline whose output is simply the most frequently occurring label in a set of data.

To sum-up, for the Model Configuration experiment, PREL is used in the following configurations:

- **Data**: Gold-PR dataset manually annotated according to the principles of PREAP v2.

| Model | Emb. Dim. | F-Score | Accuracy |
|---|---|---|---|
| Occurrence | 5 | 73.75 | 69.65 |
| | 10 | **74.79** | **70.36** |
| | 15 | 73.7 | 69.19 |
| | 30 | 73.11 | 67.97 |
| | avg | 73.84 | 69.30 |
| Burst Intervals | 5 | 71.75 | 65.54 |
| | 10 | **73.91** | **69.49** |
| | 15 | 72.97 | 67.77 |
| | 30 | 71.37 | 65.06 |
| | avg | 72.5 | 66.96 |
| Most Relevant Burst Interval | 5 | **73.06** | **67.8** |
| | 10 | 72.04 | 66.52 |
| | 15 | 71.58 | 64.43 |
| | 30 | 71.49 | 64.48 |
| | avg | 72.04 | 65.80 |
| Baseline | | 66.66 | 50 |

Table 8.2: Classification F-Score and Accuracy values for the three models with varying number of sentences considered for lexical features. Average and baseline values are also reported.

- **Learning Unit Extraction Module**: all three learning unit extraction approaches, each tested in a different prerequisite learning experiment.
- **Classification Module**: M3 classification model.
- **Features**: lexical features (WE) computed for the first $n$ sentences of each learning unit ($n$ equals to 5, 10, 15 and 30) and global features.
- **Evaluation**: 5-fold cross validation; results in terms of F-Score and accuracy are compared against a Zero Rule algorithm baseline.

#### 8.4.1.2 Results

Table 8.2 reports the results of the experiment. Overall, it appears that PREL obtains satisfying performances in all tested configurations, outperforming the baseline in all cases. Contrary to our expectations, best results are obtained by the Occurrence Model, in particular when the embedding dimension is 10 sentences. In general, computing the WE on 10 sentences or less allows to obtain better performances in all settings. This could be due to the fact that the definition of a concept and its contextualisation with respect to other concepts are generally discussed by the author of the book when the concept is first mentioned in the text. Thus, sentences containing the first occurrences of the term seem to be the most informative for this task. To assess this hypothesis, we manually inspected sentences containing the first mention of each concept. The

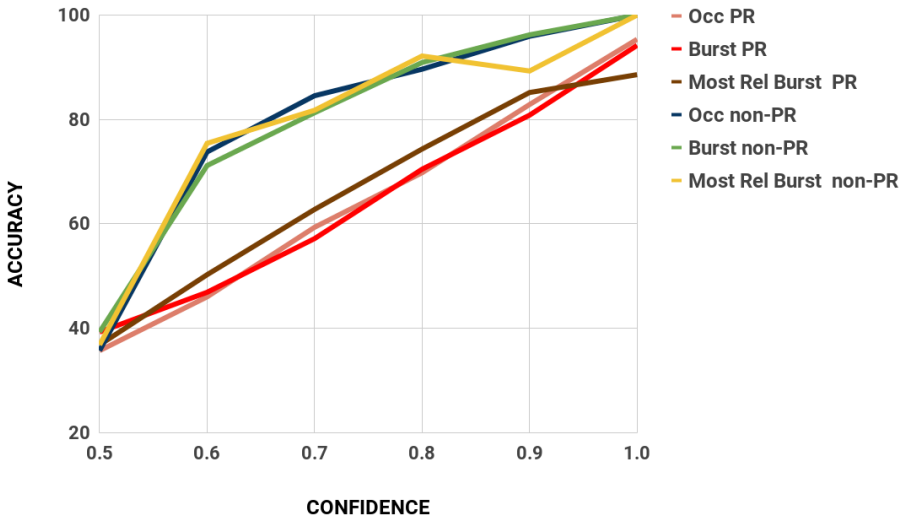



Figure 8.4: Variation of accuracy values with respect to the classifier confidence for pairs labelled as prerequisite (*PR*) and non prerequisite (*non-PR*) in all models considering 10 sentences to compute lexical features.

analysis revealed that 36.3% of the observed sentences contained a concept definition, thus supporting our intuition that the first mention is relevant for concept contextualisation.

The results obtained using the Burst Interval Model are slightly worse, although comparable, probably because, since burst intervals do not necessarily capture all the occurrences of a concept, in some cases the first mentions could be missing from the learning unit. The lowest scores are predictably those obtained using the Most Relevant Burst Interval Model: changing the order of the sentences penalises the system since the temporal order often plays an important role when a prerequisite relation is established between two concepts. Besides, the most relevant burst is not necessarily the first burst interval for that concept and, for this reason, it could contain less relevant information about the concept and its prerequisites. Interestingly, the best results for this model are obtained considering only 5 sentences for computing WE, probably because the system has less chances of observing a lexicon related to other concepts which might bring irrelevant information.

If we look at the variation of accuracy values with respect to the classifier confidence (see Figure 8.4), we observe that our system shows an expected behaviour. In fact, at high confidences correspond high accuracy scores, while at confidence around .5 (12.66% of dataset pairs) we notice that the classifier is more unsure of its decision, obtaining results below the baseline. It should be noted also that the majority of concept pairs (25%) have been classified with a confidence value around .6, while the pairs obtaining the highest confidence value (i.e. equal to 1) are only 1.21%.

The graphs in Figure 8.5 show the variation of confidence and accuracy values with respect to

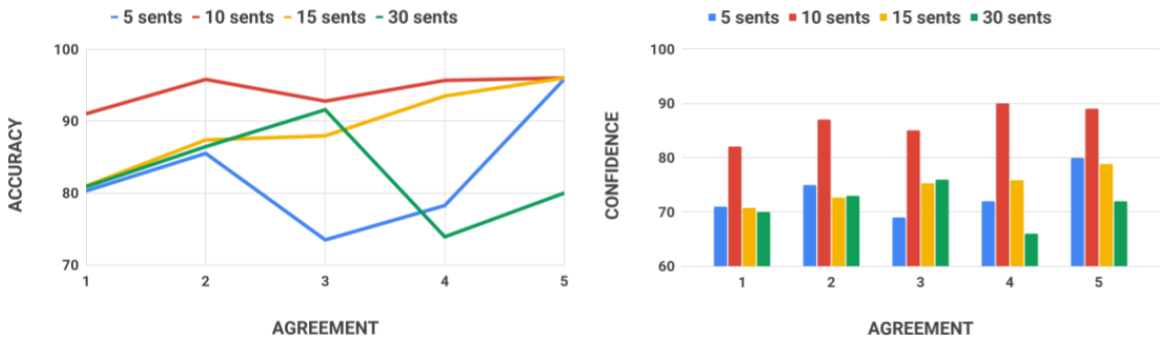



Figure 8.5: Variation of accuracy (on the left) and system confidence (on the right) with respect to the agreement of PR pairs as annotated in the Gold-PR (all possible embeddings length are considered).

the annotators agreement. We report results only for the Occurrence Model since it is the one that obtained the best scores during classification. As we can see, the concept pairs for which all the annotators agree on tend to obtain higher confidence and, consequently, the classifier shows the best performances. The only exception is the model that computes WE using the first 30 sentences, which obtains instead the best scores on the pairs annotated by only 3 experts. The reason for this behaviour will be explored in future work.

#### 8.4.1.3 Error Analysis

This Section compares the results obtained by the three models (i.e. Occurrence, Burst Interval and Most Relevant Burst Interval) when considering 10 sentences for computing WE.

The overall number of pairs assigned with a wrong label by the classifier is quite similar across each setting: 1,835 pairs for the Occurrences model, 1,923 for the Burst Interval model and 2,089 for the Most Relevant Burst model. Moreover, we observe that among these pairs more than 80% were classified as “prerequisite", suggesting that the system overestimates the prerequisite relation, assigning the label also to non–prerequisite pairs. Focusing the analysis on relations that are annotated as prerequisites in the dataset, we observe how their prediction varies across models. 126 pairs were assigned with a wrong “non-prerequisite" label by all models showing similar average confidence values: 0.66, 0.66 and 0.62 for Occurrences, Burst and Most Relevant Burst model respectively. This result suggests that these pairs are particularly complex to classify. Conducting a deeper analysis on this subset, we notice that 85.71% (108) of the pairs are transitive pairs automatically generated (see Section 8.3). Such type of relations seems thus harder to classify than manually annotated ones and might require a different set of features to be recognised considering also that they represent more distant relations. Furthermore, consider that the remaining 18 pairs (14.28%) are manually annotated relations with low agreement

values: 15, 2 and 1 were annotated by one, two and three annotators respectively.

### 8.4.2 Impact of Dataset Versions

Dataset quality can be measured by exploiting agreement metrics, but it can also be tested by verifying whether the dataset can be used to solve a certain task. The Dataset Variation Impact analysis presented in this section is aimed at verifying whether PREAP protocol revisions described in 7.6.1 brought improvements not only in dataset coherence but also to the performances of PREL model. As in the MATTER approach [229], our goal is to verify whether protocol revisions, and accordingly dataset variations, contribute to improve the performance of a system designed to address the task [119]. Furthermore, we also used the results of this analysis while developing the methodology to evaluate the changes brought at different revision steps and make adjustments on the protocol based also on PREL performances. Here, we verify whether the three dataset versions, obtained relying on the principles of PREAP at different revision steps, result in different performances of PREL when employed to train the model. The results presented here below provide interesting insights on how PR properties contribute to improve or worsen the performances of PREL.

#### 8.4.2.1 Experimental Setting

Dataset variation impact is tested here on the three versions of the Gold-PR (v1, v2, and v3). The three datasets were produced throughout the versions of the project described in chapter 7 and they reflect the principles of PREAP at different revision phases. To verify if protocol revisions brought improvements in the model performance, we trained PREL model using the three versions of the Gold-PR. Considering our goal, we used a single configuration of PREL for performing these experiments. In particular, we choose the best performing configuration of previous experiment, namely using the Occurrence model to create learning units and M3 classification model trained with global and lexical features, with the latter computed on the first 10 sentences of each learning unit. As above, the performances of the classifier are evaluated using standard metrics, namely accuracy, precision, recall and F1, computed in a 5-fold cross-validation, and compared against a Zero Rule baseline obtaining 50% accuracy and 66.66% as F1.

To sum-up, for the Dataset Variation experiment, PREL is used in the following configuration:

- **Data**: Gold-PR datasets reflecting the principles of PREAP v1, v2 and v3.
- **Learning Unit Extraction Module**: Occurrence model.
- **Classification Module**: M3 classification model.
- **Features**: lexical features (WE) computed for the first 10 sentences of each learning unit and global features.

| Dataset | Agreement | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|
| *v1* | 0.40 | 83,75% | 87.85% | 85.71% | 85.34% |
| *v2* | 0.25 | 65.08% | 87.94% | 74.79% | 70.36% |
| *v3* | **0.62** | **87.87%** | **89.67%** | **88.73%** | **88.60%** |

Table 8.3: Dataset version comparison in terms of average pairwise agreement (Cohen's $k$) and PREL model performances for each version of the dataset.

- **Evaluation**: 5-fold cross validation; results in terms of F-Score and accuracy are compared against a Zero Rule algorithm baseline.

#### 8.4.2.2 Results

Table 8.3 reports the results obtained by PREL in the current experiment on the three dataset versions. Overall, best performances are observed when the model is trained with Gold-PR v3. Such result suggests a positive effect of the final protocol revision: based on the obtained result, we can assume that Gold-PR v3 shows a higher internal coherence and homogeneity of the annotation with respect to previous versions. Interestingly, when trained on Gold-PR v2 the system looses up to 15 points of accuracy, while Gold-PR v1 obtains lower, but comparable, results with respect to v3. Since the dataset is the only variable that changes in this experiment, we assume that variations in system performances are due to the protocol revisions. By outlining which are the main differences between datasets, we aim to investigate what causes higher or lower results of the PREL system. As a side effect of this analysis, we also want to understand which are the most effective and ineffective changes introduced by each protocol version for what concerns improving PR learning model robustness. Note that, obviously, our observations apply to PREL and prerequisite learning systems based on a similar approach, while systems based on a different architecture or employing a different set of features might be affected by other factors.

The main discrepancies between the three versions concern multiple aspects, as described in 7.6.1. For the readers convenience, here we summarise the most relevant variable elements of the Gold-PRs, concerning concepts, annotators and revision. With respect to *concepts*, dataset v1 employed the domain terms automatically extracted by $T2K^2$ [81], while for dataset v2 and v3 the domain terms were manually refined by experts. Additionally, dataset v2 included also further concepts added by annotators, thus explaining its bigger dimension compared to v1 and v2. Concerning *annotators*, a different pool was recruited for each version: 4 annotators for v1, 5 in v2 and 4 in v3. Although their level of domain expertise was comparable, in v1 and v2 the annotation training phase and guidelines were more limited than in v3, where annotators received the annotation manual and an individual training. Moreover, the *revision* phase was included only in v3 of the protocol.

In light of this observations, we believe that the most impactful changes concern two aspects: *(i)* the introduction of the revision step into the PR annotation protocol and *(ii)* the number of

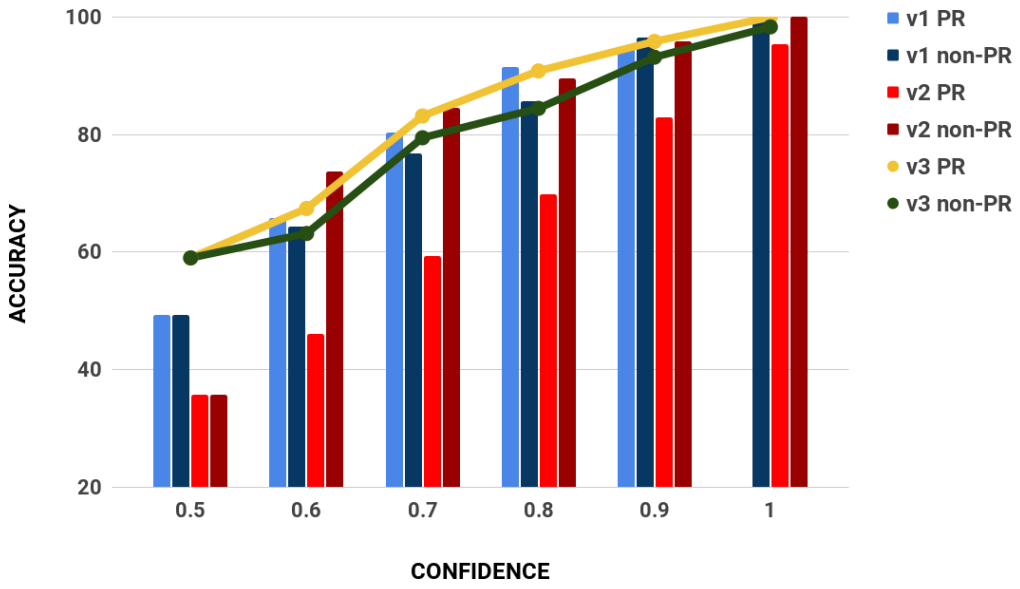



Figure 8.6: Variation of accuracy values with respect to classifier confidence for pairs labelled as prerequisite (*PR*) and non prerequisite (*non-PR*) in all dataset versions.

concepts used during the annotation. In the previous Chapter, when describing the annotation project that brought to the development of Gold-PR v3, we showed how the revision step introduced a small but consistent improvement in the annotations agreement. The results obtained by PREL on dataset v3 suggest that the increased homogeneity also benefited PREL performances, at least with respect to other two datasets, that both resulted from projects where the revision step was missing. However, the results obtained on datasets v1 and v2 offer interesting evidence about the importance of concepts for our model. The lower performances observed when dataset v2 is employed for training show that, although on the one hand adding new concepts during annotation produced a richer set of concepts, it also created a less coherent dataset, as reflected by the lower agreement of v2 ($k$=0.25) not only with respect to v3, but also to v1 ($k$=0.40).

To verify these hypotheses, we examine in more depth PREL performances on all three datasets. As a first step, as we did for the model configuration analysis, we analyse the relationship between system confidence and accuracy. The picture emerging from Figure 8.6 highlights a high similarity between PREL performances when employing datasets v1 and v3 for training. The two only differ when considering pairs where the system is less confident (confidence equal to 0.5): the model trained on v3 is better in finding the correct label for those pairs, while when using v1 the labels are randomly assigned. It should be noted that datasets v1 and v3 show similar results for what concerns the accuracy of the system for both pairs of PR and non-PRs. This is in contrast with what we observed in the previous experiment, when PREL trained on v2 proved more

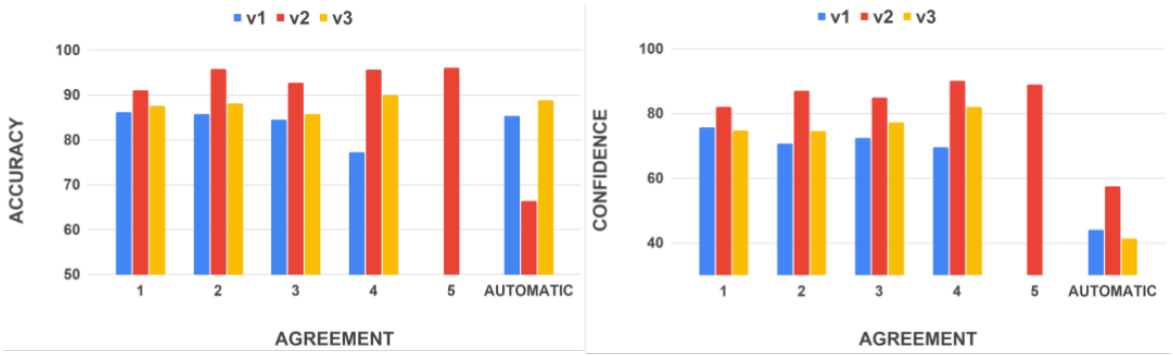



Figure 8.7: Variation of accuracy (on the left) and system confidence (on the right) with respect to the agreement of PR pairs as annotated in each version of the Gold-PR. The 'automatic' columns correspond to pairs automatically generated in the dataset expansion phase.

correct on PR pairs. To better understand the latter result, we observe how models confidence and accuracy vary with respect to pairs agreement. This time, we introduce in the analysis also pairs automatically generated during the dataset expansion phase ('AUTOMATIC' columns in Figure 8.7, including both negative and transitive pairs). Both confidence and accuracy are generally higher when considering PREL trained on dataset v2 (column '5' reports only v2 results since dataset v1 and v3 were build by 4 annotators). Counter-intuitively, when trained on v1, PREL even drops both in accuracy and confidence as agreement raises. However, if we consider the 'AUTOMATIC' column, we notice that v2 causes PREL to be less accurate then v1 and v3 on automatically generated pairs. Indeed, system confidence on automatic pairs is higher: this suggests that PREL trained on v2 tends to identify those pairs as PRs not only when they are transitive pairs, actually showing a positive label, but also when they are negative examples, thus lowering accuracy. PREL trained on v1 and v3 is is visibly more accurate when tested on automatically created pairs.

## 8.5 Discussion and Open Challenges

In this chapter we presented a deep learning model for prerequisite relation learning, PREL. The model was originally designed to be applied on Wikipedia pages, intended as Learning Objects describing each a different concept of a specific domain. PREL is based on neural network which exploits only linguistic features extracted from raw text and does not rely on Wikipedia graph or Learning Object metadata information. Here, we described how the model was adapted to be applied on educational materials, such as the textbook chapter employed in the experiments. In the future the model could be tested also on other types of educational materials, like lecture notes. The results demonstrated the effectiveness of PREL on automatic PR learning, suggesting that it is possible to infer prerequisite relation out of textual educational material without

using any source of structured information. In particular, we validate PREL workflow in two different experiment: in the first experiment we tested different approaches for obtaining the learning units used to acquire the textual features referring to concepts; the second experiment investigated if the revisions of a PR annotation protocol corresponded also to PREL performance improvements.

With respect to the research question addressed by the first experiment, namely how we should select which portions of text better describe a concept, we observed that the best results are obtained when we consider all the sentences where a concept appears. This result suggests that all mentions of a concept are somehow important to define it. Furthermore, the first occurrences of a concept seem to be the most informative with respect to its prerequisite dependency relations with other concepts. This is possibly due to the fact that, when first mentioning a concept, the textbook author need to define its context, thus related concepts, before introducing new knowledge. Indeed, concepts may appear with different scopes along the text flow: first they might be just mentioned or introduced, then recalled later to explain some new information or to explain another concept. For this reason, the order of appearance of concepts in a text should be always considered when training a PR learning model, as it played a crucial role also in our model. Note, however, that Burst analysis turned out effective for uncovering prerequisite relations, as discussed in [1], hence the fact that it didn't improve the results in our model shouldn't be interpreted as a limit of the whole Burst method.

The dataset variation impact experiment was performed to investigate whether PREAP annotation protocol revisions, and accordingly dataset variations, contribute to improve the performances of PREL system. The findings emerging from the results provide insightful information about the improvements brought by protocol revision with respect to PREL robustness. In general, having a smaller but more homogeneous dataset seems more desirable than having a larger but also less coherent set of examples. The possibility to insert new concepts during annotation once again turned out to be a double-edge sword: if on the one hand it seems reasonable to add a missing concept during the annotation, the holistic nature of PRs makes the creation of gold datasets based on such principle more problematic, thus negatively affecting the performances of PR learning systems trained on these data. Further analysis could be aimed at investigating alternative methods to combine annotations in order to create a gold standard which includes both pre-validated and manually inserted concepts. Currently, results suggest that having a pre-defined list of concepts to use during annotation results in a higher ability of the system to distinguish non-PRs from actual prerequisite pairs. By comparing the results obtained by PREL trained on datasets v1 and v3, which are quite similar, we can observe the improvements brought by introducing the revision step into the annotation protocol. Similarly to what we observed for the manual annotation, revising manually created pairs brings small but consistent improvements also to the PR learning system trained with such data. In particular, it seems that revision mostly benefits the system accuracy on pairs having lower confidence, thus possibly also

more ambiguous.

Considering the promising results obtained by PREL, further work should be done in order to test the model in a out-of-domain scenario. In that case, concept pairs used in training should belong to a different domain from the one used during testing. This scenario is still challenging for all PR learning models. Furthermore, in order to identify PRs while taking into account different types of PRs appearing in a dataset, either manually or automatically created, it could be interesting to frame our task as a ranking or multi-classification task rather than a binary classification one.

CHAPTER 9

# CONCLUSIONS AND FUTURE DEVELOPMENTS

## 9.1 Conclusions

This dissertation discussed the issues related to the identification of prerequisite relations in educational materials and proposed a novel methodology to address the task. PR identification on texts indeed presents many challenges, some of them only marginally addressed by existing literature. However, incorporating educational applications with information about the content and prerequisite structure of educational materials might be crucial for the novel generation of educational applications as it would allow to improve their effectiveness in supporting students learning process.

Our proposal for tackling these issues addressed multiple tasks associated to PR identification. First of all, we tackled the problem of manually identifying PR relations and manually annotating them on educational texts. Then, we dealt with the task of exploiting only the information that can be acquired from the content of educational materials for training an automatic PR learning system based on a machine learning models. Both problems are based on this simple, although fundamental, principle: as expressed by the title of this thesis, our goal is to acquire the PR structure from texts without resorting to any external structured knowledge representation or on experts' background knowledge. Addressing the above issues forced us to deal with many related problems and challenges. Our effort towards addressing them brought to a novel methodology for dealing with PRs. The novelty of our methodology is mainly given by the fact that it incorporates multiple perspectives. Specifically, we were guided by principles defined by researches carried out in the fields of Education, Computational Linguistics and Computer Science and Engineering. All these research areas dealt with prerequisite relations in the past, as well as with other more general tasks that we faced such as textual annotation and automatic classification. However

their points of view where never systematically combined as in our work. Our perspective on the task of PR identification, either performed manually or automatically, aims to combine them in order to build a shared vision around prerequisite relations in educational materials and a common ground of discussion for advancing the research on this topic. We systematised the principles of our methodology within the PR Framework, which comprises *PREAP*, a protocol for annotating PRs on texts, *PRET*, an interface for carrying out annotation projects according to the protocol principles, and *PREL*, a model for automatic PR learning relying only on the raw text.

## Thesis Contributions

With respect to the research questions, goals and contributions outlined in Chapter 1, we are now able to discuss them in light of the results presented throughout the thesis.

**Methodology.** Our main contribution is undoubtedly the definition of a novel methodology for dealing with PRs, both for performing manual annotation and automatic extraction from texts. Defining the methodology was our first goal (G1) and it involved setting the boundaries of the research problem and outline a novel task: uncovering PR relations relying exclusively on the content of instructional materials. With respect to the research questions related to G1, we dealt with them in different chapters of this dissertation. Chapter 2 formalised our interpretation of PR relations and educational concepts, providing our answers to RQ1 and RQ2. Specifically, with respect to RQ1, we defined concepts as pieces of knowledge represented in texts by means of domain terms, while prerequisite relations express a propaedeutic dependency relation between them. Concerning the granularity degree of concepts addressed by RQ2, our experience highlighted that it is difficult to define an optimal level of granularity for educational concepts that can be consistently valid for any scenario as it strictly depends on the goals of each project. However, we noticed that using fine-grained concepts (i.e., representing highly specific pieces of knowledge of the domain) should be done carefully as it consistently increases data sparseness.
Chapter 4 discusses the main challenges and the basic principles related to our methodology while addressing RQ3 (i.e., which are the advantages of our methodology as opposed to currently employed ones) and RQ4 (i.e. investigating the relationship between text form and manual PR identification). As widely discussed, uncovering PRs using texts as only source of information, which is the core principle of our methodology, has both practical and theoretical advantages. First of all, researchers relying on our method are not bounded to a specific knowledge representation or domain ontology, making the approach feasible also in cases where such resources are not available. Furthermore, our approach turns the task of PR identification into a text modelling task. RQ4, wondering about the interaction between text complexity and content understanding, allowed us to carry out a study which represents an unicum in the scenario of the research on prerequisite relations, at least to the best our knowledge. In order to better investigate the

issues related to the task of prerequisite identification on texts, we reported the results of the first investigation, as far as we know, aimed at exploring the relationship between the linguistic complexity of texts and the identification of PR relations. Our PR annotation and extraction methodology is grounded in the results of such study, which are discussed in 4.2, that show that humans are influenced by the linguistic complexity of texts when uncovering prerequisite relations between concepts.

**Manual PR Annotation Protocol.** Defining the principles for manual annotation tasks is a common procedure in Computational Linguistics and NLP and it represented the most natural outcome of our second research goal (G2), i.e. defining guidelines for annotating PRs in texts. However, prerequisite relation annotation was only rarely thoroughly addressed, thus the community was missing a shared strategy to represent such information. This thesis provided guidelines and examples for carrying out PR annotation on educational tests. Defining the protocol involved also finding solutions to related problems. We addressed them in our protocol as discussed in Chapter 5, which also provides our answers to the research questions related to G2. For example, concerning RQ1, referring to the ideal textual resource for performing annotation, we eventually concluded that textbooks are one of the most suitable, although somehow challenging, resources. As a matter of fact, almost all textual educational materials would be suited for the task, despite resource-specific issues might be addressed individually. Evaluating the homogeneity and coherence between annotations and defining approaches for combining multiple annotation in order to obtain a gold standard dataset, addressed by RQ2 and RQ3, were thoroughly discussed in 5.2.3. For what concerns RQ2 and agreement computation, we developed our own solution to the problem by proposing an adaptation of standard inter-annotator agreement metrics. With respect to RQ3 and individual annotation combinations for obtaining gold standard datasets, we experimented with different annotation combination approaches and provided tips for future users based on the evidence we acquired. Additionally, we provided recommendations and good practices based on our experience and comparison with the literature for obtaining good quality annotations. We put our efforts towards the definition of practices that could be shared by and adapted to different annotation projects as we aim to promote a shared vision of the PR identification task within the research community.

**PR Annotation Interface.** Since the PR annotation task is designed to be performed by domain experts rather than linguistic researchers, it is important that the annotation environment is easy to set up and use. We summarised this need into the research question RQ4 of G2. In order to provide such an environment for domain experts, this dissertation presents PRET, an annotation interface designed for supporting PR annotation tasks that want to incorporate PREAP principles into their project. PRET is tightly bound to PREAP as its modules and functionalities are aimed at supporting PR annotation according to PREAP recommendations, which are integrated in the interface as discussed in Chapter 6. The interface was also evaluated in terms of

usability with different populations of users. Results confirmed that PRET can be successfully used by a wide range of users, although different backgrounds are associated with different levels of satisfaction.

**Multi-purpose Annotated Resource.** Chapter 7 of this dissertation presented a case study examining how the PR annotation methodology and principles were applied to the annotation of a chapter about networking extracted from a computer science textbook. The resulting annotation underwent multiple evaluations in order to provide the answer to the fifth research question (RQ5) of G2: can we use datasets build relying on PREAP principles to investigate how prerequisite relations are organised within educational texts? We carried out both intrinsic and extrinsic evaluations, and the resource proved its effectiveness for multiple purposes. Concerning the intrinsic evaluation, we computed agreement between annotators (relying on the approach we propose in PREAP) as well as the use of the gold dataset for carrying out investigations concerning the textual realisation of PRs. With respect to the extrinsic evaluation, we exploited the resource for training and testing our deep learning model for automatic PR learning. All tests proved the usefulness and good quality of the resource for achieving our goals.

**Model for Automatic Extraction.** PREL is the last component of the PR Framework and addresses the task of automatically extracting PR relations from educational texts without relying on structured knowledge bases, as defined by our third goal (G3). The experiments described in Chapter 8 show how we can acquire PRs relying only on the content of instructional materials. The set of features we used, fully reported in 8.2.3.1, represent our answer to the first of the challenging research questions related to this goal (RQ1). Some model architecture proved more effective than others, however we were able to achieve above-baseline results in all cases. Developing our model brought us also to tackle the issue outlined by the second research question (RQ2): how to select the most relevant content referring to a concept from the whole resource, namely in our case the textbook chapter. We experimented with different approaches and concluded that a simple naive approach based on concept occurrences allows to obtain good results without putting extra effort into the task. The model was also used to evaluate the iterative approach adopted to design the PREAP annotation protocol. The evaluation was carried out by observing the variation in PREL performances when trained of different version of the dataset. Results confirmed that the current version of PREAP and of the dataset have the highest quality, at least with respect to PREL performances.

**The First Shared Task on Automatic PR Learning and the ITA-PREREQ Dataset.** As a further contribution of our research, it is worth mentioning ***Prerequisite Relation Learning (PRELEARN)***[1] [16], the first shared task on automatic classification of prerequisite relations

[1] `https://sites.google.com/view/prelearn20/home`

between educational concepts. Shared tasks are largely employed by the NLP community to challenge different systems at solving tasks in shared settings and identifying which challenges are still open. Hence, shared tasks constitute a perfect occasion for gathering researchers interested in the task, promote collaborations and sharing ideas. Located in the context of EVALITA 2020 evaluation campaign [29], we organised PRELEARN to call the attention of the research community on the issues of the task of automatic prerequisite relation learning that we outlined in this dissertation. As a matter of fact, the shared task wouldn't have been possible without the research carried out to define the methodology described in this thesis, which allowed us to identify the most critical issues and the open challenges of PR identification. Indeed, PRELEARN challenged participants to develop prerequisite learning systems that exploit either only information acquired from textual educational resources or that can combine those information with structural properties of knowledge structure. In order to satisfy the goals of the shared task, each system had to classify prerequisite relations between pairs of concepts distinguishing between *prerequisite* pairs and *non-prerequisite* pairs. With PRELEARN, we aimed to achieve different purposes: on the one hand, we wanted to offer a setting where the results of different approaches and systems for PR learning could be compared; on the other hand we also wanted to acquire evidence relevant to the wider information extraction and knowledge structure construction communities, as it offered the opportunity to test which information – either textual or extracted from a knowledge structure – are more effective for retrieving pedagogical relations in educational data. The latter is particularly interesting for us, as it might be obvious at this point, since it allowed us to directly compare the effects of adopting the pedagogical view over the ontological view for automatically uncovering PRs. In the occasion of the shared task, we also released for the research community ITA-PREREQ dataset, the first Italian resource annotated with prerequisite relations between pairs of concepts. In order to satisfy the requirements of the shared tasks, we obtained ITA-PREREQ by creating PR relations between pages of the Italian Wikipedia. More details on the creation process of ITA-PREREQ are reported in [193] and [16]. Eventually we obtained a moderate participation involving 3 international teams comprising 9 individual participants. Nevertheless, the models submitted and their results all satisfied the formal requirements for taking part in the shared task and provided interesting insights about PRs that we plan to include in further investigations. The shared task also created the occasion for initiating an international collaboration between our team and one of the participants on prerequisite relations.

## 9.2 Future Improvements

The PR methodology outlined in this thesis provides many different areas in which further research can be done. Indeed, although we were able to achieve most of our goals, the following issues remain open and constitute limits that should be addressed in future work.

**Ontological Representations.** As widely discussed, the novel PR annotation approach, anchored to text, is aimed to capture the instructional design knowledge about concepts organisation and presentation underlying the educational text, i.e., according to the author's view. This can be used to support pedagogical and linguistic analyses, such as comparing teaching approaches for the same subject or discovering if PRs appear within recurrent linguistic patterns. Conversely, the approach is not intended for generalisation of the captured knowledge, being different as such from approaches for ontology learning of teaching methods. This result could be eventually tackled by abstracting common paths from PR knowledge graphs.

**Manual Revision.** The revision step of PREAP is aimed to reduce the risk of errors in Gold-PR datasets and thus to have an impact on annotation agreement and the classification model. Revising all pairs rather than a sub-set can bring larger improvements in the homogeneity and internal coherence of annotations, even though benefits should be balanced with the revision costs. Furthermore, extending the revision by checking also non-included PR relations (e.g., pairs not manually annotated by the annotator, but included in the annotation of another one or resulting from all possible pair-wise concept combinations) would reduce inconsistencies due to "forgotten relations": since they are not taken into account in the revision, we can't tell if two concepts weren't paired because the PR does not exist, or because the annotator did not notice them (so the PR may or may not exist). Considering that the set of non-PRs is exponentially larger than PRs, revising them would be an extremely costly task. However, trade-off approaches could be found to balance benefits and costs, e.g., showing to the expert the relations added by other experts, without revealing how many raters created that pair to avoid biases in the revision. Anyway, note that a certain degree of disagreement is physiological for manual annotation. Indeed, when limited to subjective cases, disagreement can be highly informative and it can be even leveraged to make ML models more robust [221].

**Transitive Relations.** Transitive relations were employed for agreement computation and dataset augmentation for train set creation with promising results. However, the impact of using different thresholds when defining the maximum distance between concepts in a path should be further investigated. Similarly, we would like to test different approaches for obtaining negative relations and their effect in model training. In the experiments descried in Sec.8.4, negative pairs were obtained reversing positive PRs (relying on the asymmetric property), but also randomly pairing concepts. As said, the latter approach is tricky because it implies making assumptions about annotators text interpretation, possibly including inconsistencies. Addressing manual revision as discussed above could mitigate the possible inconsistency effects of automatic non-PRs on classifiers. Alternatively, performing a further manual check of automatically created non-PRs could guarantee higher annotation reliability.

**Agreement Computing.** As mentioned with respect to transitive relations, agreement computing is affected by the strategy we use to acquire non-PRs and implicit relations from the set of manually created PRs. In order to account for the peculiarities of prerequisite relations, we proposed an adaptation of standard metrics which computes agreement considering paths between pairs of concepts as emerging from the final annotation rather than individual concept pairs explicitly annotated. For future revisions of the annotation protocol we might consider experimenting with different penalisation and rewarding strategies. For instance, we could implement an agreement computation approach similar to the one proposed by [121] for lexical chains which, similarly to us, accounts for paths between concepts. Alternatively, we might exploit graph similarity metrics, such as sequence similarity, which compares the similarity between two graphs in terms of shared sequences of vertices and edges.

**Concept Extraction.** Concept extraction was tackled in the annotation project as a terminology extraction task, but other approaches could be tested, possibly accounting for term synonymity and co-reference. This would be particularly useful for texts belonging to different languages or the humanities domain – that we didn't addressed in the current study – given their minor use of technical terms. Testing our methodology on such domains would allow to evaluate the method adaptation to different contexts with respect to the overall annotation and extraction process and, at the opposite, to investigate if optimisations are feasible for specific domains.

**Integration with Educational Applications.** Moving beyond the task of PR identification, we believe that our results could benefit the research in the field of educational technologies in many ways. Not only we offered a newly-defined PR identification task, but also released a resource that could be used by the community for novel PR explorations not possible with existing datasets. For example, applying our annotation protocol on textbooks, as we did in our study, might allow to obtain resources relevant to the recent research field tackling textbook modelling and augmentation which is experiencing a renaissance as testified by two recent workshops on Intelligent Textbook [255, 256], within the Artificial Intelligence in Education (AIED) Conference. This line of research has fostered an interest towards developing electronic materials augmented with services to support their use and enhance their content with additional resources [39], such as interactive exercises [88, 195]. If such services would be embodied with explicit prerequisite structure representations (or, even better, being able to automatically acquire it), they could provide more tailored services to their users by presenting content tarting their specific needs without missing out what is most important in the learning process: the meaningful and coherent organisation of concepts and topics. We also suggested a possible exploitation of PRs in distant learning: a Question/Answering systems for interactive and mobile learning environments [7]. We imagined a scenario where a learner on a MOOC was able to pose questions to the platform through vocal interaction or typed text and the answer is the generation of a personalized learning

path that preserves the precedence of the prerequisite relations and contains all the relevant concepts for answering the user's question.

## Final Words

The work presented in this thesis discussed the issues associated to prerequisite relations identification from different points of view. The research made several contributions to the existing literature and also called the attention on problems frequently neglected by past works. This was possible thanks to our multidisciplinary approach, which offered us the opportunity of posing new questions and developing novel solutions: the different perspectives contributing to the research helped us looking at the task in a new light and we hope that the multi-faceted methodology we propose will be adopted by many other researchers. In particular, we hope that our work will inspire others to investigate open research questions in the area of prerequisite identification, but we also wish it would support those working on solutions addressing the wider scope of supporting learners on educational applications and learning platforms.

This list of possible future research outlined above is not exhaustive, obviously, as many other could be proposed. However, it does show that there is still more to be learned about the application of our methodology, and how it can be used to aid research in a variety of different disciplines and modalities.

# Appendix

APPENDIX A

ANNOTATION MANUAL

# Annotation Guidelines

### I. Concept Identification

1. The goal of the annotation is identifying a prerequisite relation between two distinct terms of a textual corpus. The two terms represent concepts described in the text and can be referred to as target and prerequisite concepts.

2. A concept can be either a single or multi–word term extracted from the corpus.

3. Insert a prerequisite relation for a target concept if you think you need to know the information related to a different concept in order to understand what you are reading about the target concept. Each of the two concepts must be present either in the initial Terminology provided by the project manager or in the manual terminology built by you (i.e. the annotator) during the annotation process. If a concept is still missing in the terminology, add the corresponding term and then insert the relation (if this is allowed by the project manager).

### II. Text Annotation

4. The relation must be inserted exactly in the context (i.e. the sentence) where you find it. A concept could be mentioned more than once along the text, each time introducing novel information and recalling different concept(s). Make sure to add the prerequisite relation between two concepts exactly where the target concept description recalls the knowledge related to the concept you identified as prerequisite.

5. Build a concept pair only if a prerequisite relation does exist between the two: if you think that a relation between two concepts does not occur in the text, do not insert any relation.

6. *Trust the text*: you must annotate only concepts and relations that can be acquired from the text. Do not consider concepts and relations recalled from your background knowledge about the topic.

**III. PR Features and Properties**

7. A concept cannot be a prerequisite of itself: self prerequisites such as "*computer* is a prerequisite of *computer*" will not be allowed by the system.

8. Do not introduce loops in the annotation. Imagine that you have already annotated that: i) "fruit" is a prerequisite of "citrus", and ii) "citrus" is a prerequisite of "orange". By annotating that "orange" is a prerequisite of "fruit", you will create a loop.

9. Every time you insert a relation you must also define its weight. Allowed values comprise: *strong* (the prerequisite is absolutely necessary to understand the other term) and *weak* (the prerequisite is very useful but not strictly necessary).

**IV. Annotation Revision**

10. After completing your annotation, you should also perform an annotation check aimed at revising you previously created PRs. By reading again the portion of text where you entered a relation, decide whether you want to confirm, delete or modify the pair.

11. Delete a prerequisite relation if you added it by mistake. Keep in mind, however, that you can delete one single instance of a pair at a time: if the same pair is annotated with the prerequisite relation in another part of the text, that will be preserved. If you think that ANY prerequisite relation between two concepts should be deleted, you must delete each of the relations having those two concepts.

12. Modify a prerequisite relation if you assigned it the wrong weight. You can modify the weight of the relation if you believe that text expressed a different relation strength than the one you originally assigned to the PR when creating it.

## Knowledge Elicitation Questions

1. Which concepts (among those mentioned in the text) you need to master in order to understand the meaning of the target concept?

2. Which concepts are recalled to define the target concept?

3. Are other concepts mentioned in the same context (e.g., sentence or paragraph) of the target concept? If so, are they useful to understand the meaning of the target concept?
4. Does the target concept represent a special case of another concept mentioned in the text (e.g., *circumference*[target] is a special case of *ellipsis*[prerequisite]?
5. Does the target concept show a part-of relation with another concept mentioned in the text (e.g., the *elbow*[target] is a part of an *arm*[prerequisite])?
6. Does the target concept consists of sub-elements already mentioned in the text (e.g., *elbow*, *forearm* and *shoulder*[prerequisites] are parts of the *arm*[target])?
7. Is the target concept caused by another previously described concept (e.g., *rain*[prerequisite] causes *floods*[target])) or vice versa (e.g., *rain*[target] is caused by *low pressure*[prerequisite])? If so, which one? Try to follow the relation proposed by the text author to understand if a prerequisite relation exists.

APPENDIX B

# PROFILING-UD FEATURES AND ANALYSIS RESULTS

## Crowd-based Experiment Results

### Results of Text Complexity Evaluation

Table B.1: The table reports the results of the textual complexity analysis discussed in Section 4.2.2 of Chapter 4. The table lists the average values of features (grouped by type), computed for the texts extracted from Simple Wikipedia, Wikipedia and Encyclopedias. Values are reported only for those features showing a significant difference (Mann-Whitney U Test p-val<0.05) between the groups. The symbol '–' is used when a feature doesn't vary significantly in that group with respect to the others.

| | | AVERAGE VALUES | | |
|---|---|---|---|---|
| **Property Type** | **Feature** | **Simple Wiki** | **Wikiepdia** | **Encycl.** |
| Deprel Distribution | *dep_dist_acl:relcl* | 0.75 | – | 1.04 |
| | *dep_dist_amod* | 4.81 | 6.87 | 9.12 |
| | *dep_dist_appos* | 0.49 | 0.70 | 1.53 |
| | *dep_dist_aux* | 1.03 | 1.11 | 1.11 |
| | *dep_dist_aux:pass* | 2.01 | – | 2.40 |
| | *dep_dist_case* | 8.41 | 12.57 | 13.40 |
| | *dep_dist_cc* | 2.58 | 3.27 | 3.27 |
| | *dep_dist_compound* | 1.99 | 2.46 | 2.58 |
| | *dep_dist_conj* | 3.31 | 4.64 | 4.64 |
| | *dep_dist_cop* | 4.76 | 3.25 | 3.25 |
| | *dep_dist_det* | 11.78 | 11.92 | 13.60 |
| | *dep_dist_discourse* | – | 0.12 | 0.12 |

| | | | | |
|---|---|---|---|---|
| | *dep_dist_fixed* | – | 0.28 | 0.15 |
| | *dep_dist_goeswith* | – | – | 0.14 |
| | *dep_dist_mark* | – | 2.05 | 0.96 |
| | *dep_dist_nmod* | 5.07 | 7.38 | 8.61 |
| | *dep_dist_nsubj* | 8.12 | 5.29 | 5.29 |
| | *dep_dist_nsubj:pass* | 2.01 | – | 2.18 |
| | *dep_dist_obl* | 2.71 | 4.47 | 4.47 |
| | *dep_dist_parataxis* | 0.40 | 0.56 | 0.56 |
| | *dep_dist_punct* | 14.20 | 13.19 | 12.39 |
| | *dep_dist_root* | 9.73 | 4.82 | 6.01 |
| Ordering of Elements | *obj_post* | 32.77 | 46.33 | 46.33 |
| | *subj_pre* | 91.47 | 97.32 | 83.31 |
| Inflectional Morphology | *aux_form_dist_Fin* | 74.08 | 78.11 | 64.83 |
| | *aux_form_dist_Ger* | – | – | 1.07 |
| | *aux_form_dist_Inf* | 3.73 | 6.78 | 6.78 |
| | *aux_mood_dist_Ind* | 71.19 | 80.51 | 80.51 |
| | *aux_tense_dist_Pres* | – | 76.78 | 62.29 |
| | *verbs_form_dist_Ger* | 7.47 | 9.34 | 13.63 |
| | *verbs_form_dist_Inf* | 7.20 | 12.36 | 12.36 |
| | *verbs_form_dist_Part* | 28.04 | 37.92 | 39.16 |
| | *verbs_mood_dist_Imp* | – | – | 2.54 |
| | *verbs_tense_dist_Past* | 33.47 | 54.86 | 54.86 |
| Parse Tree Struct | *avg_links_len* | 2.31 | 2.76 | 2.76 |
| | *avg_max_depth* | 3.35 | 4.50 | 4.66 |
| | *avg_max_links_len* | 6.85 | 12.81 | 12.81 |
| | *avg_prepositional_chain_len* | 0.58 | 0.99 | 1.13 |
| | *avg_subordinate_chain_len* | 0.61 | 0.78 | 0.78 |
| | *avg_token_per_clause* | 8.27 | 11.85 | 11.85 |
| | *max_links_len* | 6.85 | 12.81 | 12.81 |
| | *n_prepositional_chains* | 0.67 | 1.53 | 1.63 |
| | *prep_dist_1* | 40.80 | 66.10 | 69.21 |
| | *prep_dist_2* | 6.00 | 9.39 | 13.27 |
| | *prep_dist_3* | 0.85 | 3.32 | 3.32 |
| | *prep_dist_4* | – | – | 1.41 |
| | *principal_proposition_dist* | 61.03 | 56.70 | 43.88 |
| UPOS Distribution | *upos_dist_ADJ* | 6.53 | 7.76 | 10.58 |
| | *upos_dist_ADP* | 8.21 | 12.12 | 13.34 |
| | *upos_dist_AUX* | 7.83 | 6.31 | 6.31 |
| | *upos_dist_CCONJ* | 2.64 | 3.26 | 3.26 |
| | *upos_dist_DET* | 12.02 | 12.06 | 13.68 |
| | *upos_dist_NUM* | 1.48 | – | 1.97 |
| | *upos_dist_PART* | 1.48 | 1.58 | 0.44 |
| | *upos_dist_PUNCT* | 15.80 | 13.11 | 12.27 |

| | | | | |
|---|---|---|---|---|
| | *upos_dist_SYM* | 0.05 | 0.42 | 0.42 |
| | *upos_dist_X* | – | 0.22 | 0.34 |
| Raw Text Properties | *char_per_tok* | 4.55 | 4.83 | 5.02 |
| | *n_tokens* | 15.14 | 26.71 | 26.71 |
| | *tokens_per_sent* | 15.14 | 26.71 | 26.71 |
| Subordinate Structure | *subordinate_dist_1* | 41.53 | 57.84 | 57.84 |
| | *subordinate_post* | 42.51 | 56.75 | 56.75 |
| | *subordinate_pre* | 7.87 | – | 15.00 |
| | *subordinate_proposition_dist* | 30.97 | 41.61 | 42.52 |
| Verbal Predicate Struct | *avg_verb_edges* | 1.80 | 2.15 | 2.23 |
| | *verb_edges_dist_0* | 4.31 | 7.76 | 7.76 |
| | *verb_edges_dist_1* | 7.13 | 9.37 | 11.88 |
| | *verb_edges_dist_3* | 19.21 | 24.32 | 24.32 |
| | *verb_edges_dist_5* | 1.34 | 4.07 | 10.20 |
| | *verbal_head_per_sent* | 1.82 | 2.74 | 2.74 |
| | *verbal_root_perc* | 88 | 94.07 | 71.2 |

## Pedagogical Role of Concepts Analysis

Table B.2: Average values of the features showing a significant difference (Mann-Whitney U Test p-val<0.05) between the groups of texts referring to 'PN' and 'LO' concepts in the pedagogical role analysis discussed in Section 4.2.4 of Chapter 4. 'PN' refers to primary notions (i.e., first concepts in the sequences), while 'LO' refers to learning outcomes (i.e., last concepts in the sequences).

| | | **AVERAGE VALUES** | |
|---|---|---|---|
| **Property Type** | **Feature** | **PN** | **LO** |
| Deprel Distribution | *dep_dist_aux* | 1.39 | 0.85 |
| | *dep_dist_cc* | 4.05 | 2.38 |
| | *dep_dist_ccomp* | 0.40 | 0.66 |
| | *dep_dist_compound* | 1.56 | 2.83 |
| | *dep_dist_conj* | 6.60 | 3.24 |
| | *dep_dist_det* | 10.22 | 12.92 |
| | *dep_dist_discourse* | 0.00 | 0.25 |
| | *dep_dist_flat* | 0.04 | 0.25 |
| | *dep_dist_nummod* | 0.45 | 1.63 |
| UPOS Distribution | *upos_dist_CCONJ* | 4.02 | 2.37 |
| | *upos_dist_DET* | 10.37 | 13.12 |
| | *upos_dist_NOUN* | 27.65 | 25.46 |
| | *upos_dist_NUM* | 0.64 | 2.37 |
| | *upos_dist_SYM* | 0.00 | 0.66 |
| Inflectional Morphology | *aux_form_dist_Fin* | 82.81 | 93.78 |
| | *aux_form_dist_Inf* | 11.82 | 5.30 |

| | | | |
|---|---|---|---|
| | *aux_num_pers_dist_Sing+3* | 59.02 | 71.48 |
| Lexical Density | *lexical_density* | 0.57 | 0.52 |
| Ordering of Elements | *obj_pre* | 0.00 | 2.33 |
| Raw Text Properties | *char_per_tok* | 5.00 | 4.67 |
| Verbal Predicate Struct | *verb_edges_dist_1* | 9.21 | 14.53 |
| | *verb_edges_dist_2* | 33.71 | 23.97 |

# Gold-PR Dataset Analysis

## Primary Notion and Learning Outcome Analysis

Table B.3: Average values of the features showing a significant difference (Mann-Whitney U Test p-val<0.05) between the groups of sentences referring to 'PN' and 'LO' concepts in the Primary Notions and Learning Outcomes Comparison discussed in Section 7.5.2 of Chapter 7. 'PN' refers to the primary notions (i.e., concepts with only in-coming edges in the prerequisite graph structure), while 'LO' refers to the learning outcomes (i.e., concepts with only out-going edges in the prerequisite graph structure) of the manually annotated Gold-PR dataset.

| | | **AVERAGE VALUES** | |
|---|---|---|---|
| **Property Type** | **Feature** | **PN** | **LO** |
| Deprel Distribution | *dep_dist_acl:relcl* | 1.82 | 1.12 |
| | *dep_dist_det* | 10.97 | 13.39 |
| | *dep_dist_fixed* | 0.45 | 0.18 |
| | *dep_dist_nsubj* | 4.87 | 3.64 |
| | *dep_dist_obl* | 5.96 | 7.67 |
| | *dep_dist_punct* | 10.53 | 9.08 |
| Ordering of Elements | *obj_pre* | 4.24 | 0.00 |
| UPOS Distribution | *upos_dist_DET* | 11.10 | 13.57 |
| | *upos_dist_PART* | 1.88 | 2.64 |
| | *upos_dist_PRON* | 4.41 | 2.33 |
| | *upos_dist_PUNCT* | 10.60 | 9.08 |
| | *upos_dist_SCONJ* | 2.31 | 1.44 |
| Verbal Predicate Struct | *verb_edges_dist_2* | 26.78 | 19.43 |
| | *verb_edges_dist_4* | 10.62 | 20.44 |
| | *verbs_form_dist_Fin* | 28.01 | 19.93 |
| | *avg_verb_edges* | 2.36 | 2.84 |
| Inflectional Morphology | *verbs_form_dist_Part* | 29.04 | 45.31 |
| | *verbs_tense_dist_Past* | 48.62 | 64.97 |

| Profiling-UD Complete List of Features | | |
|---|---|---|
| **Annotation Level** | **Linguistic Feature** | **Label** |
| | **Raw Text Properties (*RawText*)** | |
| Raw Text | Sentence Length | sent_length |
| | Word Length | char_per_tok |
| | **Vocabulary Richness (*Vocabulary*)** | |
| Vocabulary | Type/Token Ratio for words and lemmas | ttr_form, ttr_lemma |
| | **Morphosyntactic information (*POS*)** | |
| | Distibution of UD and language–specific POS | upos_dist_, xpos_dist_ |
| POS tagging | Lexical density | lexical_density |
| | **Inflectional morphology (*VerbInflection*)** | |
| | Inflectional morphology of lexical verbs and auxiliaries | xpos_VB-VBD-VBP-VBZ, aux_* |
| | **Verbal Predicate Structure (*VerbPredicate*)** | |
| | Distribution of verbal heads and verbal roots | verbal_head_dist, verbal_root_perc |
| | Verb arity and distribution of verbs by arity | avg_verb_edges, verbal_arity_* |
| | **Global and Local Parsed Tree Structures (*TreeStructure*)** | |
| | Depth of the whole syntactic tree | parse_depth |
| Dependency Parsing | Average length of dependency links and of the longest link | avg_links_len, max_links_len |
| | Average length of prepositional chains and distribution by depth | avg_prep_chain_len, prep_dist_* |
| | Clause length | avg_token_per_clause |
| | **Order of elements (*Order*)** | |
| | Relative order of subject and object | subj_pre, obj_post |
| | **Syntactic Relations (*SyntacticDep*)** | |
| | Distribution of dependency relations | dep_dist_* |
| | **Use of Subordination (*Subord*)** | |
| | Distribution of subordinate and principal clauses | principal_prop_dist, subordinate_prop_dist |
| | Average length of subordination chains and distribution by depth | avg_subord_chain_len, subordinate_dist_1 |
| | Relative order of subordinate clauses | subordinate_post |

Table B.4: Complete List of Features Monitored by Profiling-UD .

APPENDIX C

# PRET USABILITY TEST

## Scenario

Imagine you are a researcher working in the field of Education and you want to study how prerequisite relations are established between concepts in educational materials. Your first challenge is to observe how prerequisite relations are represented in texts, thus you need to manually annotate the relations in order to create a dataset. At this point imagine that you want to use the dataset to analyse the relations and to compare your annotation with the one that can be obtained using automatic prerequisite extraction methods. In what follows you'll find the details about the steps that you need to do in order to achieve your goals using PRET.

## Tasks

Login into PRET tool by creating a new account and then use the tool to complete the following tasks.

### Upload your Data

1. Upload to PRET the following short text extracted from the Wikipedia page about Algebra. Name it "Your_Name's Text" and fill the Chapter field with "1". Fill the other metadata fields as you want. Annotate the chapter and sub-chapter titles to add information about text structure.

1 Intro
Algebra is one of the broad parts of mathematics, together with number theory, geometry and analysis. In its most general form, algebra is the study of mathematical symbols and the rules for manipulating these symbols; it is a unifying thread of almost all of mathematics. It includes everything from elementary equation solving to the study of abstractions such as groups, rings, and fields. The more basic parts of algebra are called elementary algebra; the more abstract parts are called abstract algebra or modern algebra. Elementary algebra is generally considered to be essential for any study of mathematics, science, or engineering, as well as such applications as medicine and economics. Abstract algebra is a major area in advanced mathematics, studied primarily by professional mathematicians.
1.1 Elementary Algebra
Elementary algebra differs from arithmetic in the use of abstractions, such as using letters to stand for numbers that are either unknown or allowed to take on many values. For example, in $x + 2 = 5$ the letter x is unknown, but applying additive inverses can reveal its value: $x = 3$. In $E = mc2$, the letters E and m are variables, and the letter c is a constant, the speed of light in a vacuum. Algebra gives methods for writing formulas and solving equations that are much clearer and easier than the older method of writing everything out in words. The word algebra is also used in certain specialized ways. A special kind of mathematical object in abstract algebra is called an "algebra", and the word is used, for example, in the phrases linear algebra and algebraic topology.
1.2 Algebraist
A mathematician who does research in algebra is called an algebraist.

2. Download the CoNLL file of your text.
3. Upload a terminology associated with “Your_Name’s Text” containing the following terms. Make sure that the terminology complies with the requirements!

| algebra, mathematics, number theory, symbol, equation, arithmetic, numbers, formula, algebraist |
|---|

From now on, you will use the text “Computer Science: An overview – Chapter 5”.

### Text Annotation

Create a new annotation for “Computer Science: An overview – Chapter 5”. Read **Annotation Guidelines** and perform the following tasks:

1. Add a strong relation between ALGORITHM and COMPUTER SCIENCE where ALGORITHM is the target concept and COMPUTER SCIENCE is its prerequisite.
2. Add at least 5 more prerequisite relations of your choice (using whatever concept in the text).
3. Add two new concepts to the terminology. Create at least one prerequisite relation using one of them.
4. Delete one of the two manually added concepts (possibly the one involved in fewer relations).
5. Revise your annotation:

   a Change the weight of one of the relations.
   b Delete a relation with the motivation that the two concepts are "Too far".
   c Confirm three relations and provide a motivation for keeping them.

6. Create a Gold Standard merging all annotations available for Computer Science – Chapter 5 (also if your annotation is the only one available). Choose a combination criterion that determines maximum inclusion.
7. Name the new gold dataset you created "Your_Name's Gold".

### Analyse the Annotation

1. Find the following information about "Computer Science: An overview – Chapter 5":

   a Number of sentences in the text.
   b Number of unique relations entered by you.
   c Consult the linguistic analysis of the sentence where you entered the relation between ALGORITHM and COMPUTER SCIENCE. Which is the grammatical category (verb, adposition, noun, etc.) of ALGORITHM?
   d If there are other annotations available for "Computer Science: An overview – Chapter 5", compute the agreement between your annotation and another one of your choice.

### Prerequisite Extraction using Automatic Methods

Explore the annotation produced by some automatic prerequisite extraction methods implemented in PRET and compare it with your manual annotation.

1. View the results obtained by the Relational Metric run on "Computer Science: An overview – Chapter 5" (set the threshold to 0 if not Succeeded already). Download the result file.
2. Compare the result obtained by the Temporal Pattern method (threshold defined by you if not Succeeded already) and your gold dataset.

   a Find if there are any common or opposite relations.

### Annotation Visualisation

1. Visualize your manual annotation as a matrix where concepts are ordered by frequency.
2. Visualize the result of the Temporal Pattern extraction method as a Gantt graph.

## Questionnaires

### SUS

Answers range from 1 (strongly disagree) to 5 (Strongly Agree).

1. I think that I would like to use PRET frequently [for prerequisite annotation].
2. I found PRET unnecessarily complex.
3. I thought PRET was easy to use.
4. I think that I would need the support of a technical person to use PRET.
5. I found the various functions in PRET were well integrated.

   a If not, which functions weren't well integrated?

6. I thought there was too much inconsistency in PRET.

   a If yes, what did you find inconsistent (e.g., the use of terms, the design, other)?

7. I would imagine that most people working in the field of Education would learn how to use PRET quickly.
8. I found PRET very difficult to use.

   a If yes, what did you find difficult?

9. I felt very confident using PRET.
10. I needed to learn a lot of things before I could get going with PRET.

    a If yes, were the instructions clear enough to let you understand how PRET works?

### PSSUQ

Answers range from 1 (strongly agree) to 7 (Strongly disagree).

1. Overall, I am satisfied with how easy it is to use PRET
2. It was simple to use PRET
3. I was able to complete the tasks and scenarios quickly using PRET
4. I felt comfortable using PRET
5. It was easy to learn to use PRET
6. I believe I could become productive quickly using PRET
7. PRET gave error messages that clearly told me how to fix problems.
8. Whenever I made a mistake using PRET, I could recover easily and quickly.

9. The information (such as online help, on-screen messages, and other documentation) provided with PRET was clear
10. It was easy to find the information I needed
11. The information was effective in helping me complete the tasks and scenarios
12. The organization of information on PRET screens was clear
13. The interface of PRET was pleasant
14. I liked using the interface of PRET
15. PRET has all the functions and capabilities I expect it to have.
16. Overall, I am satisfied with PRET.

## Test Results

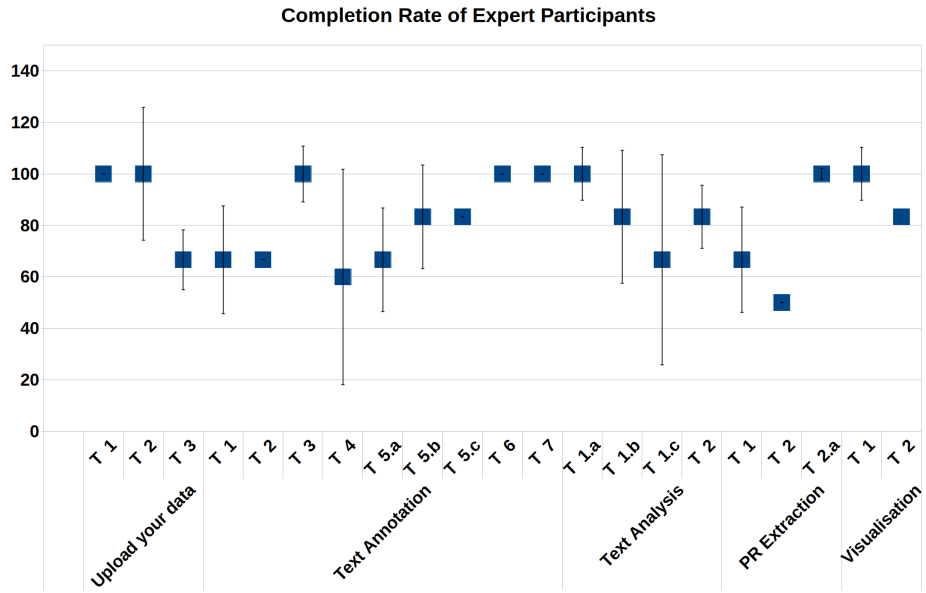



Figure C.1: Effectiveness analysis: completion rate for each sub-task for *expert users* of the usability test (i.e. each dot represents a task). Error bars represent standard deviation values.

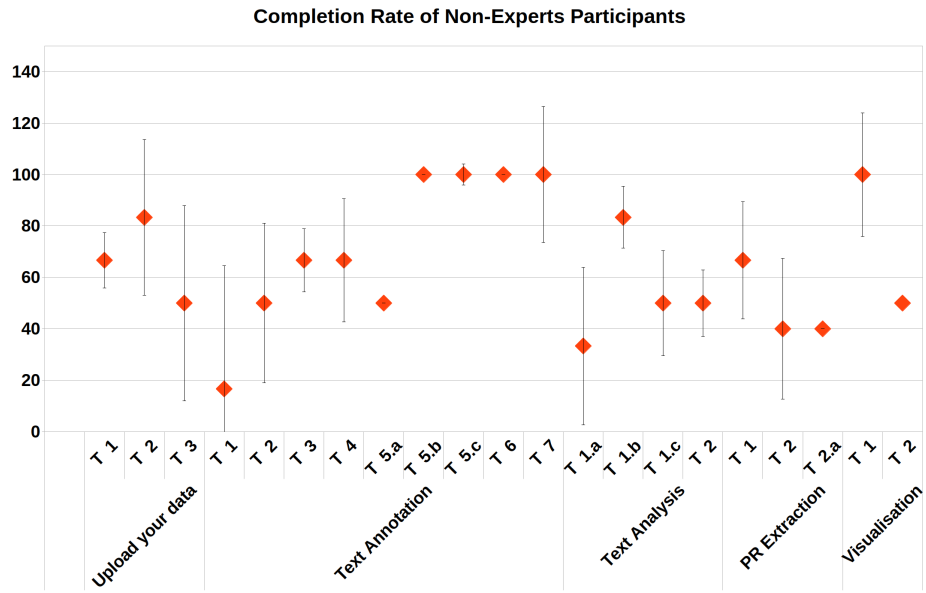



Figure C.2: Effectiveness analysis: completion rate for each sub-task for *non–expert users* of the usability test (i.e. each dot represents a task). Error bars represent standard deviation values.

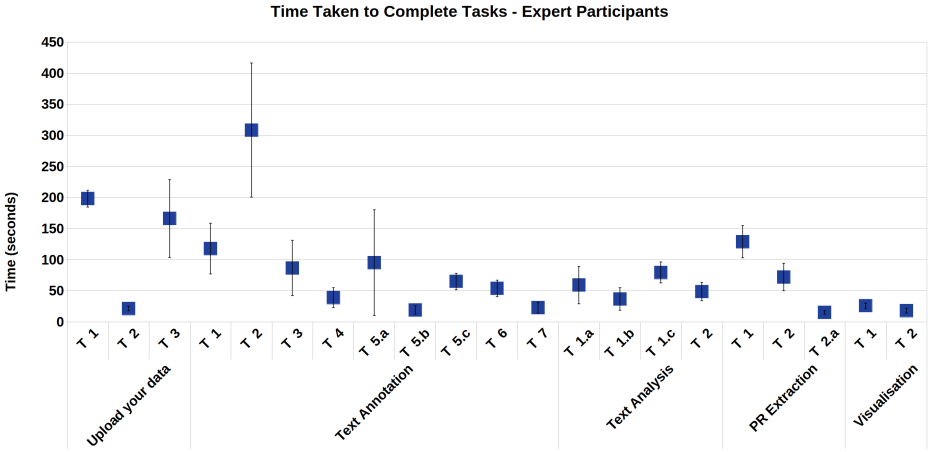



Figure C.3: Efficiency analysis: time taken to complete each sub-task for *expert users* of the usability test (i.e. each dot represents a task). Error bars represent standard deviation values.

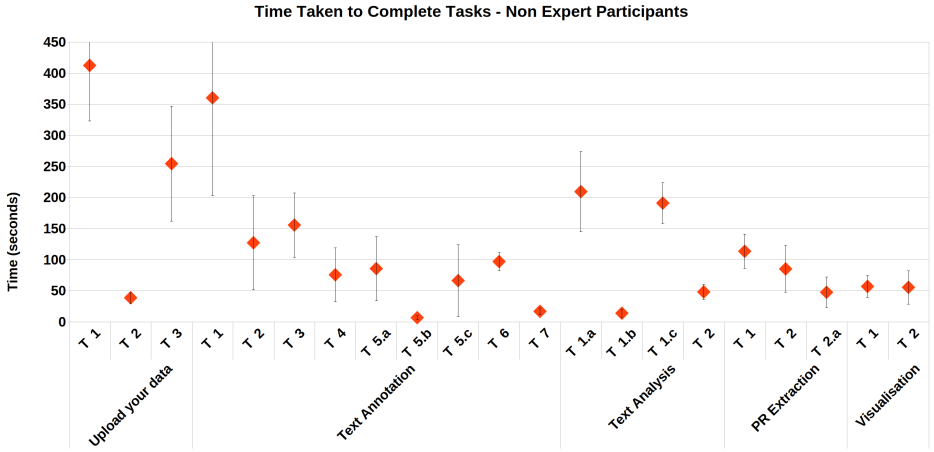



Figure C.4: Efficiency analysis: time taken to complete each sub-task for *non–expert users* of the usability test (i.e. each dot represents a task). Error bars represent standard deviation values.

# BIBLIOGRAPHY


[1] G. ADORNI, C. ALZETTA, F. KOCEVA, S. PASSALACQUA, AND I. TORRE, *Towards the identification of propaedeutic relations in textbooks*, in International Conference on Artificial Intelligence in Education, Springer, 2019, pp. 1–13.

[2] G. ADORNI, F. DELL'ORLETTA, F. KOCEVA, I. TORRE, AND G. VENTURI, *Extracting dependency relations from digital learning content*, in Italian Research Conference on Digital Libraries, Springer, 2018, pp. 114–119.

[3] G. ADORNI AND F. KOCEVA, *Designing a knowledge representation tool for subject matter structuring*, in International Workshop on Graph Structures for Knowledge Representation and Reasoning, Springer, 2015, pp. 1–14.

[4] R. AGRAWAL, B. GOLSHAN, AND E. PAPALEXAKIS, *Toward data-driven design of educational courses: a feasibility study*, Journal of Educational Data Mining, 8 (2016), pp. 1–21.

[5] J. F. ALLEN, *Maintaining knowledge about temporal intervals*, Communications of the ACM, 26 (1983).

[6] F. ALSAAD, A. BOUGHOULA, C. GEIGLE, H. SUNDARAM, AND C. ZHAI, *Mining mooc lecture transcripts to construct concept dependency graphs.*, Int. Educ. Data Mining Society, (2018).

[7] C. ALZETTA, G. ADORNI, I. CELIK, F. KOCEVA, AND I. TORRE, *Toward a user-adapted question/answering educational approach*, in Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, 2018, pp. 173–177.

[8] C. ALZETTA, F. DELL'ORLETTA, S. MONTEMAGNI, P. OSENOVA, K. SIMOV, AND G. VENTURI, *Quantitative linguistic investigations across universal dependencies treebanks*, in Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it), 2020.

[9] C. ALZETTA, F. DELL'ORLETTA, S. MONTEMAGNI, AND G. VENTURI, *Uncovering typological context-sensitive features*, in Proceedings of the Second Workshop on Computational Research in Linguistic Typology, 2020.

[10] C. ALZETTA, F. DELL'ORLETTA, S. MONTEMAGNI, M. SIMI, AND G. VENTURI, *Assessing the impact of incremental error detection and correction. a case study on the italian universal dependency treebank*, in Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), 2018, pp. 1–7.

[11] C. ALZETTA, F. DELL'ORLETTA, S. MONTEMAGNI, AND G. VENTURI, *Universal dependencies and quantitative typological trends. a case study on word order*, in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[12] ——, *Inferring quantitative typological trends from multilingual treebanks. a case study*, Lingue e linguaggio, 18 (2019), pp. 209–242.

[13] C. ALZETTA, I. GALLUCCIO, F. KOCEVA, S. PASSALACQUA, AND I. TORRE, *Digging into prerequisite relations*, in Second Workshop on Intelligent Textbooks, 2020.

[14] C. ALZETTA, F. KOCEVA, S. PASSALACQUA, I. TORRE, AND G. ADORNI, *Pret: Prerequisite-enriched terminology. a case study on educational texts*, in Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018, 2018.

[15] C. ALZETTA, A. MIASCHI, G. ADORNI, F. DELL'ORLETTA, F. KOCEVA, S. PASSALACQUA, AND I. TORRE, *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, vol. 2481, CEUR-WS, 2019.

[16] C. ALZETTA, A. MIASCHI, F. DELL'ORLETTA, F. KOCEVA, AND I. TORRE, *Prelearn@ evalita 2020: Overview of the prerequisite relation learning task for italian*, in Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online. CEUR. org, 2020.

[17] J. AMIDEI, P. PIWEK, AND A. WILLIS, *Rethinking the agreement in human evaluation tasks*, in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3318–3329.

[18] J. ANGEL, S. T. AROYEHUN, AND A. GELBUKH, *Nlp-cic @ prelearn: Mastering prerequisites relations, from handcrafted features to embeddings*, in Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), V. Basile, D. Croce, M. Di Maro, and L. C. Passaro, eds., Online, 2020, CEUR.org.

[19] R. ARTSTEIN, *Inter-annotator agreement*, in Handbook of linguistic annotation, Springer, 2017, pp. 297–313.

[20] R. ARTSTEIN AND M. POESIO, *Inter-coder agreement for computational linguistics*, Computational Linguistics, 34 (2008), pp. 555–596.

[21] D. J. ARYA, E. H. HIEBERT, AND P. D. PEARSON, *The effects of syntactic and lexical complexity on the comprehension of elementary science texts*, International Electronic Journal of Elementary Education, 4 (2011), pp. 107–125.

[22] M. N. ASIM, M. WASIM, M. U. G. KHAN, W. MAHMOOD, AND H. M. ABBASI, *A survey of ontology learning techniques and applications*, Database, 2018 (2018).

[23] S. ATKINS, J. CLEAR, AND N. OSTLER, *Corpus design criteria*, Literary and linguistic computing, 7 (1992), pp. 1–16.

[24] I. AUGENSTEIN, M. DAS, S. RIEDEL, L. VIKRAMAN, AND A. MCCALLUM, *Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications*, in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 546–555.

[25] D. P. AUSUBEL, J. D. NOVAK, H. HANESIAN, ET AL., *Educational psychology: A cognitive view*, Holt, Rinehart and Winston: New York, 1968.

[26] M. C. AYTEKIN, S. RABIGER, AND Y. SAYGIN, *Discovering the prerequisite relationships among instructional videos from subtitles*, in International conference on artificial intelligence in education, 2020, pp. 569–573.

[27] N. BACH AND S. BADASKAR, *A review of relation extraction*, Literature review for Language and Statistics II, 2 (2007), pp. 1–15.

[28] M. BARONI, S. BERNARDINI, A. FERRARESI, AND E. ZANCHETTA, *The wacky wide web: a collection of very large linguistically processed web-crawled corpora*, Language resources and evaluation, 43 (2009), pp. 209–226.

[29] V. BASILE, D. CROCE, M. DI MARO, AND L. C. PASSARO, *Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian*, in Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), V. Basile, D. Croce, M. Di Maro, and L. C. Passaro, eds., Online, 2020, CEUR.org.

[30] P. S. BAYERL AND K. I. PAUL, *Identifying sources of disagreement: Generalizability theory in manual annotation studies*, Computational Linguistics, 33 (2007), pp. 3–8.

[31] ——, *What determines inter-coder agreement in manual annotations? a meta-analytic investigation*, Computational Linguistics, 37 (2011), pp. 699–725.

[32] B. BEIGMAN KLEBANOV AND E. BEIGMAN, *Difficult cases: From data to learning, and back*, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, Maryland, June 2014, Association for Computational Linguistics, pp. 390–396.

[33] E. M. BENNETT, R. ALPERT, AND A. GOLDSTEIN, *Communications through limited-response questioning*, Public Opinion Quarterly, 18 (1954), pp. 303–308.

[34] N. BEVAN, *Iso 9241: Ergonomic requirements for office work with visual display terminals (vdts)-part 11: Guidance on usability*, Tc, 159 (1998), p. 61.

[35] D. BIBER, *Representativeness in corpus design*, Literary and linguistic computing, 8 (1993), pp. 243–257.

[36] C. BIEMANN, K. BONTCHEVA, R. E. DE CASTILHO, I. GUREVYCH, AND S. M. YIMAM, *Collaborative web-based tools for multi-layer text annotation*, in Handbook of Linguistic Annotation, Springer, 2017, pp. 229–256.

[37] J. BINDER, *Package 'bursts': Markov model for bursty behavior in streams*, 2014. R package version 1.0-1.

[38] F. BONIN, F. DELL'ORLETTA, G. VENTURI, AND S. MONTEMAGNI, *A contrastive approach to multi-word term extraction from domain corpora*, in Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010.

[39] D. BOULANGER AND V. KUMAR, *An overview of recent developments in intelligent e-textbooks and reading analytics.*, in iTextbooks@ AIED, 2019, pp. 44–56.

[40] R. J. BRACHMAN, H. J. LEVESQUE, AND R. REITER, *Knowledge representation*, MIT press, 1992.

[41] J. BROOKE, *Sus: a "quick and dirty'usability*, Usability evaluation in industry, 189 (1996).

[42] G. BROOKSHEAR AND D. BRYLOW, *Computer Science: An Overview, Global Edition*, Pearson Education Limited., 2015, ch. 4 Networking and the Internet.

[43] D. BRUNATO, A. CIMINO, F. DELL'ORLETTA, G. VENTURI, AND S. MONTEMAGNI, *Profiling-ud: a tool for linguistic profiling of texts*, in Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 7145–7151.

[44] D. BRUNATO, L. DE MATTEI, F. DELL'ORLETTA, B. IAVARONE, AND G. VENTURI, *Is this sentence difficult? do you agree?*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2690–2699.

[45] E. BRUNSKILL, *Estimating prerequisite structure from noisy data.*, in EDM, Citeseer, 2011, pp. 217–222.

[46] P. BRUSILOVSKY AND E. MILLÁN, *User models for adaptive hypermedia and adaptive educational systems*, in The adaptive web, Springer, 2007, pp. 3–53.

[47] P. BRUSILOVSKY AND J. VASSILEVA, *Course sequencing techniques for large-scale web-based education*, International Journal of Continuing Engineering Education and Life Long Learning, 13 (2003), pp. 75–94.

[48] S. BUCHHOLZ AND E. MARSI, *Conll-x shared task on multilingual dependency parsing*, in Proceedings of the tenth conference on computational natural language learning (CoNLL-X), 2006, pp. 149–164.

[49] P. BUITELAAR, P. CIMIANO, AND B. MAGNINI, *Ontology learning from text: An overview*, Ontology learning from text: Methods, evaluation and applications, 123 (2005), pp. 3–12.

[50] J. BURSTEIN, *Opportunities for natural language processing research in education*, in International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2009, pp. 6–27.

[51] T. BYRT, J. BISHOP, AND J. B. CARLIN, *Bias, prevalence and kappa*, Journal of clinical epidemiology, 46 (1993), pp. 423–429.

[52] M. T. CABRÉ, *Terminology: Theory, methods and applications*, vol. 1, John Benjamins Publishing, 1999.

[53] S. CAREY, *The origin of concepts*, Oxford University Press, 2009.

[54] P. F. CARVALHO, M. GAO, B. A. MOTZ, AND K. R. KOEDINGER, *Analyzing the relative learning benefits of completing required activities and optional readings in online courses.*, International Educational Data Mining Society, (2018).

[55] J. M. CEJUELA, P. MCQUILTON, L. PONTING, S. J. MARYGOLD, R. STEFANCSIK, G. H. MILLBURN, AND B. ROST, *tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles*, Database, 2014 (2014).

[56] D. S. CHAPLOT, Y. YANG, J. G. CARBONELL, AND K. R. KOEDINGER, *Data-driven automated induction of prerequisite structure graphs*, in Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016, Raleigh, North Carolina, USA, June 29 - July 2, 2016, 2016, pp. 318–323.

[57] H. CHAU, I. LABUTOV, K. THAKER, D. HE, AND P. BRUSILOVSKY, *Automatic concept extraction for domain and student modeling in adaptive textbooks*, International Journal of Artificial Intelligence in Education, (2020), pp. 1–27.

[58] P. CHEN, Y. LU, V. W. ZHENG, AND Y. PIAN, *Prerequisite-driven deep knowledge tracing*, in 2018 IEEE International Conference on Data Mining (ICDM), IEEE, 2018, pp. 39–48.

[59] W. CHEN, A. S. LAN, D. CAO, C. BRINTON, AND M. CHIANG, *Behavioral analysis at scale: Learning course prerequisite structures from learner clickstreams.*, International Educational Data Mining Society, (2018).

[60] Y. CHEN, J. P. GONZÁLEZ-BRENES, AND J. TIAN, *Joint discovery of skill prerequisite graphs and student models.*, International Educational Data Mining Society, (2016).

[61] N. CHOMSKY, *Aspects of the Theory of Syntax*, vol. 11, MIT press, 1965.

[62] D. V. CICCHETTI AND A. R. FEINSTEIN, *High agreement but low kappa: Ii. resolving the paradoxes*, Journal of clinical epidemiology, 43 (1990), pp. 551–558.

[63] P. CIMIANO AND J. VÖLKER, *Text2onto. natural language processing and information systems*, in 10th International Conference on Applications of Natural Language to Information Systems, NLDB, 2005, pp. 15–17.

[64] J. COHEN, *A coefficient of agreement for nominal scales*, Educational and psychological measurement, 20 (1960), pp. 37–46.

[65] M. E. COLOSIMO, A. A. MORGAN, A. S. YEH, J. B. COLOMBE, AND L. HIRSCHMAN, *Data preparation and interannotator agreement: Biocreative task 1b*, BMC bioinformatics, 6 (2005), pp. 1–8.

[66] A. T. CORBETT AND J. R. ANDERSON, *Knowledge tracing: Modeling the acquisition of procedural knowledge*, User modeling and user-adapted interaction, 4 (1994), pp. 253–278.

[67] N. R. COUNCIL ET AL., *How people learn: Brain, mind, experience, and school: Expanded edition*, National Academies Press, 2000.

[68] W. CROFT, D. NORDQUIST, K. LOONEY, AND M. REGAN, *Linguistic Typology meets Universal Dependencies*, in Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15), Bloomington, IN, USA, January 20-21, 2017, vol. 1779 of CEUR Workshop Proceedings, CEUR-WS.org, 2017, pp. 63–75.

[69] J. G. CROMLEY, L. E. SNYDER-HOGAN, AND U. A. LUCIW-DUBAS, *Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of reading comprehension.*, Journal of Educational Psychology, 102 (2010), p. 687.

[70] C. CUCCHIARINI AND H. STRIK, *Automatic phonetic transcription: An overview*, in Proceedings of ICPHS, Citeseer, 2003, pp. 347–350.

[71] Y. DAI, M. YOSHIKAWA, AND K. SUGIYAMA, *Prerequisite-aware course ordering towards getting relevant job opportunities*, Expert Systems with Applications, (2021), p. 115233.

[72] E. DALE AND J. S. CHALL, *The concept of readability*, Elementary English, 26 (1949), pp. 19–26.

[73] Q.-V. DANG AND C.-L. IGNAT, *Measuring quality of collaboratively edited documents: the case of wikipedia*, in 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), IEEE, 2016, pp. 266–275.

[74] J. DANIEL, *Education and the covid-19 pandemic*, Prospects, 49 (2020), pp. 91–96.

[75] R. E. DE CASTILHO, E. MUJDRICZA-MAYDT, S. M. YIMAM, S. HARTMANN, I. GUREVYCH, A. FRANK, AND C. BIEMANN, *A web-based tool for the integrated annotation of semantic and syntactic structures*, in Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), 2016, pp. 76–84.

[76] O. DE CLERCQ, V. HOSTE, B. DESMET, P. VAN OOSTEN, M. DE COCK, AND L. MACKEN, *Using the crowd for readability prediction*, Natural Language Engineering, 20 (2014), pp. 293–325.

[77] M.-C. DE MARNEFFE AND C. D. MANNING, *The stanford typed dependencies representation*, in Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation, 2008, pp. 1–8.

[78] M.-C. DE MARNEFFE AND C. POTTS, *Developing linguistic theories using annotated corpora*, in Handbook of Linguistic Annotation, Springer, 2017, pp. 411–438.

[79] C. DE MEDIO, F. GASPARETTI, C. LIMONGELLI, F. SCIARRONE, AND M. TEMPERINI, *A machine learning approach to identify dependencies among learning objects.*, in CSEDU (1), 2016, pp. 345–352.

[80] S. DEERWESTER, S. T. DUMAIS, G. W. FURNAS, T. K. LANDAUER, AND R. HARSHMAN, *Indexing by latent semantic analysis*, Journal of the American society for information science, 41 (1990), pp. 391–407.

[81] F. DELL'ORLETTA, G. VENTURI, A. CIMINO, AND S. MONTEMAGNI, *T2k$^2$: a system for automatically extracting and organizing knowledge from texts*, in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), 2014.

[82] F. DELL'ORLETTA, M. WIELING, G. VENTURI, A. CIMINO, AND S. MONTEMAGNI, *Assessing the readability of sentences: which corpora and features?*, in Proceedings of the

Ninth Workshop on Innovative Use of NLP for Building Educational Applications, 2014, pp. 163–173.

[83] M. C. DESMARAIS, A. MALUF, AND J. LIU, *User-expertise modeling with empirically derived probabilistic implication networks*, User modeling and user-adapted interaction, 5 (1995), pp. 283–315.

[84] M. DICKINSON, *Detection of annotation errors in corpora*, Language and Linguistics Compass, 9 (2015), pp. 119–138.

[85] S. DIPPER, M. GÖTZE, AND S. SKOPETEAS, *Towards user-adaptive annotation guidelines*, in Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora, 2004, pp. 23–30.

[86] S. DIPPER, M. GÖTZE, AND M. STEDE, *Simple annotation tools for complex annotation tasks: an evaluation*, in Proceedings of the LREC Workshop on XML-based richly annotated corpora, 2004, pp. 54–62.

[87] J.-P. DOIGNON AND J.-C. FALMAGNE, *Spaces for the assessment of knowledge*, International journal of man-machine studies, 23 (1985), pp. 175–196.

[88] B. J. ERICSON, M. J. GUZDIAL, AND B. B. MORRISON, *Analysis of interactive features designed to enhance learning in an ebook*, in Proceedings of the Eleventh Annual International Conference on International Computing Education Research, 2015, pp. 169–178.

[89] E. ESTELLÉS-AROLAS AND F. GONZÁLEZ-LADRÓN-DE-GUEVARA, *Towards an integrated crowdsourcing definition*, Journal of Information science, 38 (2012), pp. 189–200.

[90] B. D. EUGENIO AND M. GLASS, *The kappa statistic: A second look*, Computational linguistics, 30 (2004), pp. 95–101.

[91] A. R. FABBRI, I. LI, P. TRAIRATVORAKUL, Y. HE, W. TING, R. TUNG, C. WESTERFIELD, AND D. RADEV, *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), 2018, pp. 611–620.

[92] D. FAURE AND C. NEDELLEC, *Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium*, in International Conference on Knowledge Engineering and Knowledge Management, Springer, 1999, pp. 329–334.

[93] A. FERRARESI, E. ZANCHETTA, M. BARONI, AND S. BERNARDINI, *Introducing and evaluating ukwac, a very large web-derived corpus of english*, in Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google, 2008, pp. 47–54.

[94] M. A. FINLAYSON AND T. ERJAVEC, *Overview of annotation creation: Processes and tools*, in Handbook of Linguistic Annotation, Springer, 2017, pp. 167–191.

[95] J. L. FLEISS, *Measuring nominal scale agreement among many raters.*, Psychological bulletin, 76 (1971), p. 378.

[96] J. A. FODOR, *Concepts: Where cognitive science went wrong*, Oxford University Press, 1998.

[97] K. FORT, M. EHRMANN, AND A. NAZARENKO, *Towards a methodology for named entities annotation*, in Proceedings of the Third Linguistic Annotation Workshop, Association for Computational Linguistics, 2009, pp. 142–145.

[98] K. FORT, A. NAZARENKO, AND S. ROSSET, *Modeling the complexity of manual annotation tasks: a grid of analysis*, in International Conference on Computational Linguistics, 2012, pp. 895–910.

[99] K. T. FRANTZI, S. ANANIADOU, AND J. TSUJII, *The c-value/nc-value method of automatic recognition for multi-word terms*, in International conference on theory and practice of digital libraries, Springer, 1998, pp. 585–604.

[100] G. P. C. FUNG, J. X. YU, P. S. YU, AND H. LU, *Parameter free bursty events detection in text streams*, in Proceedings of the 31st international conference on Very large data bases, 2005, pp. 181–192.

[101] R. M. GAGNE, *The acquisition of knowledge.*, Psychological review, 69 (1962), p. 355.

[102] ——, *Learning hierarchies*, Educational psychologist, 6 (1968), pp. 1–9.

[103] F. GASPARETTI, C. DE MEDIO, C. LIMONGELLI, F. SCIARRONE, AND M. TEMPERINI, *Prerequisites between learning objects: Automatic extraction based on a machine learning approach*, Telematics and Informatics, 35 (2018), pp. 595–610.

[104] F. GASPARETTI, C. LIMONGELLI, AND F. SCIARRONE, *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, 2015, pp. 1–6.

[105] J. GILES, *Internet encyclopaedias go head to head*, Nature, 438 (2005), p. 7070.

[106] A. GLIOZZO AND C. STRAPPARAVA, *Semantic domains in computational linguistics*, Springer Science & Business Media, 2009.

[107] W. GOLIK, R. BOSSY, Z. RATKOVIC, AND C. NÉDELLEC, *Improving term extraction with linguistic analysis in the biomedical domain.*, Research in Computing Science, 70 (2013), pp. 157–172.

[108] J. GORDON, S. AGUILAR, E. SHENG, AND G. BURNS, *Structured generation of technical reading lists*, in Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017, pp. 261–270.

[109] J. GORDON, L. ZHU, A. GALSTYAN, P. NATARAJAN, AND G. BURNS, *Modeling concept dependencies in a scientific corpus*, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2016, pp. 866–875.

[110] S. T. GRIES AND A. L. BEREZ, *Linguistic annotation in/for corpus linguistics*, Handbook of linguistic annotation, (2017), pp. 379–409.

[111] T. R. GRUBER, *A translation approach to portable ontology specifications*, Knowledge acquisition, 5 (1993), pp. 199–220.

[112] U. GUT AND P. S. BAYERL, *Measuring the reliability of manual annotations of speech corpora*, in Speech Prosody 2004, International Conference, 2004.

[113] A. HALAVAIS AND D. LACKAFF, *An analysis of topical coverage of wikipedia*, Journal of computer-mediated communication, 13 (2008), pp. 429–440.

[114] H. V. HALTEREN, *Linguistic profiling for authorship recognition and verification*, in Proceedings ACL 2004, East Stroudsburg: Association for Computational Linguistics, 2004, pp. 199–206.

[115] S. HAN, J. YOON, AND J. YOO, *Discovering skill prerequisite structure through bayesian estimation and nested model comparison*, in Proc. International Conference on Educational Data Mining, ERIC, 2017, pp. 398–399.

[116] M. HAZMAN, S. R. EL-BELTAGY, AND A. RAFEA, *A survey of ontology learning approaches*, International Journal of Computer Applications, 22 (2011), pp. 36–43.

[117] Q. HE, K. CHANG, E.-P. LIM, AND J. ZHANG, *Bursty feature representation for clustering text streams*, in Proceedings of the 2007 SIAM International Conference on Data Mining, 2007, pp. 491–496.

[118] A. HIPPISLEY, D. CHENG, AND K. AHMAD, *The head-modifier principle and multilingual term extraction*, Natural Language Engineering, 11 (2005), pp. 129–157.

[119] L. HIRSCHMAN AND H. S. THOMPSON, *Overview of evaluation in speech and natural language processing*, in Survey of the state of the art in human language technology, S. Bird, B. Boguraev, M. Kay, D. McDonald, D. Hindle, and Y. Wilks, eds., vol. 12, Cambridge university press, 1997, pp. 409–414.

[120] N. HOLLENSTEIN, N. SCHNEIDER, AND B. WEBBER, *Inconsistency detection in semantic annotation*, in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 3986–3990.

[121] B. HOLLINGSWORTH AND S. TEUFEL, *Human annotation of lexical chains: Coverage and agreement measures*, in ELECTRA Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications (Beyond Bag of Words), 2005, p. 26.

[122] E. HOVY AND J. LAVID, *Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics*, International journal of translation, 22 (2010), pp. 13–36.

[123] G. HRIPCSAK AND A. WILCOX, *Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance*, Journal of the American Medical Informatics Association, 9 (2002), pp. 1–15.

[124] R. HÜBSCHER, *What's in a prerequisite*, in In International Conference on Advanced Learning Technology (ICALT 2001, Citeseer, 2001.

[125] N. IDE, N. CALZOLARI, J. ECKLE-KOHLER, D. GIBBON, S. HELLMANN, K. LEE, J. NIVRE, AND L. ROMARY, *Community standards for linguistically-annotated resources*, in Handbook of Linguistic Annotation, Springer, 2017, pp. 113–165.

[126] N. IDE AND J. PUSTEJOVSKY, *What does interoperability mean, anyway? toward an operational definition of interoperability for language technology*, in Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China, 2010.

[127] ——, *Handbook of linguistic annotation*, Springer, 2017.

[128] N. IDE AND K. SUDERMAN, *The linguistic annotation framework: a standard for annotation interchange and merging*, Language Resources and Evaluation, 48 (2014), pp. 395–418.

[129] *Language resource management – Linguistic Annotation Framework (LAF).*, standard, International Organization for Standardization, 2012.

[130] R. JACKENDOFF, *Toward an explanatory semantic representation*, Linguistic inquiry, 7 (1976), pp. 89–150.

[131] M. JAKUBÍČEK, A. KILGARRIFF, V. KOVÁŘ, P. RYCHLÝ, AND V. SUCHOMEL, *The tenten corpus family*, in 7th International Corpus Linguistics Conference CL, 2013, pp. 125–127.

[132] A. JATOWT AND K. TANAKA, *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, 2012, pp. 2607–2610.

[133] P. K. JENA, *Challenges and opportunities created by covid-19 for odl: A case study of ignou*, International Journal for Innovative Research in Multidisciplinary Field (IJIRMF), 6 (2020).

[134] C. JIA, Y. SHEN, Y. TANG, L. SUN, AND W. LU, *Heterogeneous graph neural networks for concept prerequisite relation learning in educational data*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 2021, Association for Computational Linguistics, pp. 2036–2047.

[135] I. T. JOLLIFFE AND J. CADIMA, *Principal component analysis: a review and recent developments*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374 (2016), p. 20150202.

[136] D. H. JONASSEN, *On the role of concepts in learning and instructional design*, Educational Technology Research and Development, 54 (2006), p. 177.

[137] R. KATAGALL, R. DADDE, R. GOUDAR, AND S. RAO, *Concept mapping in education and semantic knowledge representation: an illustrative survey*, Procedia computer science, 48 (2015), pp. 638–643.

[138] J.-D. KIM, T. OHTA, AND J. TSUJII, *Corpus annotation for mining biomedical events from literature*, BMC bioinformatics, 9 (2008), pp. 1–25.

[139] W. KINTSCH, *Comprehension: A paradigm for cognition*, Cambridge university press, 1998.

[140] J. KLEINBERG, *Bursty and hierarchical structure in streams*, Data Mining and Knowledge Discovery, 7 (2003), pp. 373–397.

[141] ——, *Temporal dynamics of on-line information streams*, in Data Stream Management, Springer, 2016, pp. 221–238.

[142] B. KLUGA, M. S. JASTI, V. NAPLES, AND R. FREEDMAN, *Adding intelligence to a textbook for human anatomy with a causal concept map based its.*, in iTextbooks@ AIED, 2019, pp. 124–134.

[143] K. R. KOEDINGER, A. T. CORBETT, AND C. PERFETTI, *The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning*, Cognitive science, 36 (2012), pp. 757–798.

[144] R. KOPER AND B. OLIVIER, *Representing the learning design of units of learning*, Journal of Educational Technology & Society, 7 (2004), pp. 97–111.

[145] G. KOUTRIKA, L. LIU, AND S. SIMSKE, *Generating reading orders over document collections*, in 2015 IEEE 31st International Conference on Data Engineering, IEEE, 2015, pp. 507–518.

[146] S. D. KRASHEN, *The input hypothesis: Issues and implications*, Addison-Wesley Longman Ltd, 1985.

[147] K. KRIPPENDORFF, *Reliability in content analysis: Some common misconceptions and recommendations*, Human communication research, 30 (2004), pp. 411–433.

[148] P. KVĚTOŇ AND K. OLIVA, *Achieving an almost correct pos-tagged corpus*, in International Conference on Text, Speech and Dialogue, Springer, 2002, pp. 19–26.

[149] I. LABUTOV, Y. HUANG, P. BRUSILOVSKY, AND D. HE, *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, 2017, pp. 907–915.

[150] J. R. LANDIS AND G. G. KOCH, *The measurement of observer agreement for categorical data*, Biometrics, 33 (1977), pp. 159–174.

[151] M. LAPATA AND A. LASCARIDES, *Learning sentence-internal temporal relations*, Journal of Artificial Intelligence Research, 27 (2006), pp. 85–117.

[152] M. LARRANAGA, A. CONDE, I. CALVO, J. A. ELORRIAGA, AND A. ARRUARTE, *Automatic generation of the domain module from electronic textbooks: method and validation*, IEEE transactions on knowledge and data engineering, 26 (2014), pp. 69–82.

[153] W. S. LASECKI, L. RELLO, AND J. P. BIGHAM, *Measuring text simplification with the crowd*, in Proceedings of the 12th Web for All Conference, 2015, pp. 1–9.

[154] K. H. LAU, T. LAM, B. H. KAM, M. NKHOMA, J. RICHARDSON, AND S. THOMAS, *The role of textbook learning resources in e-learning: A taxonomic study*, Computers & Education, 118 (2018), pp. 10–24.

[155] C.-H. LEE, G.-G. LEE, AND Y. LEU, *Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning*, Expert Systems with applications, 36 (2009), pp. 1675–1684.

[156] S. LEE, Y. PARK, AND W. C. YOON, *Burst analysis for automatic concept map creation with a single document*, Expert Systems with Applications, 42 (2015), pp. 8817–8829.

[157] ——, *Burst analysis for automatic concept map creation with a single document*, Expert Systems with Applications, 42 (2015), pp. 8817–8829.

[158] G. LEECH, *Corpus annotation schemes*, Literary and linguistic computing, 8 (1993), pp. 275–281.

[159] ——, *New resources, or just better old ones? the holy grail of representativeness*, in Corpus linguistics and the web, Brill Rodopi, 2007, pp. 133–149.

[160] J. R. LEWIS, *Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use*, International Journal of Human-Computer Interaction, 7 (1995), pp. 57–78.

[161] I. LI, A. R. FABBRI, S. HINGMIRE, AND D. RADEV, *R-vgae: Relational-variational graph autoencoder for unsupervised prerequisite chain learning*, in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 1147–1157.

[162] I. LI, A. R. FABBRI, R. R. TUNG, AND D. R. RADEV, *What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning*, Proceedings of AAAI 2019, (2019).

[163] Y. LI, Z. SHAO, X. WANG, X. ZHAO, AND Y. GUO, *A concept map-based learning paths automatic generation algorithm for adaptive learning systems*, IEEE Access, 7 (2018), pp. 245–255.

[164] C. LIANG, S. WANG, Z. WU, K. WILLIAMS, B. PURSEL, B. BRAUTIGAM, S. SAUL, H. WILLIAMS, K. BOWEN, AND C. GILES, *Bbookx: Building online open books for personalized learning*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, 2016.

[165] C. LIANG, Z. WU, W. HUANG, AND C. L. GILES, *Measuring prerequisite relations among concepts*, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1668–1674.

[166] C. LIANG, J. YE, S. WANG, B. PURSEL, AND C. L. GILES, *Investigating active learning for concept prerequisite learning*, Proc. EAAI, (2018).

[167] C. LIANG, J. YE, Z. WU, B. PURSEL, AND C. L. GILES, *Recovering concept prerequisite relations from university course dependencies.*, in AAAI, 2017, pp. 4786–4791.

[168] C. LIANG, J. YE, H. ZHAO, B. PURSEL, AND C. L. GILES, *Active learning of strict partial orders: A case study on concept prerequisite relations*, in Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5,

2019, M. C. Desmarais, C. F. Lynch, A. Merceron, and R. Nkambou, eds., International Educational Data Mining Society (IEDMS), 2019.

[169] C. LIMONGELLI, F. GASPARETTI, AND F. SCIARRONE, *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, 2015, pp. 1–6.

[170] C. LIMONGELLI, F. SCIARRONE, M. LOMBARDI, A. MARANI, AND M. TEMPERINI, *A framework for comparing concept maps*, in 2017 16th International Conference on Information Technology Based Higher Education and Training (ITHET), IEEE, 2017, pp. 1–6.

[171] S. LINDBLOM-YLÄNNE, K. TRIGWELL, A. NEVGI, AND P. ASHWIN, *How approaches to teaching are affected by discipline and teaching context*, Studies in Higher education, 31 (2006), pp. 285–298.

[172] N. LIPKA AND B. STEIN, *Identifying featured articles in wikipedia: writing style matters*, in Proceedings of the 19th international conference on World wide web, 2010, pp. 1147–1148.

[173] H. LIU, W. MA, Y. YANG, AND J. CARBONELL, *Learning concept graphs from online educational data*, Journal of Artificial Intelligence Research, 55 (2016), pp. 1059–1090.

[174] J. LIU, L. JIANG, Z. WU, Q. ZHENG, AND Y. QIAN, *Mining learning-dependency between knowledge units from text*, The VLDB Journal, 20 (2011), pp. 335–345.

[175] W. LU, P. MA, J. YU, Y. ZHOU, AND B. WEI, *Metro maps for efficient knowledge learning by summarizing massive electronic textbooks*, International Journal on Document Analysis and Recognition (IJDAR), (2019), pp. 1–13.

[176] W. LU, Y. ZHOU, J. YU, AND C. JIA, *Concept extraction and prerequisite relation learning from educational data*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 9678–9685.

[177] A. LÜDELING, M. WALTER, E. KROYMANN, AND P. ADOLPHS, *Multi-level error annotation in learner corpora*, in Proceedings of corpus linguistics, vol. 1, 2005, pp. 14–17.

[178] H. P. LUHN, *Key word-in-context index for technical literature (kwic index)*, American documentation, 11 (1960), pp. 288–295.

[179] I. MANI AND B. SCHIFFMAN, *Temporally anchoring and ordering events in news*, Time and Event Recognition in Natural Language, 9 (2005).

[180] C. MANNING AND H. SCHUTZE, *Foundations of statistical natural language processing*, MIT press, 1999.

[181] R. MANRIQUE, B. PEREIRA, O. MARINO, N. CARDOZO, AND S. WOLFGAND, *Towards the identification of concept prerequisites via knowledge graphs*, in 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), vol. 2161, IEEE, 2019, pp. 332–336.

[182] R. MANRIQUE, J. SOSA, O. MARINO, B. P. NUNES, AND N. CARDOZO, *Investigating learning resources precedence relations via concept prerequisite learning*, in 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, 2018, pp. 198–205.

[183] M. P. MARCUS, B. SANTORINI, AND M. A. MARCINKIEWICZ, *Building a large annotated corpus of English: The Penn Treebank*, Computational Linguistics, 19 (1993), pp. 313–330.

[184] S. J. MARGOLIN, C. DRISCOLL, M. J. TOLAND, AND J. L. KEGLER, *E-readers, computer screens, or paper: Does reading comprehension change across media platforms?*, Applied cognitive psychology, 27 (2013), pp. 512–519.

[185] J. L. MARTINEZ-RODRIGUEZ, A. HOGAN, AND I. LOPEZ-AREVALO, *Information extraction meets the semantic web: a survey*, Semantic Web, (2020), pp. 1–81.

[186] D. MAYNARD, *Benchmarking textual annotation tools for the semantic web*, in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 2008.

[187] T. MCENERY AND A. WILSON, *Corpus Linguistics: An Introduction*, Edinburgh University Press, 2001.

[188] T. MCENERY, R. XIAO, AND Y. TONO, *Corpus-based language studies: An advanced resource book*, Taylor & Francis, 2006.

[189] M. D. MERRILL, *First principles of instruction*, Educational technology research and development, 50 (2002), pp. 43–59.

[190] M. D. MERRILL, R. D. TENNYSON, AND L. O. POSEY, *Teaching concepts: An instructional design guide*, Educational Technology, 1992.

[191] Z. A. MERROUNI, B. FRIKH, AND B. OUHBI, *Automatic keyphrase extraction: a survey and trends*, Journal of Intelligent Information Systems, (2019), pp. 1–34.

[192] W. D. MEURERS AND S. MÜLLER, *Corpora and syntax*, Corpus linguistics: An international handbook, 29 (2009), pp. 920–933.

[193] A. MIASCHI, C. ALZETTA, F. A. CARDILLO, AND F. DELL'ORLETTA, *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), 2019.

[194] T. MIKOLOV, K. CHEN, G. CORRADO, AND J. DEAN, *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781, (2013).

[195] B. MILLER AND D. RANUM, *Runestone interactive: tools for creating interactive course materials*, in Proceedings of the first ACM conference on Learning@ scale conference, ACM, 2014, pp. 213–214.

[196] P. MONACHESI, T. MARKUS, V. POSEA, S. TRAUSAN-MATU, P. OSENOVA, AND K. SIMOV, *Supporting knowledge discovery in an elearning environment having social components*, in Technological Developments in Networking, Education and Automation, Springer, 2010, pp. 157–162.

[197] P. MONACHESI, K. SIMOV, E. MOSSEL, P. OSENOVA, AND L. LEMNITZER, *What ontologies can do for elearning*, Proceedings of IMCL 2008, (2008).

[198] S. MONTEMAGNI, *Strategie linguistiche della divulgazione scientifica: una prospettiva linguistico-computazionale*, in La linguistica della divulgazione, la divulgazione della linguistica. Atti del IV Convegno Interannuale SLI nuova serie, 2020, pp. 79–104.

[199] N. MOSTAFAZADEH, A. GREALISH, N. CHAMBERS, J. ALLEN, AND L. VANDERWENDE, *Caters: Causal and temporal relation scheme for semantic annotation of event structures*, in Proceedings of the Fourth Workshop on Events, 2016, pp. 51–61.

[200] M. L. MURPHY, *Lexical meaning*, Cambridge University Press, 2010.

[201] T. NAKAGAWA AND Y. MATSUMOTO, *Detecting errors in corpora using support vector machines*, in COLING 2002: The 19th International Conference on Computational Linguistics, 2002.

[202] S. NASH, *Learning objects, learning object repositories, and learning theory: Preliminary best practices for online courses*, Interdisciplinary Journal of E-Learning and Learning Objects, 1 (2005), pp. 217–228.

[203] M. NELSON, *Building a written corpus*, The Routledge handbook of corpus linguistics, (2010), p. 53.

[204] M. NEVES AND J. ŠEVA, *An extensive review of tools for manual annotation of documents*, Briefings in Bioinformatics, (2019).

[205] P.-T. NGUYEN, A.-C. LE, T.-B. HO, AND V.-H. NGUYEN, *Vietnamese treebank construction and entropy-based error detection*, Language Resources and Evaluation, 49 (2015), pp. 487–519.

[206] Q. NING, Z. FENG, H. WU, AND D. ROTH, *Joint reasoning for temporal and causal relations*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2278–2288.

[207] J. NIVRE, *Towards a universal grammar for natural language processing*, in International conference on intelligent text processing and computational linguistics, Springer, 2015, pp. 3–16.

[208] R. NKAMBOU, R. MIZOGUCHI, AND J. BOURDEAU, *Advances in intelligent tutoring systems*, vol. 308, Springer Science & Business Media, 2010.

[209] J. D. NOVAK, *Concept mapping: A useful tool for science education*, Journal of research in science teaching, 27 (1990), pp. 937–949.

[210] J. D. NOVAK AND A. J. CAÑAS, *The theory underlying concept maps and how to construct and use them*, (2008).

[211] R. E. O'CONNOR, K. M. BELL, K. R. HARTY, L. K. LARKIN, S. M. SACKOR, AND N. ZIGMOND, *Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty.*, Journal of Educational Psychology, 94 (2002), p. 474.

[212] P. OSENOVA AND K. SIMOV, *Semantic annotation for semi-automatic positioning of the learner*, in Proceedings of the First Workshop on Supporting eLearning with Language Resources and Semantic Data, at LREC, Citeseer, 2010, pp. 46–50.

[213] M. PALMER AND N. XUE, *Linguistic annotation*, The Handbook of Computational Linguistics and Natural Language Processing, (2010), pp. 238–270.

[214] L. PAN, J. CHEN, S. LIU, C.-W. NGO, M.-Y. KAN, AND T.-S. CHUA, *A hybrid approach for detecting prerequisite relations in multi-modal food recipes*, IEEE Transactions on Multimedia, (2020).

[215] L. PAN, C. LI, J. LI, AND J. TANG, *Prerequisite relation learning for concepts in moocs*, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1447–1456.

[216] L. PAN, X. WANG, C. LI, J. LI, AND J. TANG, *Course concept extraction in moocs via embedding-based graph propagation*, in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2017, pp. 875–884.

[217] L. PAPPANO, *The year of the mooc*, The New York Times, 2 (2012), p. 2012.

[218] S. PASSALACQUA, F. KOCEVA, C. ALZETTA, I. TORRE, AND G. ADORNI, *Visualisation analysis for exploring prerequisite relations in textbooks*, in First Workshop on Intelligent Textbooks, 2019.

[219] R. PELÁNEK, *Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques*, User Modeling and User-Adapted Interaction, 27 (2017), pp. 313–350.

[220] C. K. PEREIRA, J. F. MEDEIROS, S. W. SIQUEIRA, AND B. P. NUNES, *How complex is the complexity of a concept in exploratory search*, in 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), vol. 2161, IEEE, 2019, pp. 17–21.

[221] J. C. PETERSON, R. M. BATTLEDAY, T. L. GRIFFITHS, AND O. RUSSAKOVSKY, *Human uncertainty makes classification more robust*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9617–9626.

[222] S. PETROV, D. DAS, AND R. MCDONALD, *A universal part-of-speech tagset*, in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 2089–2096.

[223] C. PIECH, J. BASSEN, J. HUANG, S. GANGULI, M. SAHAMI, L. GUIBAS, AND J. SOHL-DICKSTEIN, *Deep knowledge tracing*, in Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1, 2015, pp. 505–513.

[224] B. PLANK, D. HOVY, AND A. SØGAARD, *Linguistically debatable or just plain wrong?*, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers), 2014, pp. 507–511.

[225] E. M. PONTI, H. O'HORAN, Y. BERZAK, I. VULIĆ, R. REICHART, T. POIBEAU, E. SHUTOVA, AND A. KORHONEN, *Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing*, Computational Linguistics, 45 (2019), pp. 559–601.

[226] M. PONTIKI, D. GALANIS, H. PAPAGEORGIOU, I. ANDROUTSOPOULOS, S. MANANDHAR, M. AL-SMADI, M. AL-AYYOUB, Y. ZHAO, B. QIN, O. DE CLERCQ, ET AL., *Semeval-2016 task 5: Aspect based sentiment analysis*, in 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016.

[227] M. PRINCE, *Does active learning work? a review of the research*, Journal of engineering education, 93 (2004), pp. 223–231.

[228] V. PUNCREOBUTR, *Education 4.0: New challenge of learning*, St. Theresa Journal of Humanities and Social Sciences, 2 (2016).

[229] J. PUSTEJOVSKY, *Unifying linguistic annotations: A timeml case study*, in Proceedings of Text, Speech, and Dialogue Conference, 2006.

[230] J. PUSTEJOVSKY AND A. STUBBS, *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*, " O'Reilly Media, Inc.", 2012.

[231] S. PYYSALO, T. OHTA, M. MIWA, H.-C. CHO, J. TSUJII, AND S. ANANIADOU, *Event extraction across multiple levels of biological organization*, Bioinformatics, 28 (2012), pp. 575–i581.

[232] L. A. RAMSHAW AND M. P. MARCUS, *Text chunking using transformation-based learning*, in Natural language processing using very large corpora, Springer, 1999, pp. 157–176.

[233] M. RANI, A. K. DHAR, AND O. VYAS, *Semi-automatic terminology ontology learning based on topic modeling*, Engineering Applications of Artificial Intelligence, 63 (2017), pp. 108–125.

[234] K. RAYNER AND S. A. DUFFY, *Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity*, Memory & cognition, 14 (1986), pp. 191–201.

[235] J. REICH AND J. A. RUIPÉREZ-VALIENTE, *The mooc pivot*, Science, 363 (2019), pp. 130–131.

[236] D. REIDSMA AND J. CARLETTA, *Reliability measurement without limits*, Computational Linguistics, 34 (2008), pp. 319–326.

[237] D. REIDSMA, D. HOFS, AND N. JOVANOVIC, *Designing focused and efficient annotation tools*, in Measuring Behaviour, 5th International Conference on Methods and Techniques in Behavioral Research, 2005, p. 4.

[238] C. M. REIGELUTH, M. D. MERRILL, AND C. V. BUNDERSON, *The structure of subject matter content and its instructional design implications*, Instructional science, 7 (1978), pp. 107–126.

[239] M. ROSVALL, D. AXELSSON, AND C. T. BERGSTROM, *The map equation*, The European Physical Journal Special Topics, 178 (2009), pp. 13–23.

[240] S. ROY, M. MADHYASTHA, S. LAWRENCE, AND V. RAJAN, *Inferring concept prerequisite relations from online educational resources*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 9589–9594.

[241] M. A. RUIZ-PRIMO, *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 (2000), pp. 29–53.

[242] J. C. SAGER, *Practical course in terminology processing*, John Benjamins Publishing, 1990.

[243] J. SAURO AND J. R. LEWIS, *Quantifying the user experience: Practical statistics for user research*, Morgan Kaufmann, 2016.

[244] M. SAYYADIHARIKANDEH, J. GORDON, J.-L. AMBITE, AND K. LERMAN, *Finding prerequisite relations using the wikipedia clickstream*, in Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 1240–1247.

[245] P. SAZEDJ AND H. S. PINTO, *Time to evaluate: Targeting annotation tools*, Proc. of Knowledge Markup and Semantic Annotation at ISWC, 2005 (2005).

[246] R. SCHEINES, E. SILVER, AND I. M. GOLDIN, *Discovering prerequisite relationships among knowledge components.*, in EDM, 2014, pp. 355–356.

[247] W. A. SCOTT, *Reliability of content analysis: The case of nominal scale coding*, Public opinion quarterly, (1955), pp. 321–325.

[248] A. SETZER, R. GAIZAUSKAS, AND M. HEPPLE, *The role of inference in the temporal annotation and analysis of text*, Language Resources and Evaluation, 39 (2005), pp. 243–265.

[249] A. SHEN, J. QI, AND T. BALDWIN, *A hybrid model for quality assessment of wikipedia articles*, in Proceedings of the Australasian Language Technology Association Workshop 2017, 2017, pp. 43–52.

[250] W. SHEN, J. WANG, AND J. HAN, *Entity linking with a knowledge base: Issues, techniques, and solutions*, IEEE Transactions on Knowledge and Data Engineering, 27 (2014), pp. 443–460.

[251] D. SHI, T. WANG, H. XING, AND H. XU, *A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning*, Knowledge-Based Systems, 195 (2020), p. 105618.

[252] S. SIEGEL, *Nonparametric statistics for the behavioral sciences.*, (1956).

[253] J. SINCLAIR, *Corpus and text-basic principles in developing linguistic corpora: a guide to good practice, ed. m. wynne*, 2005.

[254] R. SNOW, B. O'CONNOR, D. JURAFSKY, AND A. Y. NG, *Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks*, in Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, 2008, pp. 254–263.

[255] S. SOSNOVSKY, P. BRUSILOVSKY, R. AGRAWA, R. G. BARANIUK, AND A. S. LAN, *itextbooks 2019*, in Proceedings of the First Workshop on Intelligent Textbooks on Intelligent Textbooks co-located with 20th International Conference on Artificial Intelligence in Education (AIED 2019), 2019.

[256] S. SOSNOVSKY, P. BRUSILOVSKY, R. G. BARANIUK, AND A. S. LAN, *itextbooks 2020*, in Proceedings of the Second Workshop on Intelligent Textbooks on Intelligent Textbooks co-located with 21th International Conference on Artificial Intelligence in Education (AIED 2020), 2020.

[257] S. SOSNOVSKY, P. BRUSILOVSKY, AND M. YUDELSON, *Supporting adaptive hypermedia authors with automated content indexing*, in Proceedings of Second International Workshop on Authoring of Adaptive and Adaptable Educational Hypermedia at the Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2004), Eindhoven, the Netherlands, 2004.

[258] P. STENETORP, S. PYYSALO, G. TOPIĆ, T. OHTA, S. ANANIADOU, AND J. TSUJII, *Brat: a web-based tool for nlp-assisted text annotation*, in Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2012, pp. 102–107.

[259] M. STRAKA, J. HAJIC, AND J. STRAKOVÁ, *Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing*, in Proceedings of the tenth international conference on language resources and evaluation (LREC 2016), 2016, pp. 4290–4297.

[260] I. SUBASIC AND B. BERENDT, *From bursty patterns to bursty facts: The effectiveness of temporal text mining for news*, in Proceedings of ECAI 2010: 19th European Conference on Artificial Intelligence, 2010, pp. 517–522.

[261] S. SURESU AND M. ELAMPARITHI, *Probabilistic relational concept extraction in ontology learning*, Int. J. Inform. Technol, 2 (2016).

[262] A. P. SWEET AND C. E. SNOW, *Rethinking reading comprehension*, Guilford Press, 2003.

[263] Y. TAKAHASHI, T. UTSURO, M. YOSHIOKA, N. KANDO, T. FUKUHARA, H. NAKAGAWA, AND Y. KIYOTA, *Applying a burst model to detect bursty topics in a topic model*, in International Conference on NLP, Springer, 2012, pp. 239–249.

[264] P. P. TALUKDAR AND W. W. COHEN, *Crowdsourced comprehension: predicting prerequisite structure in wikipedia*, in Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics, 2012, pp. 307–315.

[265] L. TANABE, N. XIE, L. H. THOM, W. MATTEN, AND W. J. WILBUR, *Genetag: a tagged corpus for gene/protein named entity recognition*, BMC bioinformatics, 6 (2005), pp. 1–7.

[266] R. D. TENNYSON AND O.-C. PARK, *The teaching of concepts: A review of instructional design research literature*, Review of Educational Research, 50 (1980), pp. 55–70.

[267] H. THOMPSON, *Hyperlink semantics for standoff markup of read-only documents*, Proceedings of SGML97, (1997).

[268] A. TVERSKY AND D. KAHNEMAN, *Judgment under uncertainty: Heuristics and biases*, Science, 185 (1974), pp. 1124–1131.

[269] N. UZZAMAN AND J. ALLEN, *Temporal evaluation*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 351–356.

[270] S. VAJJALA AND D. MEURERS, *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, 2014, pp. 288–297.

[271] H. VAN HALTEREN, *The detection of inconsistency in manually tagged text*, in Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora, 2000, pp. 48–55.

[272] C. VAN HEE, E. LEFEVER, AND V. HOSTE, *Semeval-2018 task 3: Irony detection in english tweets*, in Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 39–50.

[273] P. VELARDI, S. FARALLI, AND R. NAVIGLI, *Ontolearn reloaded: A graph-based algorithm for taxonomy induction*, Computational Linguistics, 39 (2013), pp. 665–707.

[274] M. VERHAGEN, R. GAIZAUSKAS, F. SCHILDER, M. HEPPLE, G. KATZ, AND J. PUSTEJOVSKY, *Semeval-2007 task 15: Tempeval temporal relation identification*, in Proceedings of the 4th international workshop on semantic evaluations, Association for Computational Linguistics, 2007, pp. 75–80.

[275] Y. VERSLEY, *Disagreement dissected: Vagueness as a source of ambiguity in nominal (co-) reference*, in Ambiguity in Anaphora Workshop Proceedings, 2006, pp. 83–89.

[276] A. VUONG, T. NIXON, AND B. TOWLE, *A method for finding prerequisites within a curriculum*, in Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, July 6-8, 2011, 2011, pp. 211–216.

[277] E. VYLOMOVA, E. M. PONTI, E. GROSSMAN, A. D. MCCARTHY, Y. BERZAK, H. DUBOSSARSKY, I. VULIĆ, R. REICHART, A. KORHONEN, AND R. COTTERELL, *Proceedings of the second workshop on computational research in linguistic typology*, in Proceedings of the Second Workshop on Computational Research in Linguistic Typology, 2020.

[278] H. WAN AND J. B. BECK, *Considering the influence of prerequisite performance on wheel spinning.*, International Educational Data Mining Society, (2015).

[279] A. WANG, A. SINGH, J. MICHAEL, F. HILL, O. LEVY, AND S. R. BOWMAN, *Glue: A multi-task benchmark and analysis platform for natural language understanding*, in 7th International Conference on Learning Representations, ICLR 2019, 2019.

[280] M. WANG, H. CHAU, K. THAKER, P. BRUSILOVSKY, AND D. HE, *Concept annotation for intelligent textbooks*, arXiv preprint arXiv:2005.11422, (2020).

[281] S. WANG, C. LIANG, Z. WU, K. WILLIAMS, B. PURSEL, B. BRAUTIGAM, S. SAUL, H. WILLIAMS, K. BOWEN, AND C. L. GILES, *Concept hierarchy extraction from textbooks*, in Proceedings of the 2015 ACM Symposium on Document Engineering, ACM, 2015, pp. 147–156.

[282] S. WANG AND L. LIU, *Prerequisite concept maps extraction for automatic assessment*, in Proceedings of the 25th International Conference Companion on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 519–521.

[283] S. WANG, A. ORORBIA, Z. WU, K. WILLIAMS, C. LIANG, B. PURSEL, AND C. L. GILES, *Using prerequisites to extract concept maps from textbooks*, in Proceedings of the 25th acm international on conference on information and knowledge management, ACM, 2016, pp. 317–326.

[284] C. WESTON, T. GANDELL, J. BEAUCHAMP, L. MCALPINE, C. WISEMAN, AND C. BEAUCHAMP, *Analyzing interview data: The development and evolution of a coding system*, Qualitative sociology, 24 (2001), pp. 381–400.

[285] J. WIEBE, T. WILSON, AND C. CARDIE, *Annotating expressions of opinions and emotions in language*, Language resources and evaluation, 39 (2005), pp. 165–210.

[286] L. WISSLER, M. ALMASHRAEE, D. M. DÍAZ, AND A. PASCHKE, *The gold standard in corpus annotation.*, in IEEE GSC, 2014.

[287] B. T. WONG AND S. Y. LEE, *Annotating legitimate disagreement in corpus construction*, in Proceedings of the 11th Workshop on Asian Language Resources, 2013, pp. 51–57.

[288] X. XIONG, S. ZHAO, E. G. VAN INWEGEN, AND J. E. BECK, *Going deeper with deep knowledge tracing.*, International Educational Data Mining Society, (2016).

[289] Y. YANG, H. LIU, J. CARBONELL, AND W. MA, *Concept graph learning from educational data*, in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, ACM, 2015, pp. 159–168.

[290] S. M. YIMAM, I. GUREVYCH, R. E. DE CASTILHO, AND C. BIEMANN, *Webanno: A flexible, web-based and visually supported system for distributed annotations*, in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2013, pp. 1–6.

[291] W. C. YOON, S. LEE, AND S. LEE, *Burst analysis of text document for automatic concept map creation*, in International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, 2014, pp. 407–416.

[292] M. V. YUDELSON, K. R. KOEDINGER, AND G. J. GORDON, *Individualized bayesian knowledge tracing models*, in International conference on artificial intelligence in education, Springer, 2013, pp. 171–180.

[293] O. ZAWACKI-RICHTER, V. I. MARÍN, M. BOND, AND F. GOUVERNEUR, *Systematic review of research on artificial intelligence applications in higher education–where are the educators?*, International Journal of Educational Technology in Higher Education, 16 (2019), pp. 1–27.

[294] D. ZEMAN, *Reusable tagset conversion using tagset drivers*, in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 2008.

[295] D. ZEMAN, J. HAJIČ, M. POPEL, M. POTTHAST, M. STRAKA, F. GINTER, J. NIVRE, AND S. PETROV, *CoNLL 2018 shared task: Multilingual Parsing from Raw Text to Universal Dependencies*, in Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium, Oct. 2018, Association for Computational Linguistics, pp. 1–21.

[296] D. ZEMAN, J. NIVRE, AND ET AL., *Universal Dependencies 2.7*, 2020.
LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[297] Z. ZHANG, S. CHAPMAN, AND F. CIRAVEGNA, *A methodology towards effective and efficient manual document annotation: addressing annotator discrepancy and annotation quality*, in International Conference on Knowledge Engineering and Knowledge Management, Springer, 2010, pp. 301–315.

[298] G. ZHAO AND X. ZHANG, *Domain-specific ontology concept extraction and hierarchy extension*, in Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval, ACM, 2018, pp. 60–64.

[299] Z. ZHAO, Y. YANG, C. LI, AND L. NIE, *Guessuneed: Recommending courses via neural attention network and course prerequisite relation embeddings*, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16 (2020), pp. 1–17.

[300] L. ZHOU, *Ontology learning: state of the art and open issues*, Information Technology and Management, 8 (2007), pp. 241–252.

[301] Y. ZHOU AND K. XIAO, *Extracting prerequisite relations among concepts in wikipedia*, in 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[302] H. ZINSMEISTER, A. WITT, S. KÜBLER, AND E. HINRICHS, *Linguistically annotated corpora : Quality assurance, reusability and sustainability*, Publ. in: Corpus Linguistics. An International Handbook. Bd. 1 (Reihe Handbücher zur Sprach- und Kommunikationswissenschaft) / Anke Lüdeling und Merja Kytö (eds.). Berlin: Mouton de Gruyter, 2008, pp. 759-776, (2009).

[303] A. ZOUAQ, R. NKAMBOU, AND C. FRASSON, *An integrated approach for automatic aggregation of learning knowledge objects*, Interdisciplinary Journal of E-Learning and Learning Objects, 3 (2007), pp. 135–162.