TECHNISCHE
UNIVERSITÄT
DARMSTADT

AUTOMATIC QUESTION GENERATION *to* SUPPORT READING COMPREHENSION *of* LEARNERS

Content Selection, Neural Question Generation, and Educational Evaluation

Vom Fachbereich Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation

von

TIM STEUER, M.SC.

Vorsitz: Prof. Dr.-Ing. Tran Quoc Khanh
Referent: Prof. Dr.-Ing. Ralf Steinmetz
Korreferent: Prof. Dr.-Ing. Ulrik Schroeder

Tag der Einreichung: 19. September 2022
Tag der Disputation: 16. Dezember 2022

Darmstadt 2022

Tim Steuer, M.Sc.
Automatic Question Generation to Support Reading Comprehension of Learners
- Content Selection, Neural Question Generation, and Educational Evaluation
Technische Universität Darmstadt, Darmstadt

Jahr der Veröffentlichung der Dissertation auf TUprints: 2023
Tag der mündlichen Prüfung: 16. Dezember 2022

ABSTRACT

Simply reading texts passively without actively engaging with their content is suboptimal for text comprehension since learners may miss crucial concepts or misunderstand essential ideas. In contrast, engaging learners actively by asking questions fosters text comprehension. However, educational resources frequently lack questions. Textbooks often contain only a few at the end of a chapter, and informal learning resources such as Wikipedia lack them entirely. Thus, in this thesis, we study to what extent questions about educational science texts can be automatically generated, tackling two research questions. The first question concerns selecting learning-relevant passages to guide the generation process. The second question investigates the generated questions' potential effects and applicability in reading comprehension scenarios.

Our first contribution improves the understanding of neural question generation's quality in education. We find that the generators' high linguistic quality transfers to educational texts but that they require guidance by educational content selection. In consequence, we study multiple educational context and answer selection mechanisms.

In our second contribution, we propose novel context selection approaches which target question-worthy sentences in texts. In contrast to previous works, our context selectors are guided by educational theory. The proposed methods perform competitive to related work while operating with educationally motivated decision criteria that are easier to understand for educational experts.

The third contribution addresses answer selection methods to guide neural question generation with expected answers. Our experiments highlight the need for educational corpora for the task. Models trained on noneducational corpora do not transfer well to the educational domain. Given this discrepancy, we propose a novel corpus construction approach. It automatically derives educational answer selection corpora from textbooks. We verify the approach's usefulness by showing that neural models trained on the constructed corpora learn to detect learning-relevant concepts.

In our last contribution, we use the insights from the previous experiments to design, implement, and evaluate an automatic question generator for educational use. We evaluate the proposed generator intrinsically with an expert annotation study and extrinsically with an empirical reading comprehension study. The two evaluation scenarios provide a nuanced view of the generated questions' strengths and weaknesses. Expert annotations attribute an educational value to roughly 60% of the questions but also reveal various ways in which the questions still fall short of the quality experts desire. Furthermore, the reader-based evaluation indicates that the proposed educational question generator increases learning outcomes compared to a no-question control group.

In summary, the results of the thesis improve the understanding of the content selection tasks in educational question generation and provide evidence that it can improve reading comprehension. As such, the proposed approaches are promising tools for authors and learners to promote active reading and thus foster text comprehension.

KURZFASSUNG

Alleiniges Durchlesen eines Textes ohne aktive Auseinandersetzung mit dessen Inhalt führt oft zu mangelndem Textverständnis, denn Lernende überlesen häufig Kernkonzepte oder missverstehen wesentliche Ideen. Um das Verständnis eines Textes zu verbessern, sind Fragen hilfreich. Allerdings mangelt es vielen Texten an Fragen: Lehrbücher enthalten oft nur wenige Fragen und informelle Lernressourcen wie Wikipedia enthalten meist überhaupt keine.

In der nachfolgenden Arbeit wird deshalb erforscht, inwieweit Fragen zu Fachtexten automatisch generiert werden können und es werden zwei Forschungsfragen betrachtet: In der ersten Forschungsfrage wird die automatische Auswahl lernrelevanter Inhalte zur Steuerung des Fragegenerierungsprozesses untersucht. Die zweite Forschungsfrage befasst sich mit dem Einsatz der generierten Fragen in Lernszenarien.

Im Rahmen der Forschungsfragen wird zunächst die linguistische Güte neuronaler Fragegeneratoren in der Bildung analysiert. Bei der Analyse ergibt sich, dass die hohe sprachliche Qualität der generierten Fragen auch auf Fachtexten besteht. Jedoch werden durch übliche Fragegeneratoren viele nicht-lernrelevante Fragen erzeugt.

Infolgedessen wird die automatische Auswahl lernrelevanter Sätze und Antworten zur Verbesserung der Fragegenerierung untersucht. Es wird eine neuartige, pädagogisch motivierte Kontextselektion zur Auswahl lernrelevanter Sätze konzipiert, implementiert und evaluiert. Der vorgeschlagene Ansatz erzielt eine vergleichbare Auswahlgenauigkeit wie der Stand der Forschung und ist durch seine pädagogisch motivierten Entscheidungskriterien für Anwender leichter verständlich.

Weiterhin wird in der Dissertation die Antwortselektion untersucht, um den Fragegenerierungsprozess durch erwartete Antworten zu verbessern. Die Experimente verdeutlichen die Relevanz der Korpora-Auswahl: Modelle, die auf allgemeinen Antwortselektionskorpora trainiert werden, selektieren oftmals Antworten, die nicht zwangsläufig lernrelevant sind. Infolgedessen wird ein Ansatz zur automatischen Konstruktion von Korpora mithilfe lernrelevanter Konzepte aus Lehrbüchern vorgeschlagen. Die Nützlichkeit des Ansatzes zeigt sich dadurch, dass neuronale Modelle, die auf den konstruierten Korpora trainiert wurden, lernrelevante Konzepte in Texten erkennen.

Basierend auf den vorangegangenen Experimenten wird ein automatischer Fragegenerator zur Verbesserung des Leseverständnisses konzipiert, implementiert und evaluiert. Dieser neuartige Ansatz wird mit einer intrinsischen Evaluation mit Bildungsexperten und einer extrinsischen Evaluation mit einer empirischen Studie zum Leseverständnis evaluiert. Beide Evaluationen bieten einen differenzierten Blick auf die Stärken und Schwächen des Fragegenerators. Experten bewerten ca. 60% der generierten Fragen als hilfreich, auch wenn manche Fragen in verschiedener Hinsicht hinter der gewünschten Qualität zurückblieben. Weiterhin erhöhen die Fragen des vorgeschlagenen Ansatzes den Lernerfolg im Vergleich zu einer Kontrollgruppe.

Zusammenfassend tragen die Ergebnisse dieser Dissertation zu einem tiefergehenden Verständnis von automatischer Inhaltsauswahl für Fragegeneratoren im Bildungsbereich bei und liefern dazu neuartige Ansätze zur Kontextselektion, Antwortselektion und zur lernrelevanten Fragegenerierung. In einer durchgeführten Studie verbesserte der vorgeschlagene Fragegenerierungsprozess das Leseverständnis. Somit sind die vorgeschlagenen Ansätze zur Fragegenerierung wahrscheinlich ein nützliches Hilfsmittel für Autoren und Lernende, um aktives Lesen und somit das Leseverständnis zu fördern.

CONTENTS

Previous Publications 1

1 introduction 5
- 1.1 Overview of the State-of-the-Art 6
- 1.2 Research Challenges 7
- 1.3 Research Questions 8
- 1.4 Approach & Contributions 9
- 1.5 Structure of the Thesis 11

2 background and related work 13
- 2.1 Terminology 13
- 2.2 Automatic Question Generation Preliminaries 13
- 2.3 State-of-the-Art 20
- 2.4 Summary and Identified Research Gap 27

3 overall concept and approach 29
- 3.1 Effects of Manually Authored Literal Questions 29
- 3.2 Conceptual Content Selection Considerations 31
- 3.3 Conceptual Question Realization Considerations 34
- 3.4 Approach 35

4 piloting linguistic generation quality in education 39
- 4.1 Methodology 39
- 4.2 Datasets 40
- 4.3 Question Realization Models 42
- 4.4 Content Selection Conditions 43
- 4.5 Automatic Evaluation Results 45
- 4.6 Annotation Study Design 48
- 4.7 Annotation Study Results 49
- 4.8 Discussion of the Main Insights 52

5 educationally motivated context selection 55
- 5.1 Definition & Causal Sentence Selection 56
- 5.2 Empirical Evaluation Study 59
- 5.3 Results 61
- 5.4 Validity Considerations 66
- 5.5 Discussion of the Main Insights 66

6 TRANSFERRING EDUCATIONAL ANSWER SELECTION FROM NONEDUCATIONAL TEXTS 69
6.1 Implicitly Learnt Answer Selection Transfer 70
6.2 Evaluation Setup 71
6.3 Datasets 72
6.4 Answer Selection Models 74
6.5 Results 75
6.6 Validity Considerations 81
6.7 Discussion of the Main Insights 82

7 ANSWER SELECTION FROM AUTOMATICALLY CONSTRUCTED TEXTBOOK CORPORA 85
7.1 Automatic Index Extraction 86
7.2 Automatic Corpus Construction 88
7.3 Answer Selection Model Comparision 89
7.4 Evaluation Results 90
7.5 Validity Considerations 92
7.6 Discussion of the Main Insights 93

8 QUESTION GENERATION IN READING COMPREHENSION SCENARIOS 97
8.1 The Educational Automatic Question Generator 97
8.2 Evaluation Studies 103
8.3 Intrinsic Evaluation Study 104
8.4 Extrinsic Evaluation Study 113
8.5 Discussion of the Main Insights 124

9 SUMMARY, CONCLUSIONS, AND OUTLOOK 129
9.1 Summary of the Thesis 129
9.2 Conclusions 131
9.3 Outlook 132

BIBLIOGRAPHY 133

A APPENDIX 151
A.1 RACE Plots of the Pilot Study 151
A.2 Cause and Definition Annotation Schema 152
A.3 Overview of All Model Hyperparameters 156
A.4 List of Acronyms 157
A.5 Supervised Student Theses 159

B AUTHOR'S PUBLICATIONS 161

C ERKLÄRUNGEN LAUT PROMOTIONSORDNUNG 165

PREVIOUS PUBLICATIONS

This chapter details the author's previous publications in the field, i.e., conference and journal articles included in this thesis. Table 1 lists the relevant previous publications, none of which were directly reprinted in this thesis, except for tables and figures, particularly for evaluation data. Each source was made explicit by providing its reference in the corresponding caption. A full list of the author's publications can be found in Appendix B.

Investigating educational technologies is inherently interdisciplinary and requires experts from multiple fields to collaborate. During this dissertation work, I experienced science as a collaborative effort, and all results achieved in the thesis were derived from scientific teamwork. Consequently, I disclose the contributions of all relevant collaborators and co-authors with their respective affiliations. Whenever no dedicated affiliation is provided, the person is or has been a colleague at the Multimedia Communications Lab of the Technical University of Darmstadt. In the following chapters, I will use the pronoun *we* to acknowledge the collaborative team effort.

Chapter	Publications
2	[150], [153], [155], [156]
3	[150], [153], [155], [156]
4	[153]
5	[149, 156]
6	[154]
7	[151]
8	[150], [155]

Table 1: Author's publications in relation to the thesis chapters.

Chapter 2 provides an in-depth overview of the related work in Automatic Question Generation and analyzes the state-of-the-art to derive the relevant research gaps. While working on previously publications, I conducted multiple literature reviews for the respective related work sections. Thus, much of the related work analysis has already been published, with the main analysis conducted in [150], [153], [155], [156]. The corresponding related work chapter of the thesis is a revised, reorganized, and restructured essence of these sections. I was supported in conducting the respective literature reviews by Anna Filighera, Nina Mouhammad, Gianluca Zimmer, and Christoph Rensing, with whom I shared and discussed relevant papers and research gaps. The idea of investigating educational Automatic Question Generation using neural networks is a consequence of frequent discussions with Anna Filighera and Christoph Rensing about the recent developments in educational technology and natural language processing.

Chapter 3 analyzes the linguistic and educational concepts relevant to an educational Automatic Question Generation system. Two collaborative circles contributed to the analysis. First, the regular meetings with Christoph Rensing and Anna Filighera were constructive and supported my analysis efforts. We mainly discussed technologies and their linguistic evaluation. Many of the initial ideas seen in the chapter can be traced back to these discussions. Second, during the dissertation, I led a Software Campus project with the Holtzbrinck Publishing Group, in which I discussed potential design ideas with two linguistical research assistants (Friederike Lenke and Kris Schilling) and two psychological research assistants (Monja Huhle and Liza Mekschrat). The group discussions facilitated my understanding of key ideas relevant to the thesis. After providing Monja Huhle and Liza Mekschrat with relevant technological and psychological literature, they suggested valuable additional literature concerning text understanding and textual learning relevance features.

The initial study investigating the linguistic capabilities of Automatic Question Generation models on educational data in Chapter 4 was conducted together with Anna Filighera and Christoph Rensing [150]. Anna Filighera provided valuable technical assistance and supported the dataset and model selection steps. Christoph Rensing provided feedback on the conceptual level. He gave feedback on the soundness of my research questions and operationalization approaches. Besides, he supported me while authoring the initially published version of the material by proposing helpful manuscript revisions. Furthermore, the discussion with both fostered our shared understanding of the results' scope and their potential interpretation.

The unsupervised baseline models in Chapter 5 were initially explored in a master thesis proposed by me to Tianyang Zhou. Based on the results and the corresponding related work, I hypothesized that the unsupervised approaches have shortcomings in finding learning-relevant contexts. After analyzing the linguistic and educational requirements, I proposed using semantic classification for sentence selection. Gianluca Zimmer and I explored an early version of such a system based on causal sentences in his bachelor thesis, with promising results. Given these promising results, Gianluca Zimmer, Anna Filighera, Thomas Tregel, and I developed the idea further in Steuer et al. [156]. Anna Filighera and Thomas Tregel supported the paper's conceptual level, whereas Gianluca Zimmer supported us during the technical challenges.

In Chapter 6, we investigate the transfer of machine learning-based answer selection models trained on noneducational data to the educational domain. I developed the idea for the master thesis of Lin Li, where we tried to transfer models from the noneducational Stanford Question Answering Dataset (SQuAD) to an educational domain. For the thesis evaluation, I, therefore, supervised the dataset collection of the educational Textbook Question Answering with Answer Spans (TQA-A) dataset, which Nina Mouhammad and Kris Schilling annotated. Lin Li's initial results hinted that transferring SQuAD-trained models to TQA-A does not provide much benefit. I followed up the thesis with a larger-scale extensive evaluation of model transferability published in Steuer, Filighera, and Tregel [154]. Anna Filighera and Thomas Tregel supported the research design. We jointly developed the operationalization for the dataset and model comparisons.

In Chapter 7, an answer selection strategy for extracting learning-relevant concepts using automatically constructed corpora is proposed. The technical prerequisite of extracting the concepts from back-of-the-book PDF indices was initially explored in a student lab I had proposed to Ercan Akar and Julian Barthel. I rewrote and extended the software after the lab, resulting in the final version of the pdf-index-extractor. Nina Mouhammad, Gianluca Zimmer, and I collaborated on the research challenge while working on my Software Campus project. I conducted the research study and experiments, whereas Nina Mouhammad provided evaluation procedures and Gianluca Zimmer computed dataset statistics under my supervision. Anna Filighera and Thomas Tregel supported the research efforts in our weekly research meetings with iterative feedback. Anna Filighera, Thomas Tregel, and I discussed and confirmed the plausibility of the dataset construction procedures and result metrics. It resulted in helpful suggestions such as including the unseen concepts evaluation split. The result of these efforts was published in Steuer et al. [151].

We propose an educational Automatic Question Generation system in Chapter 8 and assess the generated questions' quality in an annotation study and a reading comprehension study. The final system design was a consequence of research meetings at the Multimedia Communications Lab. In weekly research meetings, Tobias Meuser, Anna Filighera, and Christoph Rensing supported my design efforts. I proposed the initial design and refined it based on the multifaceted feedback given by my colleagues. For instance, we considered different model architectures for the realization of the context selection and question realization components. After I had realized the design in a technical prototype, the intrinsic question quality study was conducted in collaboration with Anna Filighera and Christoph Rensing [150]. The extrinsic study was conducted in collaboration with Anna Filighera, Thomas Tregel, and André Miede from the Saarland University of Applied Sciences [155]. We collaborated on the study's design and the data analysis. I assembled the study's initial reading materials and test items and wrote the first draft of the manuscript. Anna Filighera, André Miede, and Thomas Tregel analyzed and revised the initial selection of the study's material. André Miede and I orchestrated the study's data collection.

1 INTRODUCTION

Reading is an ancient and precious cultural technique omnipresent in education. Even if subject-matter experts are absent, reading allows us to follow their thoughts and learn from them. Hence, reading fosters learning in various stages in our educational careers, ranging from elementary to graduate school and from basic science understanding to advanced quantum physics. Consequently, educational careers are greatly affected by learners' reading comprehension and their retention of the material.

However, passively reading educational texts without actively engaging with their content, is not an optimal learning strategy [4, 9, 37, 73]. Reading comprehension theories provide ample evidence that readers shift their attentional focus [103], have limited mental memory capacity [103], and often fail to construct a coherent mental representation [9]. Most people have experienced this themselves, reading a few text pages only to realize they do not remember what was stated on the first page. Moreover, even when readers form a coherent mental representation of the text's content, recalling this information later is not guaranteed because good retention benefits from active recall training [39]. Thus, texts are best studied actively instead of passively to improve learning outcomes [4, 37, 133].

One well-established active reading method is questioning learners about the texts they read [4, 13, 51, 133]. Questions after reading increase learning outcomes [4, 13, 51] and improve delayed retention compared to no-question control groups [39]. Interestingly these benefits are caused by transfer and inference questions [48] but also by literal questions [4, 39, 51, 73], inquiring about facts directly stated in the text. Both transfer and literal questions can significantly improve learners' text comprehension.

Nevertheless, many textual learning resources in higher education still lack questions. Among them are textbooks, which often only have a few deep comprehension questions at the end of a chapter, and online resources such as Wikipedia, which lack questions entirely. This undersupply of questions is unsurprising, as authoring relevant questions is time-consuming and requires educational experience. Yet, time, educational experience, and sometimes even the willingness to author questions are usually scarce among authors. They often work on a royalty basis or voluntarily and are subject-matter experts but are not necessarily trained in educational issues. Hence, they may not be aware of the benefits provided by questioning readers.

In consequence, this work investigates educational Automatic Question Generation (AQG) to support learners and resource authors. The goal of educational AQG is to automatically generate meaningful questions about reading materials. Authors may profit from generated question recommendations during drafting manuscripts, offering them a quick way to add questions to their writing. Learners may profit from generated questions because the questions transform their passive reading session into an active reading session, improving text comprehension and prolonged retention.

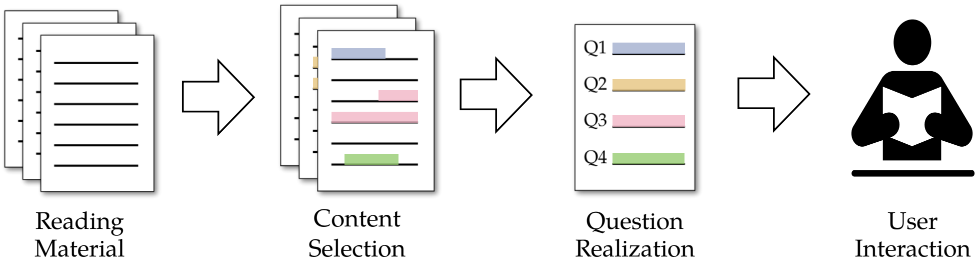


Figure 1: An illustration of educational Automatic Question Generation investigated in this work. It encompasses selecting learning-relevant content from educational reading material such as textbook chapters and transforming the selected information into question-form (Question Realization) to foster learners' text comprehension.

1.1 OVERVIEW OF THE STATE-OF-THE-ART

In educational AQG, we aim to generate questions about learning-relevant text contents to foster comprehension (see Figure 1). The task usually comprises a content selection step, extracting learning-relevant content from the text, and a question realization step, transforming the selected information into question form. The state-of-the-art in question realization stems mainly from research outside educational application, whereas the content selection has specifically been researched for education.

Question realization comprises three main lines of research: rule-based, template-based, and neural network-based approaches. First, rule-based approaches receive a declarative sentence as input and apply syntactic rules to transform it into question-form [56, 61]. However, they fail for more complex declarative sentences often encountered in educational texts which comprise multiple pieces of information [55]. Second, predefined question templates are utilized [86, 100, 101]. A question template encompasses variables to be filled by the corresponding AQG algorithm (e.g. *"What is X?"*). Compared to rule-based approaches, templates shift the task focus slightly towards information extraction. Still, syntactic rules or filters are often necessary to simplify the detected information before filling the template variable [101]. Consequently, they generate repetitive questions and do not necessarily generalize to more complex sentences. Third, neural network-based approaches recently outperformed template-based and rule-based approaches regarding the generated questions' linguistic quality for single sentences [36] and small paragraphs [185]. Neurally generated questions comprise fewer grammatical errors, sound more natural, and require more reasoning before answering them [32, 36, 186]. This is a consequence of the neural networks' strong language modeling capability, removing the necessity of handcrafted transformation rules by approximating the conditional word or token distribution of the training data [36]. However, it is currently not well-established to what extent these superior language skills generalize the educational domain and whether they can support learners' text comprehension.

Furthermore, the content selection step is substantial for educational AQG, where we aim to ask learning-relevant and not arbitrary questions. Therefore, content selection mechanisms are required to detect learning-relevant information inside the larger input text [19, 34]. In template-based and rule-based approaches, content selection mainly relies on information extraction techniques like automatic summarization [19] or educational priors [29, 61]. However, there is a lack of research evaluating these content selectors for the linguistically better-performing neural network-based question realization approaches. Thus, whether they provide helpful guidance for neural question realization is uncertain. In contrast, neural network-based approaches primarily rely on either separately or jointly learning of what constitutes relevant information from a training corpus [117, 158]. Yet, these content selection approaches are frequently trained on noneducational text corpora, and it is unclear if the learned selection criteria transfer to education.

Finally, if we seek to support learners with AQG we need to evaluate the educational quality of the generated questions. However, while there has been annotation study-based research outside the educational domain [32, 36, 62, 158, 186], we observe that AQGs potential educational effects on learners are not well understood. Studies evaluate only linguistic aspects of the generated questions [32, 36, 62] or focus on the expert-based judgement of the questions' educational quality [59, 174]. However, the judgement of questions' quality is difficult even for human experts [2]. Research suggests that reader-based, task-specific evaluation of text generation pipelines yields a complementary picture to an expert-based output assessment [165]. Hence, more research exploring how the generated questions affect learning is needed.

1.2 RESEARCH CHALLENGES

We identified the following challenges in the current research landscape of educational AQG.

Research Challenge: *Learning-relevant content selection*

Educational texts are densely written and comprise various sentences. Educational questions need to concentrate on those text aspects essential for comprehension. Thus, selecting learning-relevant, question-worthy sentences is vital for educational AQG. However, automatically identifying the essential text aspects poses serious challenges. Learning relevance is an intuitive concept inherently challenging to operationalize. Moreover, our state-of-the-art analysis revealed that content selection mechanisms are underinvestigated. Some content selection methods have been applied with template-based and rule-based question realization in education [29, 100, 120]. Yet, the overall effect of content selectors in neural network-based question realization is unknown. Furthermore, machine learning-based content selectors are frequently trained on noneducational text corpora. However, it is unclear whether such training leads to valid content selection in educational settings. Finally, a content selector's usefulness is usually estimated through the quality of the generated questions, distorting the evaluation results with errors from the actual question realization algorithm.

Research Challenge: *Educational automatic question generation to foster text comprehension*

Educational questions are not exclusively selected for their linguistic quality. Consequently, generating questions with high linguistic quality for a given input is not enough. Instead, the questions' educational properties must be considered. However, the educational value of neural network-based AQG approaches is not well understood. So far, research suggests their linguistic superiority, yet only few studies investigated the educational aspects of generated questions with mixed results. More evidence is needed to what extent the generated questions can support text comprehension. There are many dimensions to measure questions' educational quality, and designing a meaningful evaluation scenario is challenging [2, 59]. Ideally, new evidence is gathered from the learners' and the authors' perspectives because both groups may profit from educational AQG, but may have different application requirements.

1.3 RESEARCH QUESTIONS

We address the following research questions for the identified research challenges.

Research Question 1: *To what extent can content selectors, based on textual features, extract learning-relevant information from educational science texts in different domains?*

The research question is based on the observation that learning-relevant inputs are a prerequisite for learning-relevant questions. Our research is scoped as follows. We deem information as learning-relevant if it improves text comprehension according to empirical data from learners or educational experts. We believe this results in higher external and ecological validity [33, p. 106] than previous studies, which mainly defined the concept through automatic measures [19, 63] on noneducational corpora [19, 34, 177]. However, it is clear that measuring learning relevance with this approach requires emprical data and a thorough empirical evaluation design.

Scientific texts are particularly challenging in reading comprehension scenarios and are frequently encountered in learners' educational careers [9]. Thus, the research question focuses on educational texts written to convey domain knowledge, for instance, textbook chapters or encyclopedic articles. These texts may stem from different domains such as biology, anatomy or economics. Other writings such as stories, news, or essays are explicitly excluded from the research context. Consequently, the thesis gathers insights into the applicability of AQG to science reading comprehension scenarios to support domain understanding. Other potential reading comprehension scenarios are out of the scope of this thesis.

Moreover, the research question solely relies on text characteristics. Other factors such as the learners' mental state or learning goals will not be considered. This is a consequence of many of the excluded factors relying on the individual reader's mental state [103], vastly complicating their automatic measurement. In contrast, text properties such as the author's style of expression are independent of the individual reader. Nevertheless, text characteristics affect reading comprehension scenarios [27, 70, 104] and learners rely on them for content relevance estimation [27, 90], rationalizing this restriction.

Research Question 2: *To what extent can educational AQGs generate literal questions supporting reading comprehension of educational science texts in different domains?*

The research question addresses the missing evidence of the generated questions' effect on reading comprehension. We explore the question under the following constraints.

We only investigate neural network-based methods for question realization because they consistently outperform rule-based and template-based approaches in noneducational AQG research [36]. Hence, in this work, different neural approaches and their educational AQG suitability are investigated in the context of educational science texts.

We concentrate on literal short-answer questions with answers stated explicitly in the text and do not incorporate external knowledge into the generation process. This limitation inevitably leads to simpler questions, as deeper questions usually combine knowledge sources or incorporate external knowledge [48]. However, it has been repeatedly shown that learners benefit from literal, short-answer questions compared to passive reading [4, 39, 51].

Finally, the generated questions' influence on reading comprehension may be manifold [13, 39, 51]. Measuring every imaginable effect the questions may cause is out of the scope of this thesis. We concentrate our research efforts on dimensions that, we hypothesize, have an imminent influence on the learning outcomes. Other effects, for example, of an emotional or motivational nature, are not examined in this thesis.

1.4 APPROACH & CONTRIBUTIONS

Contributions to Research Question 1: *to improve the understanding of textual features-based content selectors' strengths and weaknesses in reading comprehension scenarios.*

We design, implement and evaluate content selectors that explicitly model educational priors. In contrast to the approaches previously employed, they are not directly trained to detect arbitrary learning-relevant content but rely on a priori educational assumptions for their decisions. We conduct a joint reader-based and system performance study, published in [156], showing that the approach's selections are reliably associated with learners' perception of relevance and yield competitive results for classifying learning-relevant sentences.

Next, we analyse the transfer of noneducational neural network content selectors to education. For that, we conduct a study of the AQG answer selection subtask, published in [154]. We collect a mid-sized educational corpus and compare the approaches' performance characteristics and generalization capabilities on noneducational and educational data. Our contribution provides evidence that the models only possess a limited transfer capability from the noneducational to the educational task. Thereby, the analysis improves the understanding of whether text characteristics from noneducational data help improve educational AQG.

Furthermore, we investigate neural network methods for learning-relevant concept selection based on automatically constructed textbook corpora, published in [151]. The results reveal that neural networks learn to select relevant concepts from these corpora. This could, for instance, enable human-in-the-loop approaches for learning-relevant content selection.

In summary, our contributions improve understanding of content selection approaches in education. We show the benefits of content selectors based on educational priors and highlight the need for educational datasets and evaluation.

Contributions to Research Question 2: *to gather insights into how educational AQG is perceived by experts and how it affects learners in reading comprehension scenarios.*

Neural question generators come in various architectures and models. We study to what extent these approaches transfer to education even though trained on noneducational data, published in [153]. The employed comparison parameters measure both linguistic and selected educational quality characteristics. The results suggest that pre-trained, answer-aware neural generators transfer much of their linguistic quality to out-of-distribution educational data but often ask educationally irrelevant questions.

We then design and implement an educational AQG pipeline based on content selection insights gathered from the first research question and the best neural question generator found in the initial study. The proposed educational AQG approach utilizes a content selection based on educational priors and syntactic patterns and generates questions using an answer-aware pre-trained neural generator. The approach's linguistic and educational qualities are evaluated in an in-depth expert-based annotation study published in [150]. Moreover, we conduct an empirical, learner-based reading comprehension study to investigate to what extent the generated questions affect learners' reading comprehension learning outcomes, published in [155]. The study finds that questions from the proposed educational AQG approach improve reading comprehension compared to a no-question control group.

In summary, we contribute vital insights into how the proposed approaches may be applied for educational AQG. Notably, our results suggest that the questions stemming from the proposed educational AQG approach improved learners' reading comprehension but are also frequently perceived as imperfect by expert evaluators.

1.5 STRUCTURE OF THE THESIS

The thesis structure is summarized in Figure 2, mapping the thesis chapters to the corresponding research questions and AQG tasks.

Chapter 2 introduces the necessary background knowledge and relevant related works before identifying technical research gaps in AQG. Chapter 3 discusses the benefits of literal questions on learning and analyses related conceptual works on learning-relevancy and linguistic quality of questions. It combines this analysis with the technical research gap from the previous chapter to deduce our educational AQG research approaches. Next, in Chapter 4, we describe a piloting study on the linguistic quality of neural question realization approaches in education, providing us with initial evidence regarding the feasibility of these approaches in the educational context (RQ 2). Afterward, Chapter 5 addresses the selection of learning-relevant, question-worthy sentences based on educational criteria (RQ 1), whereas Chapter 6 and Chapter 7 address different methods for selecting relevant answers to guide the question generation process (RQ 1). Finally, Chapter 8 proposes a novel educational AQG approach for textbooks and reports expert evaluation results as well as empirical reading comprehension study results on the questions' quality and their effects on learning outcomes (RQ 2). We conclude by summarizing our core contributions in Chapter 9.

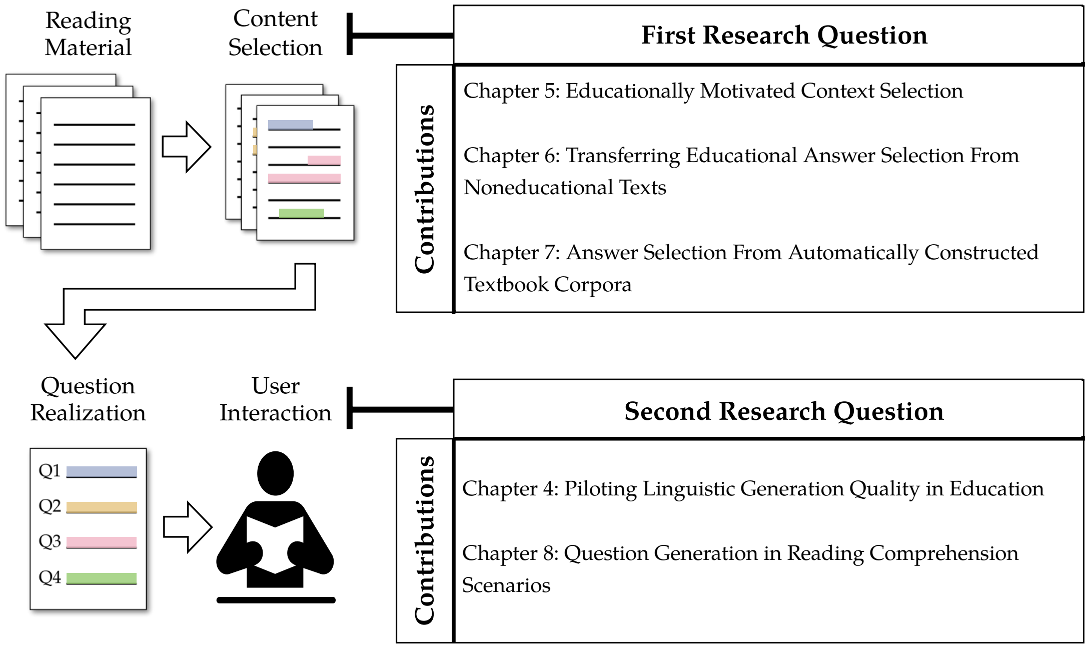

Reading Material
Content Selection
First Research Question
Contributions
Chapter 5: Educationally Motivated Context Selection
Chapter 6: Transferring Educational Answer Selection From Noneducational Texts
Chapter 7: Answer Selection From Automatically Constructed Textbook Corpora
Question Realization
User Interaction
Q1
Q2
Q3
Q4
Second Research Question
Contributions
Chapter 4: Piloting Linguistic Generation Quality in Education
Chapter 8: Question Generation in Reading Comprehension Scenarios

Figure 2: An overview of the research reported in this thesis and its relation to the educational Automatic Question Generation process. The research questions focus on content selection and the potential of educational Automatic Question Generation in reading comprehension scenarios.

2 BACKGROUND AND RELATED WORK

We investigate educational Automatic Question Generation (AQG), and this chapter provides the necessary foundations and an in-depth analysis of the state-of-the-art. We initially define the AQG process and characterize the challenges involved. We describe the three subtasks of educational AQG, their technological foundations, and typical evaluation methodologies. Next, the state-of-the-art for the AQG subtasks and AQG systems used in education is analyzed, and we derive relevant technical research gaps.

2.1 TERMINOLOGY

Initially, we would like to briefly clarify important terminology.

Literal questions are factoid short answer questions. Answering them does not require text external information. Instead, the answer is stated in the text in a single sentence or small paragraph. We defer a discussion of the reasoning behind this question type to Chapter 3. Furthermore, we often refer to *implicit* content selection or learning in the related work discussion and the remainder of the thesis. We use the term *implicit* whenever a system determines what constitutes important content solely from data and without theoretical grounding. That is, no explicit definition of what constitutes relevant information is applied a priori, and no theoretical grounding of why the contents selected by the system are relevant is given. Instead, only the data distribution determines relevancy. We contrast that with explicit or educationally-motivated content selection, where the system authors a priori define what they consider relevant and then build a content selector around these relevancy assumptions. We will see why such approaches might be beneficial when discussing the different content selection approaches in this chapter. Besides, we refer to datasets not explicitly collected in educational scenarios or with educational intent as *noneducational*. This includes datasets that might be used as training data for educational AQG systems in hindsight. However, their initial purpose was noneducational. That is, educational datasets have to be collected in an educational scenario and with educational theories or tasks in mind a priori. Particularly Chapter 6 will highlight why this differentiation is crucial.

2.2 AUTOMATIC QUESTION GENERATION PRELIMINARIES

The educational AQG task can be analyzed as a three-step process composed of context selection, answer selection, and question realization. The first two steps identify relevant information in text, whereas the third step transposes relevant information into question form. The subtasks have similar but unique technological backgrounds and challenges. Moreover, the overall educational AQG task encompasses significant evaluation challenges.

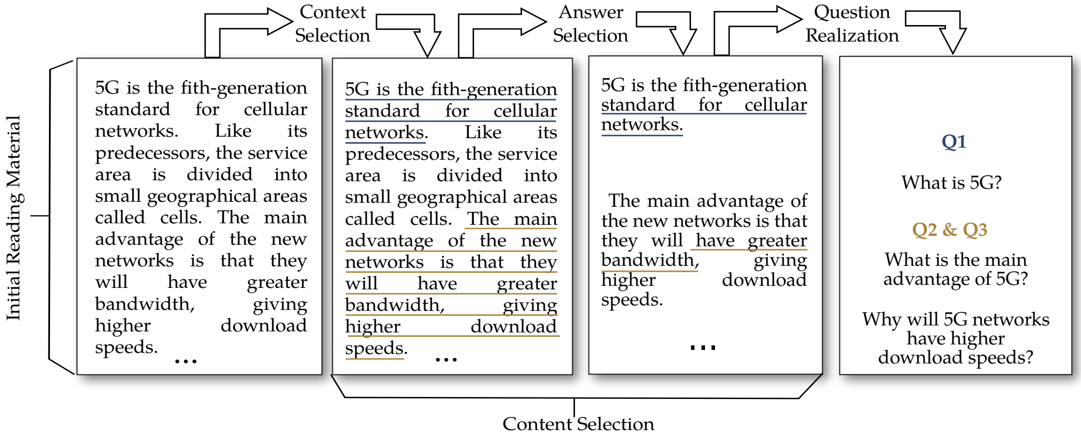

Context Selection
Answer Selection
Question Realization
Initial Reading Material
5G is the fith-generation standard for cellular networks. Like its predecessors, the service area is divided into small geographical areas called cells. The main advantage of the new networks is that they will have greater bandwidth, giving higher download speeds. ...
5G is the fith-generation standard for cellular networks. Like its predecessors, the service area is divided into small geographical areas called cells. The main advantage of the new networks is that they will have greater bandwidth, giving higher download speeds. ...
5G is the fith-generation standard for cellular networks.
The main advantage of the new networks is that they will have greater bandwidth, giving higher download speeds.
...
Q1
What is 5G?
Q2 & Q3
What is the main advantage of 5G?
Why will 5G networks have higher download speeds?
Content Selection

Figure 3: An example of the technical tasks involved in educational AQG. They comprise context selection, answer selection, and question realization. In the example, the blue and yellow sentences are selected during context selection. Afterward, the answer selection marks a single expected answer in every selected sentence. The question realization phase generates the question Q1 for the blue (context, answer) pair as well as Q2 and Q3 for the yellow (context, answer) pair.

2.2.1 *The Automatic Question Generation Process*

We introduced in Chapter 1 that texts used for educational AQG in this thesis are self-contained scientific educational materials. Thus, the AQG process receives reading comprehension material as input and produces one or multiple written short-answer questions concerning the texts as output. From a technical viewpoint, the generation process encompasses Natural Language Understanding (NLU) and Natural Language Generation (NLG) components to select question-relevant content and to transform declarative texts into question form. The technical task may be divided into three subtasks (see Fig. 3):

- *Context Selection* [19, 22, 34]
- *Answer Selection* [22, 63, 177]
- *Question Realization* [32, 36, 123]

When the question realization step does not require an answer as input, the generation process is called *answer-agnostic*. In case an answer is required, it is called *answer-aware*.

The first subtask is context selection. Its goal is to reduce the initial text to question-worthy sentences or small paragraphs, which we call *contexts*. For instance, given a textbook chapter of hundreds of sentences, context selection seeks to select the learning-relevant sentences. That is, questions will target only information in selected contexts, and unselected sentences will not become a question target. The step exists for multiple reasons.

From a conceptual point of view, the generated questions aim to foster learning, and the time learners are willing to spend with questions is limited. Hence, important information must be pre-selected because one cannot ask learners indefinitely many questions. Furthermore, the input text's size must be tractable. Current NLG models often have a maximum processing window of a few hundred tokens [11, 30, 32, 123]. For larger inputs, the text is truncated, and valuable context information is lost. Therefore, the AQG model is usually unable to process whole textbook pages. Besides, approaches exist that generate multiple questions per sentence and keep only the most promising questions according to a ranking metric [56, 84]. This results in the initial generation of multiple questions per sentence. Since learning irrelevant sentences or paragraphs lead to mediocre questions at best, using every sentence as input for those algorithms would waste valuable computational resources. Context selection mitigates this issue.

Second, answer selection is applied. The step accounts for the fact that, given a context, a variety of questions may be asked. The step marks expected answers in the previously selected contexts. In turn, the marking of the expected answer eliminates the generation of some implausible questions. For example, in the sentence: *"Barack Hussein Obama II is an American politician and lawyer who served as the 44th president of the United States."*, we could ask for Obama's profession, his nationality but also for more unlikely things: *"Which type of states was Obama president of?"*; Answer: *"United"*. Such syntactically correct but semantically dubious questions are prevented by answer selection. Moreover, answer selection influences which type of question is generated (e.g. [147]). A Who-question may be most appropriate if the selected answer is a name. However, if the chosen answer comprises a complete causal clause, What-, Why- or How- questions may be more appropriate. Finally, answer selection reduces the subsequent question realization subtask's ambiguity because it brings it closer to a one-to-one mapping task [84].

Finally, the question realization subtask generates a question given the selected context and answer [36]. While the previous subtasks determine the learning-relevancy of the textual information, this subtask determines the questions' textual appearance. Consequently, it considerably affects the questions' characteristics. Depending on the realization algorithm, it paraphrases or copies more or fewer words from the input text segments, influencing the questions' perceived naturalness [36, 62]. Additionally, the subtask decides to what degree the generated questions comprise contextual information from the text, affecting the questions' ambiguity. Consequently, the choice of the question realization algorithm is vital to the desired question characteristics.

As with many conceptual views on technical processes, the described subtasks are neither always sequential nor clearly separated. Common exceptions are answer-agnostic systems skipping the answer selection (e.g. [34]), or systems jointly solving answer and context selection [117]. Besides this, several question generation processes employ further steps such as overgenerating and ranking [56, 84]. They generate numerous questions, rank them, and select only the most promising for end-users. Still, the conceptual view is helpful as it decomposes the task into the respective NLU and NLG challenges and separately describes the different information extraction concerns involved in context and answer selection.

2.2.2 *Token and Sparse Text Representations*

AQG is a natural language processing task and, as such, needs a way of representing the input texts. Texts are composed of tokens like words, punctuation, or numbers. These tokens consist of character symbols. However, modern natural language processing relies heavily on the statistical properties of texts [46]. Consequently, a meaningful numerical representation of the input text is required.

Typical numerical representations vectorize a given text based on words or smaller units. A common vectorization method is the *Bag-Of-Words*. It represents every token in the vocabulary with a vector element, and the element's value represents the occurrence count of a specific token. Another more specialized type of vectorization is the *TF-IDF* format [144]. It is based on the Bag-Of-Words concept but introduces a weighting scheme for every vector element. For that, it weights every token's text frequency (TF) with its frequency in the overall document corpus (IDF). Hence, the more specific a token for the given text is, the higher its vector elements value. In consequence, the resulting vector representation encodes the token's importance to a given document better than the Bag-Of-Words approach. Different weighting schemes exist [96], but they all assign document-specific tokens a higher value.

Both approaches have in common that the vector dimension grows linearly with the size of the vocabulary as each vector element represents a single token. Moreover, vector instances often comprise only comparably few non-zero entries as most texts only use a small part of the available vocabulary. Thus, this kind of representation is called *sparse*.

Word-level token representation schemes fail to represent words outside of their vocabulary [139]. For this reason, character-level and subword-level representations exist. For instance, a character-level Bag-Of-Words is constructed the same way as a word-level Bag-Of-Words, except that it assigns a vector element to every character and not every word. This results in a vocabulary that is much smaller than a word-based vocabulary. Furthermore, there are rarely any out-of-vocabulary tokens. However, character-level representations usually have lower expressiveness because they do not represent words explicitly. Subword-level representations [139, 179] apply a statistically learned splitting scheme to every word in the text, leaving it as is or splitting it into multiple subwords. For instance, the term *"unintended"* may be divided into the three tokens *"un ##intend ##ed"*. In many cases, the final subword set has higher representational power than a character-level encoding and does not suffer from out-of-vocabulary terms [139]. It is, therefore, a good compromise between character- and word-level representations.

2.2.3 *Dense Text Representations and Embeddings*

Neural networks employed in AQG are frequently based on dense vector representations called *embeddings*, popularized with the advent of neural networks for NLU and NLG [46]. Embeddings represent every token with a distinct n-dimensional vector. The construction algorithms aim to represent similar tokens with similar vectors [106]. That

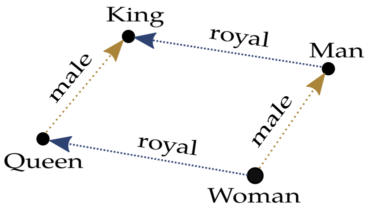


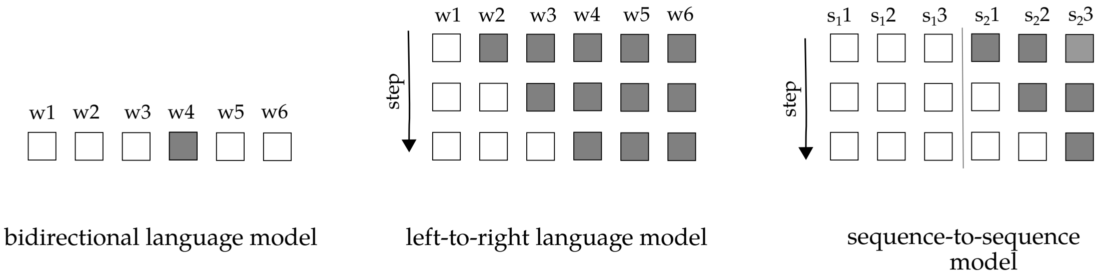
Figure 4: The canonical example of the properties of well-adjusted embedding vectors [88]. Note that this illustration is idealized, and the actual embedding vector addition is noisy and does not perfectly capture semantic relations.

is, the vector for *"dog"* is closer to the vector for *"hound"* than to the vector of *"pencil"*. Moreover, embeddings enable a certain level of arithmetic. The sum of vectors is often close to the combined meaning (Figure 4: *"Woman"* + *"royal"* = *"Queen"*). Embedding vectors represent single tokens, and their single dimensions are real-valued and cannot easily be interpreted. Therefore, embeddings are considered a *dense* text representation. Having such a vector space has advantages over sparse word representations. First, the vector space dimension is chosen a priori, whereas the dimension of sparse representations increases with the vocabulary size. Usual embedding spaces have only a few hundred dimensions, whereas sparse representations may have thousands of words as vocabulary. These results are a design and a computational advantage because many machine learning algorithms are easier to design and train in dense, low-dimensional feature spaces [31, 46].

Second, the embedding-based word representations usually generalize better as soon as they are pre-trained on large text corpora [46]. The pre-training calibrates the embedding vectors such that words that occur in similar sentences and positions have similar vectors. Hence, the embedding vectors of synonyms are close together, making it easier for machine learning algorithms to associate similar words with similar outputs. Once trained, these *pre-trained embeddings* can be shared among research groups as feature representations for texts without further training. Pre-trained embeddings are available for download in a numerous languages and have been successfully used as feature representation for a variety of tasks [106, 121].

2.2.4 *Transformer Architectures in Natural Language Processing*

The transformer neural network architecture [168] specializes in effectively processing long sequences and is key to most modern neural network-based AQG systems. Compared to formerly used sequence processing techniques such as Long Short-Term Memory (LSTM) [57] or Gated Recurrent Unit (GRU) [23], it can capture longer sequences and is easier to parallelize and thus, in practice, more computational effective [168]. It is currently used in many state-of-the-art systems, including applications in computer vision [69], NLU [30] and NLG [32, 123]. Its versatility stems from the different masking schemes [32] and pretraining objectives [123] that may be applied with the architecture. We outline three common architecture use cases (see Figure 5). In every use case, a token may be a word, subword, or a character.



bidirectional language model

left-to-right language model

sequence-to-sequence model

Figure 5: The different attention schemes frequently used in transformer architectures. Gray tokens cannot be attended, and white tokens can be attended by the model. In the sequence-to-sequence case, s_1 represents the tokens of the input sequence, whereas s_2 represents the tokens of the output sequence.

First, transformer models may approximate a *left-to-right language model* [126]. A left-to-right language model predicts the next output token w_{i+1} given the previous k tokens: $P(w_{i+1}|w_i, ..., w_{i-k+1})$. In other words, the model receives the sentence *"To be or not to __"* and predicts the blank *"be"* based on the previous tokens. This can be used for text generation by predicting w_{i+1} first and than autoregressively predicting w_{i+2} by feeding w_{i+1} back into the model.

Second, transformers may also approximate a *bidirectional language model* where they rely on the left and right context tokens for predicting w_k [30]. Given *"To be or __ to be"* the masked language model uses the tokens left and right to predict the blank token. The resulting models provide so-called *contextualized embeddings*, an embedding for the masked tokens contextualized on the surrounding tokens. These representations are helpful in many token-level classification tasks where the input sequence does not need to be autoregressively extended. A typical use case is, for instance, part-of-speech tagging, where the model predicts which word class a word belongs to (e.g. verb or adjective) [113, 172].

Third, transformers can be used to model *sequence-to-sequence* tasks. In a sequence-to-sequence task, the model is given a source sequence of tokens and predicts a target sequence. Typical use cases are, for example, machine translation, where the source sequence may be an English text, and the target sequence is the German translation [112]. Also, question realization is often modeled similarly. The declarative sentence is the source sequence, and the question to be generated is the target sequence [36]. In sequence-to-sequence tasks, the model masks the source and target sequence differently [32]. It is able to attend to all source sequence tokens in both directions to construct a meaningful source representation for the source tokens. Furthermore, it masks the target sequence similar to a conditional language model. Hence, the model relies on all source sequence tokens and the already predicted target sequence tokens for the next token prediction.

As with embeddings, these transformer models have the advantage that their model weights can be shared once they are pre-trained. Thus, as a well-adjusted transformer model has been trained, it can adapted to other tasks with minimal change to the architecture and without fully retraining all network weights [126].

2.2.5 *Evaluation Methodologies in AQG*

Syntactical and semantical aspects contribute to the quality of generated questions. In consequence, intricate automatic and human evaluation methods have to be used to capture question quality. We divide these evaluation methodologies into automatic, intrinsic and extrinsic evaluation because a thorough NLG evaluation methodology comprises automatic and human evaluation steps [85]. The intrinsic evaluation aims to evaluate the system's objective (e.g., generating high-quality questions), whereas extrinsic evaluation seeks to measure the system's task support (e.g., supporting reading comprehension) [65, p. 19].

Automatic evaluation is a cheap and fast measure of the generated texts' plausibility and the executed AQG subtasks quality. Information extraction methods are employed during the content selection subtasks. They are therefore frequently evaluated using precision, recall, and F1 score measures. The measures are built on the applied classifiers' respective false and true-positive classifications (fp/tp) and false and true-negative classifications (fn/tn). Precision is defined as $p = \frac{tp}{tp+fp}$ and measures the ratio of correct predictions given all positive predictions of a given system. The recall is defined as $r = \frac{tp}{tp+fn}$ and measures the ratio of correct predictions given all possible positive items for a class in the dataset. Finally, the class level F1 score is the harmonic mean of precision and recall $F = 2 \cdot \frac{p*r}{p+r}$. These metrics are used instead of plain accuracy, as the problems are often imbalanced. They allow us to estimate the quality of the educational AQG's content selection subtasks but cannot estimate the generated questions quality.

Besides content selection quality, question realization quality must be measured. A commonly used automatic evaluation metric for question quality in AQG is the Bilingual Evaluation Understudy (BLEU) [119] score that has been invented in the context of machine translation. It is a precision-oriented metric based on reference texts ranging between zero and one. A value of one indicates perfect overlap between reference and generated text, and zero indicates no overlap. Intuitively, it computes how often the tokens in the generated text are present in a given reference text. Tokens are counted only as often as they appear in the reference. As an example, if we have a prediction of *"the the the"* for the reference *"the cat is good"* the corresponding precision is $\frac{1}{3}$ and not $\frac{3}{3} = 1$. Besides this adaptation, BLEU also incorporates n-gram precisions into the calculation and adjusts to the predicted text's length. It is common to talk about BLEU-n scores (e.g., BLEU-4), where the n indicates the maximum n-gram length considered in the computation of the score. Hence, a BLEU-4 score indicates that the score includes n-gram overlaps between the generated text and the reference text up to a length of $n = 4$. However, automatic methods such as BLEU provide only a coarse view of the actual question quality and by construction do not measure actual text semantics. Studies have shown that automatically judging the quality of generated questions is only possible if one accepts a weak correlation between the evaluation metric and human perception of text quality [85]. After all, text comprehension happens in the readers' minds and is subject to factors like prior knowledge, attention or readers' beliefs which are hardly automatically measurable [2].

Next, besides the automatic measures, sound evaluation design contains intrinsic evaluation studies. In an intrinsic evaluation, an annotation study tests the generated texts' intrinsic characteristics and does not require an upstream task [8]. For instance, in AQG, question attributes such as understandability, syntactic correctness, or naturalness can usually be judged by human annotators without knowing the purpose of the questions and are thus considered intrinsic. Quality control for these intrinsic annotator studies is complex because text interpretation inevitably includes subjective aspects [2]. Intrinsic studies, therefore, ideally comprise multiple annotators judging multiple generated texts [85]. According to Van der Lee et al. [165], studies frequently employ one to four annotators. Moreover, it is common for studies to evaluate approximately 100 output texts [165]. In the multi-annotator setting is vital to measure Inter-Annotator Agreement (IAA) to quantify the subjectivity of the different ratings. Various chance-adjusted agreement measures are employed, depending on the used rating scale and annotator setup [122].

Finally, whereas intrinsic studies focus on task-independent quality measures, extrinsic studies use the generated texts during an upstream task and measure the task support provided by the generated texts. Extrinsic studies are empirical and often comprise a larger sample of participants trying to achieve a shared task. The specific study design is most variable relative to the previous methods and is highly dependent on the defined objective. This results in a considerable effort for the experimental design, operationalization of variables, and experimental execution. However, it is often worthwhile to carry out extrinsic evaluation because texts' intrinsic and extrinsic qualities are usually not congruent. Instead, the intrinsic and extrinsic evaluation may be complementary. It has, for example, been shown that for some tasks, the intrinsic judgement of the generated texts' human likeness does not necessarily relate to extrinsic task performance [7, 8]. In the context of educational AQG, this may also be the case. One may argue that questions may not have to be perfectly worded to support learners during reading comprehension.

2.3 STATE-OF-THE-ART

In this section, we will discuss the state-of-the-art for each of the different AQG steps and present relevant systems in educational AQG. We divide the discussion of context selection into three main directions: machine learning-based approaches, summarization-based approaches, and others. The presentation of the state-of-the-art answer selection is divided into two parts: machine learning-based approaches and other systems. Next, the question realization step is discussed, focusing on state-of-the-art neural systems as they outperformed templates and rules regularly. Thereafter, we provide an overview of the current uses of AQG systems in educational scenarios. Finally, we conclude by describing the identified technical research gap.

2.3.1 *Machine Learning-based Implicit Context Selection*

A main line of research investigates learning context selection implicitly and directly from a given dataset [34, 76, 94, 117, 173, 185]. That is, no explicit definition of what constitutes relevant information is applied. Instead, relevancy is implicitly given by the training data's relevant sentences.

The methods can be distinguished by the types of features they apply and to what degree they jointly learn other aspects of the AQG task. Initial research by Du and Cardie [34] models the context selection task as a sentence labeling task similar to supervised summarization. Their approach relies on a hierarchical neural network based on word embeddings. The authors use the selected sentences as input for an answer-agnostic question generator and report that applying their context selection improves the generator's performance compared to other machine learning-based context selectors. Given this initial success, most follow-up machine learning-based context selection addresses the joint task of context selection and question realization [76, 94, 117, 185]. In the joint task, a machine learning pipeline that jointly learns to ask relevant questions and extract contexts is built. It has been investigated to what extent restricting the question realization models' vocabulary [76] or allowing it to copy relevant words [76, 185] from the input text is sufficient for context selection. These adaptations improve the automatic evaluation metrics such as BLEU-4 [76, 185] and the perceived relevancy [76] of the generated questions. Furthermore, approaches have been developed that rely on the document and the answer for the joint learning task [94, 117]. Pan et al. [117] introduced context selection via dependency and semantic role labeling graph features to jointly learn to generate deeper questions given a document and an answer. They encode these graphs in a Graph Neural Network [83] and allow the question realization algorithm to attend to the graph features. Their human annotation study provides initial evidence that such features lead to the generation of questions requiring reasoning over multiple facts of the text, showing that their context selection slightly improves questions' depth. Additionally, Wang et al. [173] aim to extract contexts to generate more diverse questions and model context selection as a continuous multi-dimensional latent variable during the question generation process. The latent variable is learned jointly during training to approximate the diverse question distribution on the given training data. Their results suggest that the latent variable content selectors lead to more diverse question types.

These results provide evidence that implicitly learned context selection is promising if one assumes that the implicitly given relevancy definition of the training dataset is similar to the actual relevant information one aims to extract. Yet, these approaches have drawbacks. First, the implicit and joint learning results in opaque context selection criteria. What constitutes learning-relevant information is never made explicit and is only represented in the model's training data and parameter configuration. Accordingly, understanding why the model selects a context is difficult. Consequently, it becomes hard to tell whether or not a model will perform well on novel out-of-distribution data. Furthermore, most of these approaches (e.g. [76, 173, 185]) are evaluated on question-answering datasets. Yet, question-answering datasets like the Stanford Question Answering Dataset (SQuAD) [127, 128] are crowd-sourced, and their construction had no educational intent. Hence, it is almost impossible to tell if the proposed approaches transfer well to actual educational datasets, particularly considering the first drawback. Finally, most approaches [34, 76, 173, 185] measure the quality of the context selection via measuring the linguistic quality of the generated questions, thereby introducing noise from the answer and question realization tasks. Consequently, it is almost impossible to estimate the influence of context selection on the overall AQG task separately, and the generated questions' educational value is unknown.

2.3.2 *Summarization-based Context Selection*

Another line of research assumes that automatic summarization algorithms generate brief texts covering learning-relevant materials. Thus, they apply summarization-based context selection in which they summarize the input document and use the summary as input for answer selection or question realization [6, 19, 22, 94, 134].

Many different algorithms have been investigated on various datasets. First, Becker, Basu, and Vanderwende [6] explore utilizing the SumBasic [111] text summarization algorithm for context selection. The algorithm exploits a text's word co-occurrence matrix to author a summary of sentences using central words. Yet, the authors only evaluate their generated questions, not their context selector. They justify the application of the summarization algorithm only theoretically, hypothesizing that summaries encompass learning-relevant contents. To address this evaluation gap, Chen, Yang, and Gasevic [19] systematically explore which of a set of unsupervised summarization algorithms may be used for context selection on different educational and noneducational datasets. They thereby evaluate the context selection performance by proxy, using the selected contexts as input for a question realization algorithm which is then evaluated. In their study, no algorithm works consistently best on all datasets and sometimes heuristics like selecting the longest sentence or the first sentence of the text outperform all other algorithms. In their conclusion, the authors suggest that the *LexRank* [41] sentence similarity graph-based summarization algorithm gives a robust performance on a variety of datasets. In connection to this work, Rüdian, Heuts, and Pinkwart [134] show in a different study that the summarization algorithm of Edmundson [38] outperforms LexRank on their dataset. However, they evaluate the context selection in a German language setting on a single corpus with 48 texts. With the observation that no unsupervised algorithm performs well on all datasets in prior work, Mahdavi et al. [94] take multiple unsupervised summarization results as input for a trained sentence classifier. They show that this approach improves the quality of their generated questions on two datasets.

In summary, although the use of summarization-based context selectors has been investigated over the years, no study found an algorithm performing consistently on multiple datasets. Furthermore, the evaluation studies for those context selection approaches repeatedly intertwine question realization and context selection evaluation. Thus, similar to the machine learning-based context selectors, it becomes difficult to tell which effect the context selector had on the overall system performance.

2.3.3 *Other Approaches for Context Selection*

Besides machine learning and automatic summarization, various other context selection approaches have been investigated [29, 61, 74, 120, 147, 161]. These approaches are harder to categorize into a specific type of algorithm. Instead, their commonality is more general in that they try to define a helpful, automatically detectable proxy measure for context relevancy based on educational or linguistical premises. They then apply different algorithms to select sentences according to their defined metric.

In the linguistic realm, it has been proposed to define learning-relevant sentences through topic modeling, extracting sentences most likely to express distinct topics [74]. Moreover, Pavlik Jr. et al. [120] detect learning-relevant sentences via co-reference chains under the assumption that sentences in a longer chain are more important. Yet, the respecitve authors only provide theoretical justification for their sentence selection approaches and no distinct evaluation. Other authors define custom word vector-based metrics for the keyness, completeness, and independence of sentences' to extract relevant but not redundant information [61]. Their evaluation shows that this approach results in a better context selection than a simple baseline that selects the first sentence of every text's quarter.

In the educational domain, some research groups limit the set of relevant sentences by some explicitly stated educational prior [29, 147, 161]. Desai, Dakle, and Moldovan [29] work under the assumption that their input is tagged via *Rhetorical Structure Theory* [95], indicating the discourse information of a text. Given this annotated corpus, they generate questions to contexts depending on the relation expressed by the annotations. Furthermore, Stasaski et al. [147] detect cause and effect relations automatically in texts and generate questions based on these relations. In their argument, cause and effect questions are higher in Bloom's learning taxonomy [10], thus more helpful for learners. Their evaluation shows that the selected contexts are more likely to generate cause and effect questions. Finally, Syed et al. [161] propose an eye-tracking-based approach for active learning. They first determine the learners' attention state via eye-tracking and pose a question when learners' attention falls below a certain threshold. The evaluation indicates that using gaze tracking indicators for context selection may improve learning for low knowledge learners and hinders learning slightly in high knowledge learners.

The educationally motivated context selectors have the advantage that they are grounded in theory. Therefore, in our opinion, the overall system becomes more predictable than the summarization and direct learning approaches. After all, even if the selection fails, the applied selection criteria are transparent and easier to communicate to users and experts. Moreover, it becomes easier to reason whether or not a system transfers to a particular domain, as the selection criterion is known in advance. However, these systems also have drawbacks. As with the previous research directions for context selection, mainly the generated questions and not the context selection itself is evaluated [29, 74, 120, 147, 161]. Moreover, some approaches are only applicable when additional inputs are provided [29, 161].

2.3.4 *Machine Learning-based Answer Selection*

Several works have modeled answer selection as a supervised machine learning task simliarly to implicit context selection [6, 35, 74, 75, 117, 125, 158, 171, 174, 177]. In an early study Becker, Basu, and Vanderwende [6] used a logistic regression classifier for answer selection, whereas more recent works mainly rely on neural networks [35, 75, 117, 125, 158, 171, 177]. The underlying problem formulation is diverse. Some researchers investigated the task as a classification problem [6, 74, 75, 177], extracting candidate phrases via pos tagging [74], Named Entity Recognition (NER) [74, 75, 177], or semantic role labeling [6] before using the classifier to decide whether they constitute a potential answer. Moreover, the problem may be seen as a sequence labeling task [35] where every token in the context's token sequence is marked as either being part of an answer or not. Other approaches [75, 158] tackle the task as an information extraction problem and apply neural network architectures such as pointer networks [169] to extract the indices of potential answers. Furthermore, Willis et al. [177] compared classification-based approaches with generative keyphrase construction approaches. In their experiment on the SQuAD dataset, the generative system outperformed the classification approaches in phrase-level F1 score. Finally, the answer selection has been learned jointly with question realization [117, 125, 171]. For that, auxiliary learning goals [171] or an iterative question-keyphrase generation module [125] have been investigated.

Independent of the problem formulation, most studies investigate noneducational datasets [6, 35, 74, 75, 117, 158, 171, 177]. Only Wang et al. [174] and Qu, Jia, and Wu [125] focus explicitly on educational data. Consequently, it is challenging to estimate if the learned knowledge is directly transferable to the educational domain. Furthermore, assuming an educational training dataset, it is unclear if the general learning approach functions because one may argue that educational answers are inherently different from answers found in question-answering datasets.

2.3.5 *Other Approaches for Answer Selection*

Aside from answer selection via machine learning directly on answer selection corpora, other answer selection approaches have been investigated [29, 93, 100, 101, 147, 174]. Likewise to context selection, the lines of research can be divided into linguistically informed and educationally informed approaches.

The linguistically informed line of research relies on extra annotations and automatic tagging for answer selection. Mazidi et al. [100, 101] rely on linguistic analysis based on syntactic parsing and semantic role labeling to infer relevant answers in given sentences. They show that this combined view leads to many acceptable questions in a template-based question realization setting. Additionally, the system by Stasaski et al. [147] employs causal sentence detection to detect the sentence parts that express cause or effect. They show that upstream question realization algorithms generate more frequent cause-and-effect questions given these sentence parts. Besides, the *Rhetorical Structure Theory* approach by Desai, Dakle, and Moldovan [29] relies on a corpus, which explicitly describes the argument structure of texts. Various questions may subsequently be posed using either the nucleus or the satellite of the sentences as question target, capitalizing on the underlying annotation structure.

In education, it has been investigated if keyphrase extraction algorithms [93] or back-of-the-book indices [174] can be used for answer selection. Both related works have been evaluated with educational data either showing promising question quality on textbook data [174] or an increase in learning outcomes compared to a control group [93]. Again, a drawback of these approaches is the missing evaluation of the standalone answer selection. Furthermore, the advantages of educationally motivated context selectors apply similarly to answer selection. We suspect these approaches to be more transparent and easier to transfer between domains (see Section 2.3.3).

2.3.6 *Question Realization*

Initial question realization algorithms were mainly based on question templates [26, 86, 87, 100, 101] or handcrafted transformation rules [56]. However, templates are more likely to result in a repetitive set of questions [116]. Moreover, matching and filling template variables with the correct text chunks involves elaborate handcrafted rules, especially when the answer is not just a noun but comprises a whole clause of the source sentence [101]. In such cases, the template variable has to be simplified via syntax transformations [77] sharing the weaknesses of rule-based approaches. Rule-based approaches perform well on short declarative sentences but struggle with complex syntactic constructs and thus often need to apply additional sentence simplification steps [56, 77]. Syntactic patterns are manifold for complex sentences, and numerous exceptional cases arise. Manually authored rules, therefore, frequently fail to capture this expressiveness.

Recently, neural network-based question realization algorithms have been introduced. In contrast to templates and rules, they usually do not rely on any symbolic representation of syntax but tread question realization as a sequence-to-sequence task [36]. Initial architectures were plain answer-agnostic generators relying solely on the information directly learned from the source dataset and did not receive the expected answer as an input [36]. Research quickly explored various additional feature sets [117, 186] and answer-aware architectures [32, 167, 186] dealing with the specifics of question realization. For instance, Kim et al. [71] introduced an answer masking scheme for the sequence-to-sequence learning process to prevent generating questions containing their answer while also focusing more on the expected answer. Moreover, Zhou, Zhang, and Wu [187] introduced a question type prediction module in their architecture to match the questioning word to the expected answer. Furthermore, joint learning or reinforcement learning between the question answering and question realization task has been explored [60, 159] as well as the introduction of semantic graph features [117]. These various architectural changes frequently improved the respective state-of-the-art as measured automatically by BLEU scores on SQuAD or other reference corpora. Finally, the task recently became more nuanced with investigations of difficulty controllable question realization [21] or multihop question realization [157].

Apart from the neural architectures trained only on the question realization task, large pre-trained language models have also been investigated for the question realization [32, 123, 167]. In contrast to the previous neural models, these models initially train a task-independent large language model that is fine-tuned to the AQG task afterward. Models like UNILM [32] or CopyBERT [167] use different attention masks while pre-training a large transformer on massive corpora. This variation in attention masks allows them to jointly learn conditional, masked, and sequence-to-sequence modeling improving their general text generation capabilities. Additionally, Qi et al. [123] introduce ProphetNet, which uses multiple attention masks and also relies on more elaborate text generation loss functions based on the model's n-gram predictions, improving the learning of language patterns from massive corpora. In consequence, these large pre-trained models outperform their non pre-trained counterparts on the question realization task in terms of automatic metrics [32]. However, they require way more computational resources for training and inference [11].

In summary, question realization research shifted recently from rule-based and template-based approaches to neural question realization outside education. In contrast, educational technology research mainly relies on rule-based and template-based question realization [77]. Noteworthy exceptions explicitly investigating neural question realization approaches under educational considerations can be found in Section 2.3.7. The novel neural systems outperform the rule-based and template-based systems measured by automatic metrics [32] and linguistic annotation studies [36]. They possess different architectures, are are either answer-aware or answer-agnostic and also differ in respect to their pretraining. However, currently, neural pre-trained answer-aware models' generation capabilities are primarily judged in terms of their question quality on reference corpora [32, 36, 123, 167, 186]. Little is known about their language generation capabilities outside the reference corpora, for instance, when applied in an actual learning setting.

2.3.7 *AQG Systems in Education*

Many AQG systems are investigated in pure natural language processing research. They are automatically [32, 167] or intrinsically evaluated via aspects such as fluency, answerability, relevance or difficulty [36, 117]. In contrast, some AQG systems have been built that are explicitly aimed toward educational use and have been evaluated with respect to educational dimensions [59, 93, 160, 161, 163, 184]. An extensive analysis of mainly educational rule-based and template-based systems can be found in Kurdi et al. [77]. Besides reading comprehension AQG systems, educational AQG also aims to generate structured question items for single domain tasks such as theoretical computer science [67] or programming [162]. This thesis will not discuss these structured AQG systems in-depth, as they are domain-focused and not geared towards text comprehension.

A few authors have experimented with neural AQG systems. Lu et al. [93] conducted a pre/posttest design study with programming language learners to investigate the influence of the generated questions on learning outcomes. Their control group studies without questions, whereas their experimental group receives generated questions and is also automatically graded during learning. Their AQG system builts on templates filled by modern neural models. Even with prior knowledge as a covariate, the experimental group received significantly higher posttest scores than the control group. Moreover, Syed et al. [161] investigate the effects of automatically generated questions via a neural generator coupled with eye-tracking in a reading comprehension scenario. They distinguish between high and low-knowledge learners and provide four different conditions: no question, automatic question, literal human questions and literal+deep human questions. They measure learning outcomes directly after the experiment and with a delayed posttest for long-term effects. The study finds evidence for improved long-term retention when learners are confronted with automatically-generated questions. Interestingly, the effect for automatic questions was more pronounced than the effect of manually authored questions. Finally, Horbach et al. [59] introduce an expert annotation scheme for evaluating neural AQG generated questions' syntactic and educational quality. Their hierarchical scheme consists of four groups of evaluation items where higher groups consider questions' educational aspects and lower groups evaluate linguistic and fundamental understanding aspects. When applying their scheme to a neural question generator, they find that experts judge the neural questions as less educationally valuable than manually authored questions.

In summary, only a limited set of studies discuss modern AQG approaches applying neural question realization. Only two studies address the influence of systems with neural components on learning outcomes to the best of our knowledge. They found positive effects of the generated questions on learning outcomes compared to no-question control groups. However, they either rely on systems that are only partly neural or employ eye tracking-based question sequencing in their experiments. In contrast, non-neural systems are evaluated on various dimensions, such as difficulty, engagement or learner persistence [163], and studies find positive effects on learning-related variables [160].

2.4 SUMMARY AND IDENTIFIED RESEARCH GAP

This chapter introduced the necessary AQG preliminaries and provided an in-depth analysis of the current technical state-of-the-art in educational question generation. This analysis results in the following identified research gaps.

From an overall systems perspective, we identified a research gap between fundamental AQG research and educational AQG research. Educational systems primarily rely on templates and rule-based question realization algorithms. In contrast, state-of-the-art AQG research suggests that neural question realization is the superior NLG paradigm and results in higher-quality generated questions. Consequently, investigating neural question realization systems in the educational domain is a promising research direction. This idea results in multiple additional research gaps to be addressed.

First, neural question realization algorithms are currently primarily evaluated on noneducational reference corpora stemming from question-answering. Hence, if they transfer their superior text generation capabilities to educational datasets is unclear because reading comprehension texts are considerably different from these corpora. Moreover, they are frequently evaluated and compared using automatic measures such as BLEU. Yet, as pointed out in Section 2.2.5, these measures do not cover all desired question attributes and frequently only weakly correlate with human perception of text quality [85]. In consequence, empirically evaluating their language generation capabilities in the educational domain is paramount before applying them in the educational AQG process.

Second, we have discussed in Section 2.2.1 that learning-relevant inputs are required for educational AQG. That is, sole neural question realization is insufficient for an actual educational AQG pipeline, and we need to acquire learning-relevant input before question realization. However, in the reviewed literature, the context selection and answer selection subprocesses are often only evaluated via proxy measures making it difficult to estimate their influence on the generation. Hence, more isolated research investigating these subprocesses is needed. The investigation has multiple promising research directions. On the one hand, machine learning approaches advanced the state-of-the-art context and answer selection on noneducational question-answering datasets. However, it has to be shown if they transfer their capabilities to educational corpora. On the other hand, we argue in Section 2.3.3 and Section 2.3.5 that linguistically and educationally motivated algorithms have advantages over implicit machine learning-based approaches. Thus, in this respect, we identify two research gaps: investigating the transfer of machine learning-based content selection to education and investigating a combination of linguistically and educationally motivated content selector and machine learning-based approaches.

Finally, although educational AQG encompasses content selection and question realization, its educational affordances can not easily be derived by analyzing the subcomponents. Instead, the output of the overall system has to be evaluated intrinsically and extrinsically in an actual learning context (see Section 2.2.5). Currently, there is a research gap in this respect as research has concentrated on the influences of non-neural AQG pipelines in education. In consequence, the affordances of neural AQG pipelines are not well understood, and more research is needed to understand possible application scenarios and their effects on reading comprehension.

3 OVERALL CONCEPT AND APPROACH

The primary thesis goal is to investigate two research questions with respect to educational Automatic Question Generation (AQG).

RQ1 To what extent can content selectors, based on textual features, extract learning-relevant information from educational science texts in different domains?

RQ2 To what extent can educational AQGs generate literal questions supporting reading comprehension of educational science texts in different domains?

The research questions are both investigated with the goal of building an educational AQG approach for literal short-answer questions, as stated in the introduction. Thus, we begin this chapter by providing an in-depth justification for investigating educational literal AQG, discussing the effectiveness of manually authored literal questions. Furthermore, we highlight critical contextual factors that must be considered when utilizing and evaluating the literal questions for text comprehension support.

Next, we move from the evidence for the effectiveness of manually authored questions to important design considerations for the successful automatic generation of educationally valuable literal questions. We discuss the design space along the question generation process (see Section 2.2.1), first reviewing conceptual works influencing content selection approaches. Afterward, we focus on the question realization step and the linguistic aspects involved. We finally link this conceptual analysis with the research gaps identified in the related work (see Section 2.4) to derive our research approach.

3.1 EFFECTS OF MANUALLY AUTHORED LITERAL QUESTIONS

Manually authored literal questions have been frequently researched. The following sections discuss their effects on learning outcomes and findings one has to consider when utilizing and evaluating such questions in learning scenarios. We discuss these findings first because we can derive the potential usefulness of automatically generated literal questions based on the effectiveness of manually authored questions.

3.1.1 *Literal Questions and Their Effect on Learning Outcomes*

There is compelling evidence that literal questions foster learning. Influencial meta-reviews by Anderson and Biddle [4] as well as Hamaker [51] show that literal short-answer questions increase learning outcomes. These meta-reviews include various empirical studies indicating the positive effects of the questions, and newer literature confirms the results [13, 39, 53, 132, 141]. The empirical evidence suggests that the questions not only increase learning outcomes but also perform better than multiple-choice questions or cloze questions [39, 51]. This may be explained by the fact that short-answer questions allow less pattern matching while answering compared to multiple-choice or cloze questions [3]. Thus, learners are more likely to actively monitor their current understanding leading to better comprehension.

Additionally, literal short answer questions cause the doer effect [73] and the testing effect [39]. The doer effect describes that learners actively engaging with material have a better learning effect than those that only consume the material passively. The testing effect states that long-term memory increases when learners actively train to recall the information [131]. In the testing effect's experimental setup, a control group reads a text several times. In contrast, the experimental group takes a literal test concerning the text's content. Suppose these groups are invited to a delayed recall test, for example, after one week. In that case, the experimental group likely remembers significantly more of what they have read than the control group. This shows that literal questions can lead to prolonged retention of the read information.

In summary, literal questions increase the direct comprehension of the text and the long-term memory of the information. Therefore, they are a sound and well established instructional means to foster learning.

3.1.2 *Embedding Literal Questions into Learning Scenarios*

Besides the direct effect on learning outcomes, the research literature also provides additional contextualizing information with respect to literal questions. Not every application scenario of these questions works equally well. The meta-reviews by Anderson and Biddle [4], as well as Hamaker [51] show that the position of the literal questions during reading is important. Pre-questions are asked before the learners start to read the texts containing the answer, and post-questions are asked after learners have read the text. Post-questions are likely to increase the learning outcome without directing the learner's attention to specific information in advance. In contrast, pre-questions influence learners' attention a priori and cause learners to focus their reading on the questions' answers. Therefore, pre-questions may actually worsen learning of unquestioned information. The effect has not been found for post-questions in the meta-reviews, which either increase learning or do not affect the learning outcomes. Furthermore, the meta-reviews suggest that the effects of questions are most evident when subjects have to write an answer to the questions.

Additionally, empirical evidence indicates that questions do not equally facilitate the comprehension of all information in the text. Hamaker [51] discusses the effect of literal short-answer questions on posttest performance in his meta-review. The author shows that an increase in learning outcomes is only significant for the information addressed by the short-answer questions. For that, the posttest items measuring the effects of the literal questions are classified into three categories: verbatim, related and unrelated to the encountered short-answer questions during reading comprehension. Verbatim items copy the experimental literal questions, and related items target the same or similar information. Unrelated items cover information not targeted by the reading comprehension questions. Thereby, a significant effect of reading comprehension questions on the verbatim and related posttest items was found in the meta-review[51], yet no significant effect emerged for the unrelated posttest items.

In summary, literal short answer questions only support learning the information related to their answer. Posttest questions are generally considered a better instructional means than pretest questions because they do not cause initial attention priming. Literal questions do not increase learners' overall learning outcomes or foster the learning of information never questioned. In consequence, targeting learning-relevant information pieces with the questions is crucial.

3.1.3 *Design Decisions*

These research results concerning manually authored literal questions lead us to two design decisions with respect to the research conducted in this thesis.

First, the results indicate that focusing on literal questions because they positively affect learning. Accordingly, in combination with the analyzed technical research gap (see Section 2.4), we will explicitly study the generation of literal questions rather than deeper comprehension questions in this thesis (see Chapter 4; Chapter 8). After all, conceptually they improve comprehension, and the potential generation algorithms for these literal questions drastically improved recently.

Second, the presented evidence informs our respective evaluation design decisions (RQ 2). In case of extrinsic learner-based studies (see Chapter 8), we will not use pretest questions because they may lead to suboptimal attention priming. Moreover, participants must provide a written answer to the questions to ensure that they engage mentally with the questioned information. Lastly, we will measure different learning outcome categories, as previous work has shown that the literal questions only improve related and verbatim learning outcomes but not unrelated learning outcomes.

We have now established that manually authored literal questions support learning. Additionally, we inferred general evaluation guidelines if one wants to test the effect of literal questions on text comprehension. However, the thesis aims to automatically generate these literal questions fostering text comprehension. Thus, the subsequent sections discuss the conceptual aspect influencing the AQG process and its quality. We start with considerations influencing the content selection and subsequently discuss concerns affecting question realization.

3.2 CONCEPTUAL CONTENT SELECTION CONSIDERATIONS

This section analyzes what is known about the reading comprehension process in learners and how comprehension builds on learning-relevant information. Moreover, we also discuss how text structure comprises clues on learning-relevant information allowing readers and potential automatic content selectors to focus on learning-relevant information.

3.2.1 *Reading Comprehension Models and Learning-relevant Information*

There are important insights given in educational literature about what constitutes relevant information given a reader's mental text representation.

According to the given reading comprehension models, comprehension is usually assumed to be concept-based [103]. Readers construct their mental representation based on propositions compiled from concepts and connecting predicates. This mental representation is vital for most text comprehension processes. The most relevant concepts are ideally well connected in the readers' propositional network and interact with many other propositions, allowing readers to use their conceptual knowledge as an aid during reading [133]. For example, in a text about computer networks, core concepts such as *"network packets"* recur and are central to comprehension.

Moreover, reading comprehension models assume that readers enrich their mental propositional network with prior knowledge [103]. For instance, if readers have prior knowledge about the *"Domain Name System"*, they probably have an easier time understanding a text describing how a web browser functions. However, the domain of science texts poses severe challenges for mental model construction and integration of prior knowledge. The texts are densly written, introduce many concepts and are mainly self-contained [9, 133]. Hence, readers can rely less on their prior knowledge about the core concepts and if they have prior knowledge, it is subject to misconceptions [9, 133]. In other words, conceptual networks form the basis of reading comprehension, and for science texts, the core concepts are often not given by readers' prior knowledge, but readers must acquire them directly from the text.

Another influencial finding is that good readers monitor the text's causal chain and infer actively to keep their mental model coherent with the text [9, 92, 140, 164]. Some authors argue that this inference, when it co-occurs over large parts of the text, is the key to deep comprehension because it leads to the reader's mental representation not being fragmented at the sentence level but reflecting the overall context [9]. That is, relevant information is connected to the text's primary causal chain and inferences thereby occur at multiple textual levels [9, 92, 164]. At the lower level, readers resolve co-references (e.g. pronouns) whereas, at the higher level, they integrate different paragraphs' content.

Related to these causal chains, science texts frequently encompass causal explanations of real-world phenomena. For example, physics texts deduce new relationships from a few laws of nature. Similarly, biology or network engineering books, describe complex causal dependencies between individual components, systems and processes. These causal chains are often more complex and harder to understand compared to simple causal chains in narrative texts [82]. Therefore, the understanding of causal chains is another influential comprehension factor and is particularly relevant and challenging in science texts.

3.2.2 *Text Structure and Learning-relevant Information*

Besides the reading comprehension models, educational psychology has researched textual factors communicating information relevance.

On the structural level, authors have numerous ways to signal information relevance [90, 91, 98, 115]. First, it is evident that the information order in texts is not arbitrary. Studies suggest that paragraphs usually begin or end with the most relevant information [70]. That is, authors either introduce something of relevance and continue with sentences supporting the thought or they build a cascade of arguments leading to a conclusion. Second, at the sentence level, additional language constructs are used [90, 98]. These constructs, for instance, increase the text's coherence (e.g. subordinating conjunctions like *"as"*), emphasize points (e.g. *"In other words..."*) or make causality explicit (e.g. *"Thus / Therefore.."*). Third, on the typographical level, different highlighting mechanisms are used [90, 91]. For example, relevant concepts appear in the heading, are italicized in the text, or are indexed in the back-of-the-book index. All these structural aids explicate the author's relevance view and assist the reader navigating the text. In contrast to the perceived relevant concepts in the readers' mind described above, the structural indicators are easily identifiable in texts. Consequently, they are likely helpful assets for automatically identifying relevant information.

Furthermore, evidence suggests that readers apply simplified content-based heuristics to determine relevant information. Dee-Lucas et al. show [27, 28] that novices and experts rely on the sentence type for judging the relevance of information in physics texts. They showed that novices regard definitions more relevant to their learning than other sentences. Likewise, subject-matter experts agreed in the study of Dee-Lucas and Larkin [28] that this definition heuristic is sound. However, novices also had great difficulty recalling definitions. That is, although they regard the information as relevant, they struggle to learn and remember it. Moreover, there is evidence that readers judge causal sentences as more relevant when the underlying causality is highlighted with pointer words such as *"because"* or *"as a result"* [98].

3.2.3 *Design Decisions*

Given the presented evidence, we make the following design decisions.

First, we have presented evidence that reading comprehension is concept-based and builds on the successful understanding of key concepts. Moreover, science texts often are more complicated as they introduce many foreign concepts, not necessarily present in the readers' prior knowledge. Thus, we argue that it is vital to support the readers' comprehension of key concepts by actively engaging them with the definitions of these concepts. Therefore, we decide to target definitions with our context selection approaches (see Chapter 5). Furthermore, we have also seen that the readers' understanding of the texts' causal chain fosters their comprehension. In consequence, we also focus on building context selectors that explicitly target causal sentences (see Chapter 5).

Second, reading comprehension is concept-based. Hence, we argue the selection of these learning-relevant concepts from texts helps to build educational AQG approaches (see Chapter 7). We have seen that these concepts and their respective sentence are sometimes associated with specific text characteristics. Therefore, we aim to exploit these characteristics on the sentence, phrase, and typographical levels to extract learning-relevant content from texts (see Chapter 6; Chapter 7).

3.3 conceptual question realization considerations

In this section, we discuss important linguistic requirements specific to the question realization phase, which also influence the overall usefulness of educational AQG.

Spelling and Grammaticality

Spelling and grammaticality are vital for linguistic question quality. They indicate if the generated text follows the standard writing rules of the output language. Reading comprehension models postulate that readers construct their comprehension based on these syntactic rules and relations [103]. After reading a sentence, readers update their mental representation based on the read concepts and rely thereby on syntactical relations to integrate how the sentence's concepts interact with each other [103, 164]. This mental representation forms the basis for the readers' overall text understanding [103, 164]. Thus, grammar and spelling errors can negatively affect comprehension by hampering the correct interpretation of the given sentence.

Moreover, the syntactic complexity of questions may affect the readers' understanding. Reading comprehension theories state that readers have a limited amount of working memory that has to be managed during reading comprehension [103]. Hence, complex syntactic structures negatively affect understanding [47] and may overload their working memory [81]. Particularly, the more propositions a question comprises, the more likely it is that the sentence parsing overflows the reader's working memory, resulting in a poor question understanding.

3.3.1 *Semantic Understandability*

Semantic understandability is essential for text comprehension. A question is semantically understandable if the reader can deduce a possible meaning from the given question. It is a minimum requirement for all generated questions because if a question is not understandable, it will not help the readers to reflect on their readings.

Furthermore, semantic understandability correlates with but differs from syntactical correctness in terms of grammar and spelling. There are various sentences with sound syntax but without appearant meaning, as the famous example: *"Colorless green ideas sleep furiously."* by Chomsky illustrates [24, p. 2]. Moreover, there are sentences which violate formal grammar rules but are nevertheless acceptable by a portion of native speakers [145]. Hence experiments ideally measure syntax and understandability as two separate dimensions.

Understandability does not require every reader to deduce the same meaning for the generated text, as questions may be ambiguous. For instance, given the question *"What is the president of the USA?"* readers may assume they understand the meaning but answer in two different directions. One writes down what constitutes the position of U.S. President, while the other names the current U.S. President. In the context of reading comprehension, both interpretations may support the readers' learning outcome depending on their learning goals and the source text. Consequently, one ideally distinguishes between ambiguity and the overall semantic understandability of a question.

3.3.2 *Answerability*

Finally, the overall answerability of the given questions influences the linguistic quality. While there are questions posed without expecting an answer, reading comprehension questions require a response. A question's answerability, thereby, does not necessarily depend on the text. Readers may answer the question using prior knowledge, common sense or external sources. Yet, constructing questions answerable by the information available from the previously read material is aspired in reading comprehension scenarios.

3.3.3 *Design Decisions*

The discussed factors in the previous sections contribute to the question's linguistic quality. Furthermore, they frequently are measured in related works, sometimes named differently (e.g. [22, 59, 62, 174]). Consequently, we will use them as the conceptual underpinning for our evaluation studies as they define vital measures for overall linguistic quality. They are essential for evaluating the second research question, as a minimum of linguistic quality is necessary to realize educational quality. However, not all subsequent evaluations in this thesis will measure all of these properties due to time and resource constraints. Notably, whenever we investigate the first research question, we will evaluate the quality of the content selectors' extracted information instead of the linguistic quality of generated questions. That is, we evaluate the first research question directly, without the proxy of generated questions as discussed in Section 2.4. Yet, the evaluation of the final proposed educational AQG system provides an in-depth intrinsic assessment of all these factors on a variety of scales (see Chapter 8).

3.4 APPROACH

Our overall research approach is guided by the previous conceptual analysis, the discussed design decisions, and the technical analysis of the related work and its research gap in Chapter 2. We address the given research questions with novel approaches for context and answer selection (RQ 1) and combine these content selectors with neural network-based Natural Language Generation (NLG) to investigate the affordances of educational AQG (RQ 2).

3.4.1 *Context Selection*

With respect to context selection, we focus on causal and definitory sentence selection (see Chapter 5). We have analyzed in Section 3.2 that both selections are educationally valuable AQG inputs. Notably, they focus on vital aspects in the science reading comprehension process. Furthermore, both selection categories frequently rely on specific syntax patterns and wording, enabling automatic detection possibilities. Hence, they are a promising context selection target with sound educational underpinnings and detectable textual grounding.

Alternatively, we could have focused on implicit machine learning-based context selectors. However, as the related work analysis revealed (see Section 2.4) implicit context selectors are more opaque in their decision criteria, complicating estimating their performance a priori on unseen out-of-domain datasets. Additionally, this complicates their use in human-in-the-loop scenarios because humans do not understand their decision criteria. Moreover, while noneducational research and machine learning-based context selection datasets exist, the research field lacks large-scale, high-quality datasets for the educational domain to train and compare systems. Besides, summarization-based context selectors could have been investigated. However, they share the opaqueness issue with the implicit machine learning-based approaches (see Section 2.4). Thus, we will mainly use them to compare the proposed approaches.

3.4.2 *Answer Selection*

For the answer selection subtask, linguistic and machine learning-based approaches are investigated. Initially, we determine if machine learning-based answer selection models transfer well from noneducational to educational corpora or if they differ in their selection criteria (see Chapter 6). We conduct this evaluation because in Section 2.4 we analyzed that there is a need for studies that directly investigate the transfer of content selection from noneducational to educational data. Results from these experiments have immediate implications for the first research question.

In the light of the results, we then investigate the extent to which we can generate large educational corpora for answer selection fully automatically from textbooks, alleviating the need for a direct transfer from models trained on noneducational corpora to the educational domain (see Chapter 7). Finally, we integrate the insights into our design for an educational AQG system, in which we investigate a linguistically motivated answer selection approach based on the findings from the previous context selection and answer selection experiments (see Chapter 8). The proposed design thereby relies on structural text information to select relevant answers from the given contexts (see Section 3.2).

With respect to alternative approaches, our investigation of answer selection starts broadly evaluating different alternatives. We then decide on the most promising method for our final system based on the insights gathered. However, the investigation primarily considers machine learning-based, named entity-based and dependency graph features. It excludes alternative tagging schemes, such as annotations from *Rhetorical Structure Theory* (see Section 2.3.5) because it relies on less precise automatic taggers and requires manual annotation [29]. Moreover, we do not investigate abstractive and joint learned answer generation (see Section 2.3.4) because the subsequent question realization models are trained on extractive inputs. Accordingly, we assume that the question realization phase works best when the answers provided are as accurate as possible, reflecting the actual information provided in the text. Furthermore, joint learning implies that the underlying corpus has educationally valuable answers annotated, which is either unclear or often not the case as we will see in Chapter 6. Hence, we focus on extractive answer selection without joint learning in subsequent chapters.

3.4.3 *Question Realization and Educational Evaluation*

In the question realization phase, we evaluate the principle transferability of different neural architectures to the educational domain in a pilot study (see Chapter 4) because as indicated in Section 2.4, it is unclear if neural generators transfer to out-of-distribution data. Particularly, our analysis in Section 3.3 argues that minimal linguistic quality is a prerequisite for helpful reading comprehension questions. With the insights gathered in the pilot evaluation, we focus on a pre-trained, answer-aware neural generator for the question realization phase because of its high-linguistic qualities in subsequent chapters.

Alternatively, we could have focussed on template and rule-based approaches. However, it was ruled out because, although widespread in education, their linguistic quality has been shown to be inferior to neural approaches in noneducational research (see Section 2.3.6).

Finally, we combine the insights gathered while investigating the different AQG subtasks in this thesis and deduce an architecture for an educational AQG approach generating literal short-answer questions (see Chapter 8). As we discuss in Section 3.1, there is evidence for the potential educational usefulness of such a literal AQG. We evaluate the educational AQG system in expert- and learner-based studies to investigate the second research question and to understand its affordances and effects on learning outcomes. Thereby, we rely on the deduced best practices for NLG evaluation (see Section 2.2.5) and educational evaluation (see Section 3.1).

4 PILOTING LINGUISTIC GENERATION QUALITY IN EDUCATION

This thesis investigates educational Automatic Question Generation (AQG) systems applying neural network-based methods to foster reading comprehension. We have established that such systems require linguistic and educational qualities to foster learning. Both qualities are interdependent, and a minimum linguistic quality is necessary to pose meaningful questions. From a conceptual standpoint, the educational AQG process comprises two content selection subtasks and one question realization subtask (see Chapter 2). On the one hand, generated questions' educational requirements, such as learning relevance, primarily depend on the content selection subtasks. On the other hand, generated questions' linguistic quality is affected primarily by the question realization subtask. Hence, it is imperative to validate the linguistic quality of potential question realization components before investigating the content selection in more detail. After all, if neural network-based question realization algorithms fail to transfer their linguistic expressivity to the educational domain, learning-relevant inputs will not improve their performance. In this case, the linguistic quality stales the overall AQG performance. Therefore, we aim to validate the assumption that neural network-based question realization algorithms transfer their linguistic quality into the education domain in a pilot study.

4.1 METHODOLOGY

We compare an answer-agnostic and an answer-aware question realization model to provide initial linguistic and educational quality estimates of neural question realization approaches in education. The evaluation uses the same context selection for the two question realization models to ensure similar initial inputs for this shared step in both models. Moreover, four system configurations are evaluated due to the answer selection step only available in the answer-aware model (see Figure 6). The answer-agnostic model does not require answer selection and yields the first condition. Furthermore, we vary the answer selection for the answer-aware model by three conditions: *Random* selection, grammatical *Subject* selection and *Direct Object* selection. Consequently, the four different conditions allow us to measure both models' output quality while controlling for their content selection.

We do not train the question realization models, but rely on fine-tuned model snapshots that were state-of-the-art at the time of the model's respective release. They were trained on the Stanford Question Answering Dataset (SQuAD) dataset [128] for AQG and did not receive any education-specific fine-tuning. Thereby, we primarily aim to determine if their linguistic quality characteristics transfer when confronted with educational inputs. Particularly, we evaluate if the models are still reliable on

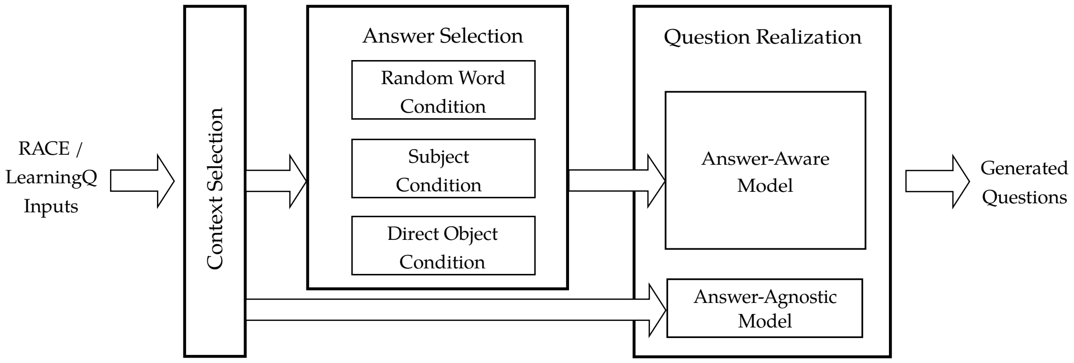


Figure 6: The pilot study's methodology. We use two educational datasets as input to an educational AQG pipeline. After selecting context sentences, we generate questions using four different answer selection conditions and two different question realization models.

educational sentences not stemming from the same distribution as the training data. The models' reliability on out-of-distribution data is not self-evident [138]. It has been observed that neural models rely on non-robust features and correlations present in their training distribution, deteriorating performance on out-of-distribution data [25].

The first question realization model under investigation is an answer-agnostic model trained on SQuAD relying on pretrained word embeddings. The second question realization model is an answer-aware question realization model that was first pre-trained on the English Wikipedia and the BookCorpus [188] and then fine-tuned on SQuAD for AQG. That is, the two models under investigation differ in their architecture and pre-training data. We select both models for the reasons provided in Section 4.3. Both models are evaluated automatically on two educational datasets. Additionally, we conduct an expert-based annotation of the generated questions on one dataset, to better understand their quality characteristics. Thereby, restricting the annotation study to a single dataset is necessary due to the expensive annotation process.

A brief overview of the datasets, models and answer selection approaches investigated in this study follows. We discuss the datasets, the models' technical characteristics and point out crucial validity constraints regarding the study and the overall thesis. After that, we present the automatic and annotation study evaluation results and discuss their consequences concerning the thesis.

4.2 DATASETS

We use two educational corpora within the study. They are sufficiently different from the Wikipedia texts included in SQuAD used to train the models.

4.2.1 *LearningQ*

The LearningQ dataset [20] is based on educational video transcripts. It was crawled from the Ted-ED platform and covers a wide range of topics. The dataset contains comprehension questions written by experts for the corresponding video and a mapping from transcript to comprehension questions. For a given transcript, multiple comprehension questions are supplied in LearningQ. However, questions are not mapped to their corresponding sentences in the transcript. Therefore, performing context selection is essential when working with the dataset because inputting the entire transcript exceeds the question realization model's input capacities. Before use, we filter the dataset and remove all questions that do not end in a question mark (cloze items). Afterwards, 1,089 texts and 5,235 questions remain.

4.2.2 *Reading Comprehension Dataset From Examinations*

The Reading Comprehension Dataset From Examinations (RACE) [79] contains text from Chinese English exams. The learning materials have been used in middle and high schools in China and have been written by educators. Each text includes questions to check a reader's text comprehension. The questions are not purely literal and can require inferences made by the reader. For example, the actual answer to the questions is not always found verbatim in the text. Hence, as with LearningQ, no mapping between the questions and the text on the sentence level exist, and context selection is necessary. The dataset used in the study contained 19,944 texts and 40,349 questions.

4.2.3 *Dataset Validity and Limitations*

The pilot study is primarily concerned with investigating the linguistic capabilities of the neural question realization models on out-of-distribution educational data. The RACE and LearningQ datasets come from educational texts sufficiently different from the Wikipedia texts on which the models have been trained on. Therefore, both datasets are well suited to study model transfer to the educational domain.

However, we would like to mention three important limitations those datasets exhibit as we aim to understand the affordances of AQG in scientific reading comprehension scenarios. First, the RACE dataset is collected for reading comprehension scenarios but in a language learning setting. Hence, it also contains narrative texts not usually present in scientific reading comprehension. Second, the LearningQ dataset comprises multiple scientific texts stemming from video transcripts. Thus, their wording is likely different from written science texts. Third, both datasets do comprise literal and deep comprehension questions.

Consequently, both datasets do not perfectly reflect the nature of texts used in scientific reading comprehension and the nature of questions we aim to generate, as we primarily aim for literal questions (see Chapter 3). Nevertheless, their dataset size and sufficiently different wording from the initial training data let us assume that they still constitute valid datasets for the piloting experiments. After all, they stem from educational texts and are sufficiently different from the Wikipedia articles used for training. If at all, actual scientific reading comprehension articles are even closer to the training data distribution than these datasets. Therefore, they provide a good compromise for piloting, alleviating the need for custom dataset construction from scratch. Such a custom dataset annotation would require a tremendous effort as educational experts would need to pose thousands of questions concerning varying educational science texts. Such annotation is not only time-consuming but also encompasses considerable legal challenges regarding the textbooks' copyrights. In using RACE and LearningQ, we circumvent these problems while still having educational data for the initial AQG evaluation.

4.3 QUESTION REALIZATION MODELS

We investigate the following question realization models in the study to examine their generated questions' linguistic quality on the introduced datasets.

4.3.1 *Answer-Agnostic Question Realization*

The answer-agnostic model used is an Long Short-Term Memory (LSTM)-based architecture proposed by Du, Shao, and Cardie [36]. We selected it for the study as it was one of the first and simplest neural question realization models that considerably outperformed rule-based methods. We will briefly describe the fundamental architecture and refer to the original paper for technical details.

The model uses a word-based vocabulary of 45,000 input and 28,000 output tokens. Unknown words are marked with a special *UNK* token. The model's sentence encoder consists of a bidirectional LSTM [57] with attention, learning to represent the input sentence given as GloVe [121] embeddings. In other words, the encoder can fully read and attend to the input sequence from both sides. An autoregressive LSTM is applied for decoding. It is initialized with the encoder's last hidden state and learns to predict the next token given its hidden state and the tokens already present in the output. In the study, we use the original model snapshot as introduced by Du, Shao, and Cardie [36]. It was trained on SQuAD and achieved a BLEU-4 score of 0.12.

4.3.2 *Answer-Aware Question Realization*

The answer-aware model used in the study is UNILM by Dong et al. [32]. The model was selected because it provided state-of-the-art question realization performance measured by automatic metrics at the time of the study. In the following we provide a quick overview of the model, and we refer to the original paper for its technical details. UNILM is a transformer-based pre-trained language model that is capable of solving Natural Language Understanding (NLU) and Natural Language Generation (NLG) tasks. It relies on a subword token representation for its inputs and thus, does not require an out-of-vocabulary token. The model is initially pre-trained unsupervised on a large text corpus. In the pre-training algorithm the authors introduced a novelty by training the model with multiple different attention masks simultaneously. They applied bidirectional language modeling, left-to-right and right-to-left language modeling, as well as sequence-to-sequence language modeling masking schemes during pre-training. Consequently, the model learns to represent all these masking schemes jointly and profits from synergies between the different masking objectives. After the pre-training, the model is fine-tuned for a given task requiring a specific masking scheme. We use a model fine-tuned on SQuAD for AQG supplied by the authors, achieving a BLEU-4 score of 0.22 on the same dataset as the previous model.

4.3.3 *Model Validity Considerations*

We selected the two models based on the following reasoning. Both models differ concerning their inputs (answer-agnostic vs. answer-aware) and are also built on fundamentally different neural architectures. The answer-agnostic model relies on LSTMs, whereas the answer-aware model utilizes transformer layers. Moreover, the input and output representation of the answer-aware model is improved by using subword tokens, and the model generally encompasses more parameters to tune.

The answer-agnostic model is provided as a baseline for evaluating the educational transfer of the models. It is a relatively small and simple model. If it already transfers well to the educational domain, this provides evidence that more advanced, larger models may not be needed in every case. This is important when computational resources are scarce, like applying the models in mobile learning scenarios. On the other hand, the answer-aware question model was state-of-the-art at the time of the study. Thus, it provides an estimate of how well complex neural network question realization models transfer their linguistic quality to the educational domain.

Consequently, a direct model comparison that explains why the models exhibit varying performance is hardly possible since the models differ greatly beyond the answer-agnostic and answer-aware aspect. We will, therefore, not strive for such a direct explanatory comparison.

4.4 CONTENT SELECTION CONDITIONS

In this section, we describe the different content selection conditions leading to the two question realization models' inputs (see Figure 6).

4.4.1 *Context Selection*

The LexRank algorithm is used for context selection [41]. We selected it based on study results by Chen, Yang, and Gasevic [19] which investigate different context selection approaches on multiple educational and noneducational datasets. They used the quality of the generated questions as a comparison metric. While LexRank did not produce the best results on all datasets, they suggest that LexRank is the most promising context selection approach across datasets. In the present study, the algorithm selects up to four sentences as a summary of every paragraph, or less if the paragraph contains fewer sentences. We apply the same LexRank configuration as the related work [19, 134].

4.4.2 *Answer Selection*

The answer-aware question realization model requires the expected answer as input. Therefore, we explore four different answer selection strategies, extracting various sentence parts as expected answers for the question to be generated (see Figure 6).

Random Condition

The simplest answer selection strategy applied selects a random word from the given sentence as the expected answer. The sentence is word-tokenized using Stanford CoreNLP [97]. It is clear that this strategy is not particularly helpful for the downstream question realization model, as the selected answer is unlikely relevant for an educational question. Therefore, we consider this strategy as a lower bound for the capabilities of the answer-aware question realization model, independent of its actual answer-awareness. In case the model still produces higher-scoring metrics than the answer-agnostic model when confronted with random input answers, it is likely that the underlying neural architecture is superior even when ignoring its advantage due to the additional inputs.

Subject Condition

The *Subject* condition selects the grammatical subject of the context sentence as an answer candidate for the upstream question realization. A sentence's subject is frequently the executing agent of the primary action of the sentence. That is, posing questions about the subjects may provoke learners to recall what initiated the action described in the given sentence. Moreover, subjects are frequent in sentences and extracting them allows us to construct a question for many context sentences. Hence, selecting the sentence's grammatical subject is plausible from a learning perspective and a text coverage perspective, as we want to generate questions about all relevant sentences detected by context selection.

We implement this strategy via dependency parsing and Semgrex matching [16] on the provided context sentence. A dependency graph describes the grammatical structure of a given sentence (see Figure 7). Semgrex matching applies regular expression-like patterns to the constructed dependency graph, extracting specified subcomponents or subgraphs. These patterns are almost as flexible as implementing the graph traversal code by hand but are less complex to write and easier to adapt. One pattern can result in multiple matching subject subgraphs because one sentence may encompass multiple subjects. For instance, the sentence: *"The computer executed the program while the network card was waiting for packets to arrive."* comprises the subjects *"The computer"* and *"the network card"*. We select the subject with the most characters as the expected answer in these cases under the heuristic that longer strings comprise more information.

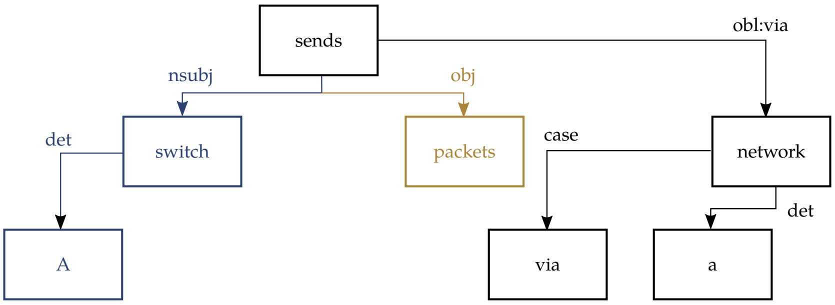


Figure 7: An example of a dependency tree for a given sentence. In this study we focus on subject (blue) and object (gold) subgraphs of the dependency graph.

Direct Object Condition

The applied *Direct Object* condition selects direct objects from the context sentence via the same technical approach as the *Subject* condition. We select direct objects for similar reasons as grammatical subjects. While the subject relation often indicates the actor of a sentence, the direct object is frequently associated with the entity that receives the executed action. Thus,questions posed regarding the direct objects should serve different purposes than questions regarding the subjects of sentences. Moreover, sentences also frequently comprise direct objects yielding a high coverage again for this answer selection condition.

4.5 AUTOMATIC EVALUATION RESULTS

The descriptive automatic evaluation metrics are reported as an initial characterization of the generated questions.

4.5.1 *Number of Generated Questions*

The four experimental conditions lead to a different numbers of generated questions per dataset (see Table 2). For instance, the answer selection in the *Subject* condition was successful for 92% of all context sentences, whereas the answer selection in the *Direct Object* condition was successful for 67% of all context sentences in RACE. In cases where the answer selection failed, the corresponding automatically generated dependency graph either contained no grammatical subject or direct object node. It induces a slight selection bias in ignoring sentences where the dependency-based answer selection could not find a target answer. Yet, for the estimation of linguistic quality, such a bias is not too dramatic as we look at the sentences relative to the selected answer. That is sentences without a selected answer neither provide a positive nor negative signal to the evaluation metrics. However, during the interpretation of the results, one has to keep in mind that there are exceptional sentences where we were unable to estimate the linguistic quality of the system with any of our conditions.

Dataset	Condition			
	Answer-Agnostic	Random	Subject	Direct Object
RACE	79,558	79,558	72,870	53,635
LearningQ	4,340	4,340	4,016	2,931

Table 2: Number of generated questions per dataset and condition.

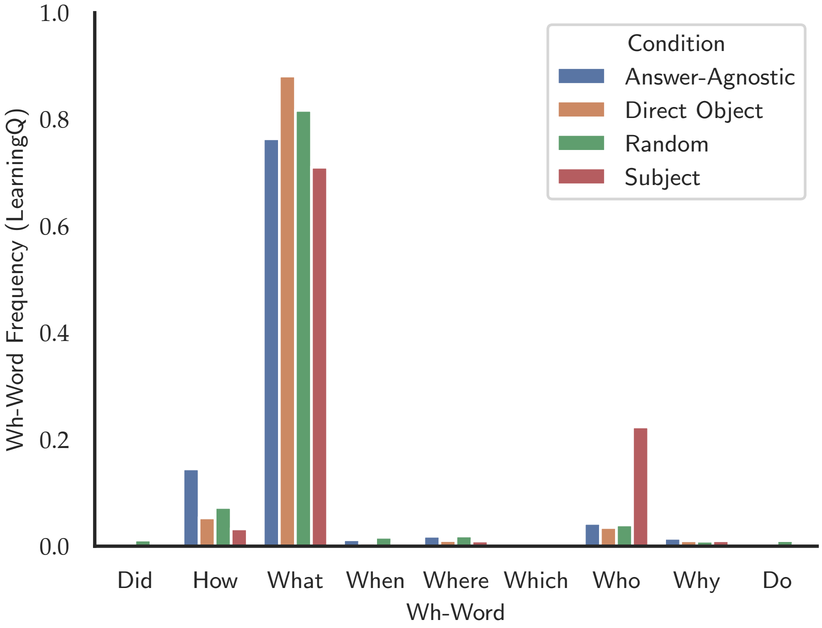


Figure 8: Distribution of Wh-Words on LearningQ. The y-axis shows relative frequency and ranges from zero to one. The respective RACE plot can be found in Appendix A. This figure has been adapted from Steuer, Filighera, and Rensing [153].

4.5.2 *Type of Questions*

The Wh-word distribution on the LearningQ dataset for the different conditions is shown in Figure 8. The models most frequently generate *What* questions. Furthermore, questions starting with more complex Wh-words are rarely generated by the different conditions (e.g. *Why* or *How*). This initial evidence suggests that the generated questions focus more on literal than deep comprehension questions. Besides this, the distribution also implies that the answer selection conditions influence the generation process. The *Subject* condition more commonly results in *Who* questions than the other conditions. Moreover, the answer-agnostic generator focuses more on *How* questions than the other generators. The corresponding plot for the RACE dataset validates these results by expressing similar patterns and can be found in Appendix A.

Dataset	Condition			
	Answer-Agnostic	Random	Subject	Direct Object
RACE	0.04	0.05	0.05	0.05
LearningQ	0.08	0.09	0.08	0.06

Table 3: BLEU-4 scores on the different datasets for the different conditions only considering the maximum scoring question per paragraph.

4.5.3 *BLEU-4 Analysis*

We calculate BLEU-4 score assuming it serves as a rough estimate for the grammaticality of the generated questions (see also Section 2.2.5). The reported BLEU scores fall between zero and one, where zero indicates no overlap with the reference question and one indicates perfect overlap with the reference question. Related work frequently uses BLEU-4 scores between the generated questions and human reference questions, to measure AQG performance (e.g. [32, 36, 167]), and thus we report them for reference. However, both datasets lack a mapping between input sentences and reference questions and thus require context selection. Therefore, we assume that all reference questions of a paragraph are meaningful references for any generated question of that paragraph. In other words, all reference questions for a paragraph are used to compute the BLEU score of any of the corresponding paragraph questions.

The BLEU related analysis relies on 4-grams and computes a the generated questions overlap with the references using BLEU-4 scores on both corpora. It only uses the maxium scoring question of each paragraph to compute the corpus BLEU-4 score similiar to Chen, Yang, and Gasevic [19]. With this evaluation method, paragraphs achieve a high score as long as the context selection has selected at least one promising sentence that results in a question similar to the reference questions. This mitigates the effect of imperfect context selection since each generator has the chance to generate questions on four selected sentences. However, it also overestimates the BLEU-4 score as we keep only the maximum score from each paragraph. We achieve the BLEU-4 scores shown in Table 3.

However, this analysis is rather noisy and does not provide a clear picture of the actual system performance, as Figure 9 indicates (see Appendix A for the RACE plot with similiar results). The figure shows the mean paragraph BLEU-4 score distirbution. It is clear from the violin plot that the scores are zero on average and rarely deviate upward. In fact, all conditions on both datasets have a median paragraph score of zero. In other words, reference questions and generated questions usually do not overlap. That is, it could be that the content selection prefers dissimilar question topics to the datasets' authors. However, this does not necessitate that the generated questions lack linguistic or educational quality. Questions may target other information and are still linguistically pleasing and improve learning. Therefore, we investigate this in a follow-up expert-based annotation study on a sample of questions generated on the RACE corpus in the next section.

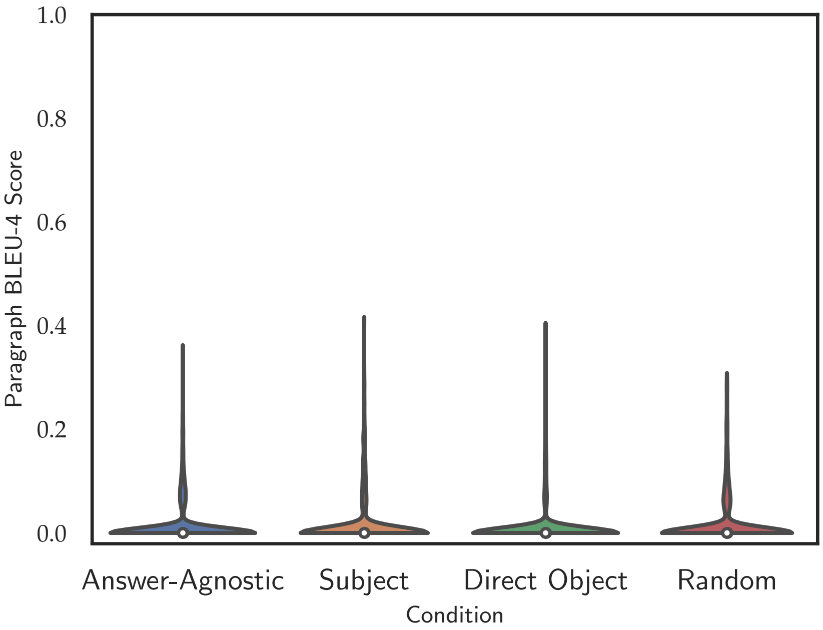


Figure 9: Violin plot of the distribution of the mean paragraph BLEU-4 scores on LearningQ. The plot visualizes the data median (white circle) and the score's distribution in the four conditions with the four colored distribution areas. Each colored area is symmetrically around it's condition, and the wider an area, the more data points fall in the corresponding range. We can see that the BLEU-4 scores are mainly close to zero in all conditions, with only occasional outliers. This figure has been adapted from Steuer, Filighera, and Rensing [153].

4.6 ANNOTATION STUDY DESIGN

The automatic evaluation yielded two main insights. First, the generated questions appear significantly different from the references. Second, the generated questions Wh-words suggest they likely only target short, factual answers, not deep comprehension. We conduct an expert-based annotation study on a subset of the RACE data to complement and validate these insights gathered by automatic evaluation. This is particularly important since we are mainly interested in the questions' linguistic quality, and the low BLEU-4 may stem from two reasons. They may result from valid but differently expressed questions or from linguistically flawed ones.

Therefore, we annotate 240 questions on the RACE dataset in three dimensions using two annotators assessing each question's quality. Both annotators have sufficient English language skills for the texts from the RACE corpus. One speaks English on B2 level according to the Common European Framework of Reference for Languages (CEFR) and the other is a native speaker. The annotators read 80 paragraphs with three questions per paragraph for 240 questions per annotator. To reduce order effects, both annotators see the paragraphs in random order. Each annotator first reads the text and then evaluates the questions shown together with the context sentence.

The annotation study measures the following three important question quality criteria. First, the questions' acceptability is measured on a five-point Likert scale. A value of one is assigned to an unacceptable question, and a value of five indicates a syntactically pleasing question. The acceptability measure is closely related to grammaticality and spelling quality criteria(see Chapter 3). We prefer it, instead of asking for these estimates directly, because it integrates both into one scale and does not require a formal grammar analysis from the annotators. This scale is most important for the study's purpose because it provides us a human judgement of the models' linguistic quality on the educational corpus. Second, the answerability of the questions is annotated on a five-point Likert scale. A value of one means unanswerable, and a value of five indicates the question is unambiguously answerable with just text knowledge. Third, the usefulness of the questions for a reading comprehension scenario is annotated on a three-point Likert scale. A value of one means the question is not hepful, a value of two means the question could be beneficial but does not necessarily target important knowledge, and a value of three indicates a useful reading comprehension question. While the first scale provides an estimate of the generated questions' linguistic quality, the remaining two scales provide initial evidence to which extent the generated questions foster reading comprehension as they are. To support text comprehension, we assume that questions must be readable, answerable, and relevant.

A shared annotation guideline is provided to both annotators to ensure a common understanding of the three annotated dimensions. The Inter-Annotator Agreement (IAA) was Krippendorff's $\alpha = 0.63$ for acceptability, Krippendorff's $\alpha = 0.78$ for answerability, and Krippendorff's $\alpha = 0.55$ for usefulness. We resolved conflicts by preferring the annotation of the native speaker.

4.7 annotation study results

The annotation study results are shown in Figure 10 and Figure 11, example questions are provided at the section's end in Table 4. We start by discussing the acceptability results.

4.7.1 *Acceptability Results*

Both question realization algorithms frequently generate linguistically acceptable questions. The median acceptability rating for the answer-agnostic approach is four, and all answer-aware conditions even score five points for acceptability in the median. This result seems counterintuitive given the BLEU results of the automatic evaluation. There it appeared that most generated questions scored low and are thus unacceptable. Hence, it provides us with the insight that for non-sentence-level references, BLEU scores in AQG do not correlate with human ratings of linguistic acceptability. In fact, the correlation coefficient between the mean paragraph acceptability scores and the mean paragraph BLEU-4 in all conditions is low (Pearson's r=0.13, p=0.26, n=80). This validates the findings of Liu et al. [85] in the context of AQG, which already have shown that BLEU-4 scores do not correlate well with human ratings in AQG.

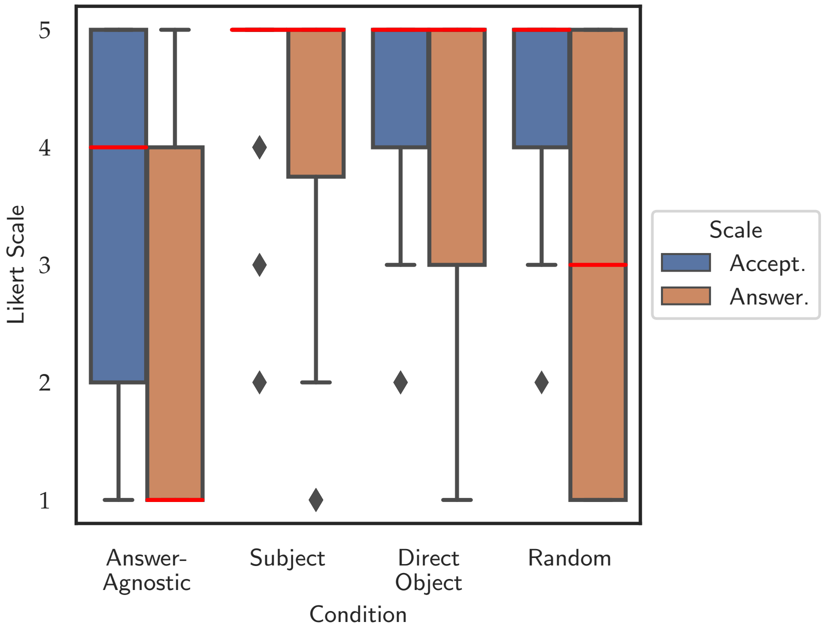


Figure 10: Expert ratings for the annotated acceptability and answerability scales. The red bars indicate the median, the whiskers the 1.5 interquartile range and diamonds outliers. This figure has been adapted from Steuer, Filighera, and Rensing [153].

When comparing the two question realization approaches with respect to acceptability, the comparison indicates that the answer-aware architecture performs better than the answer-agnostic architecture in all four conditions. The median acceptability rating differs by one point, yet the answer-agnostic architecture has a wider spread in its ratings and generates considerably more unacceptable questions. In particular, even in the *Random* condition, the answer-aware architecture generates more acceptable sentences. We interpret this as evidence that the architecture's better generation capabilities are not attributable solely to the additional input of the answer but that the higher parameter count and different model structure provide a general advantage.

Furthermore, it becomes clear from the acceptability scores that the additional answer input affects the acceptability of the generated questions in the answer-aware conditions. The *Subject* condition generates frequently linguistically acceptable questions with a very narrow spread of the acceptability scores. In contrast, the *Direct Object* and *Random* conditions produce fewer perfectly acceptable questions.

4.7.2 *Answerability Results*

The answerability of the questions varies depending on the condition (see Figure 10). In contrast to acceptability, not all systems achieve good or excellent scores in the median. The answer-aware approach leads to better answerable questions than the answer-agnostic approach, even in the *Random* condition. One reason is that questions with a low acceptability score are also unlikely to score high in answerability because they

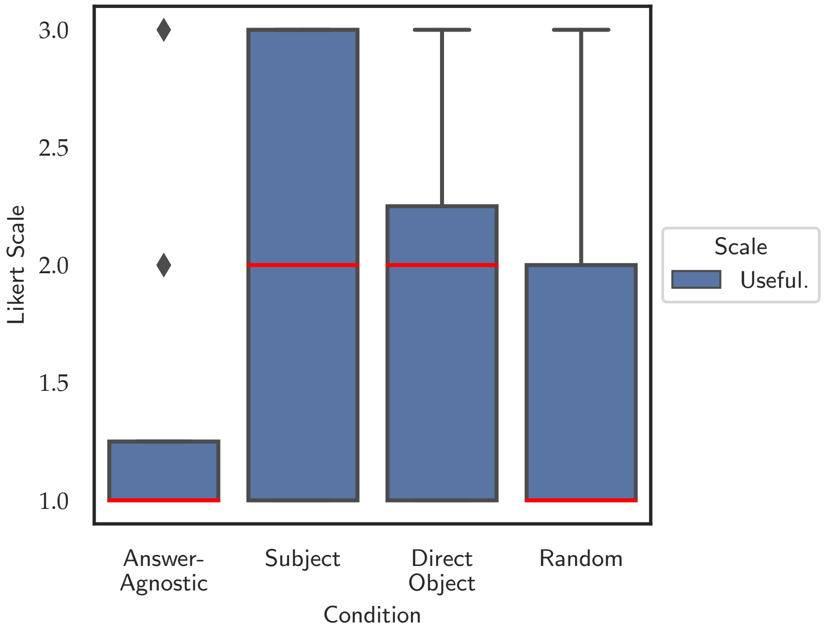


Figure 11: Expert ratings for the annotated usefulness scale (adapted from Steuer, Filighera, and Rensing [153]). It is structured as the grammar and answerability plot.

are difficult for the annotators to understand. Moreover, the *Subject* and *Direct Object* conditions result in more answerable questions. Thus, even on the out-of-distribution data of RACE, the answer-aware generator utilizes a given valid answer correctly and poses meaningful questions about them.

4.7.3 *Usefulness Results*

The usability scale gives us an initial estimate of the kind of reading comprehension questions that were generated. As in the other two dimensions, the answer-agnostic system is inferior. In fact, it generates hardly any meaningful questions for reading comprehension. The three answer-aware conditions achieve significantly better scores in the median. Only the *Random* condition questions receive a median score of one, indicating their uselessness for any educational practice. Annotators give both the *Subject* and the *Direct Object* condition a median score of two. Hence, according to the annotation guideline, the questions have a textual reference and target facts in the text, but these are not central to understanding the content of the text. This rating is congruent with the distribution of Wh-words in the automatic evaluation, which already points towards literal questions, mainly asking to recap short facts from the texts. From the perspective of the question realization algorithms, this is interesting because the algorithm generates meaningful literal questions given the provided inputs. From the content selection perspective, LexRank probably did not select the central information in the given text. Nevertheless, this study provides initial evidence of good acceptabil-

Condition	RACE	LearningQ
Random	Are many people finding it increasingly difficult to keep close relationships ?	What is one of the options for who goes across first?
Answer-Agnostic	What did Paten documents for the design describe it ?	What is the name of the feature that some entertainment on board the scuffling buffalo was?
Direct Object	What do many students see as a campus tradition ?	What is removed from a jellyfish ′ s skin to prevent it from firing?
Subject	What will provide safe driving control and route finding ?	What happened to the water connection between the pacific and the caribbean?

Table 4: Sample questions generated by the four conditions used in the pilot study.

ity and the answerability of the generated questions in the answer-aware conditions. However, the results also suggest that the generated questions' educational usefulness is still mediocre and adaptations to the overall educational AQG process have to be made to generate reading comprehension questions (see Chapter 5, Chapter 7 or Chapter 8).

4.8 DISCUSSION OF THE MAIN INSIGHTS

In summary, the following findings emerge from the pilot study. First, the linguistic quality of neural question realization algorithms on out-of-distribution data is reasonably high according to our annotators' ratings. After all, both the answer-agnostic and answer-aware conditions perform well in the median for acceptability. Additionally, the answer-aware conditions only rarely deviate to the lower end of the acceptability scale. Thus, despite the different text genres and structures present in the educational data, the question realization algorithms were nonetheless able to perform the syntactic transformation of the inputs into a plausible question. Therefore, the results support our hypothesis that neural question realization algorithms can be used in the educational domain without significantly lower linguistic quality due to the out-of-distribution inputs compared to their training data. Evidence for the hypothesis is encouraging in light of the second research question. According to our analysis in Chapter 3, solid linguistic quality forms the foundation for reading comprehension questions. Given the study's data, the investigated neural question realization systems have the potential to provide this foundation for educational Automatic Question Generation, highlighting their linguistic strengths with respect to the second research question.

Second, unfortunately, the BLEU scores only weakly correlated with the human annotation study, highlighting the weaknesses of automatic evaluation in Natural Language Generation and Automatic Question Generation. Thus, the LearningQ dataset's results must be interpreted with caution since we only collected BLEU scores and did not perform any human evaluation. The data support the results of Nema and Khapra [110] and Liu et al. [85], which already showed that BLEU scores often only allow a very noisy evaluation of Natural Language Generation and Automatic Question Generation systems. In particular, after this analysis, we assume that we will always need to include a human annotation study to discuss further results of this thesis. More generally, concerning the second research question, we learn from the experiment that works reporting only automatic scores are likely to provide poor insight into educational Automatic Question Generation characteristics even on the linguistic level. That is, common automatic scores may not measure the extent to which generated questions are helpful for reading comprehension, even on the linguistic level.

Third, according to the study, linguistically valid questions are generated. Yet, the annotators do not judge them to be particularly helpful for reading comprehension. Even conditions that perform very well linguistically receive a median rating of two out of three on the usefulness scale. A score of two indicates questions about auxiliary facts are mainly generated instead of targeting the texts' gist. Given the mediocre score, the content selection may hinder the question realization algorithms, providing inputs unsuited for education. Therefore, we will investigate content selection mechanisms for the neural question realization algorithms that explicitly target educationally relevant text contents (see Chapter 5, Chapter 7 or Chapter 8). That is, the study's results also provide further motivation to investigate the first research question. Moreover, the usefulness score scatters significantly across the scale. Possible reasons for this include that the annotators generally have difficulties assessing the usefulness of a question without a concrete learning scenario. Such annotation difficulties are a limitation of expert annotation studies. Due to the costs of larger-scale annotation projects, studies with only a few annotators are common in Natural Language Generation research [165]. However, they have the validity constraint that they usually do not reach a very high agreement on the annotated scales [165]. Thus, while their judgments are more reliable than automatic evaluation metrics such as BLEU, they do not necessarily generalize to a larger readership. Consequently, we conclude that an extrinsic learner-based evaluation is essential for investigating the second research question as it evaluates the questions' quality in a concrete learning scenario, providing complementary more objective measures on the usefulness of the questions for reading comprehension. Hence, future evaluations in this thesis will include studies with learners in addition to expert assessment (see Chapter 8).

5 EDUCATIONALLY MOTIVATED CONTEXT SELECTION

According to the previous chapter, neural question realization frequently generates plausible and linguistically pleasing questions. Thus, those systems may have great potential for educational Automatic Question Generation (AQG). However, the insights gathered in the pilot study also hint toward the importance of content selection approaches. After all, another finding of the piloting study suggests that the generated questions often are not focused on helpful text content to foster reading comprehension. The content selection process comprises context selection and answer selection. This chapter addresses context selection approaches, whereas subsequent chapters investigate answer selection approaches. Our primary goal in this chapter is to propose a context selection approach that extracts learning-relevant sentences fostering reading comprehension. Thus, the design, implementation, and empirical evaluation of such a context selection approach is described, addressing the first research question. In contrast to previous works, we introduce the following two novelties.

First, we do not rely solely on implicit machine learning for context selection, but define what constitutes relevant information a priori and not implicitly through the dataset. We then purposefully collect or choose datasets to represent these relevance criteria during our model training. Second, we train a recent state-of-the-art contextualized embedding model as a classifier on the collected datasets, exploiting its ability to capture part of each sentence's semantics [105]. The classifier acts as a context selector, using the parts matching the a priori defined question-worthy criteria as contexts' for AQG. Put together, this addresses the research gap identified in Chapter 2, where we found that machine learning algorithms for context selection are powerful, but should ideally work according to educational criteria. We will see in the result and discussion of this chapter that such an approach yields a competitive context selection with similar performance characteristics as other state-of-the-art context selection mechanisms. Additionally, the proposed context selectors are grounded in educational theory.

The study in this chapter illustrates and evaluates this approach on two relevant context types: definitions and causal sentences. We use these sentence types for our initial study because they are highly relevant to text comprehension according to educational literature (see Chapter 3). However, the general idea is extendable to other sentence types. For instance, Graesser, Rus, and Cai [48] state that questions targeting examples or comparisons may also support learning. Hence, comparing or exemplifying sentences may also make a valuable context selection target, and the proposed approach can easily be extended to such novel sentence types.

Investigating the two illustrative context types requires training two context selection classifiers. Both classifiers are trained on respective German definitions and causal sentence corpora, allowing us to conduct a direct reader-based evaluation with native speakers of their context selection quality. This approach shares some similarities with Stasaski et al. [147] in that they also investigate the use of causal sentences as contexts. However, it has been developed in parallel. Furthermore, instead of evaluating the resulting questions, we empirically assess the actual context selectors' performance with an evaluation study in German and experiment with an additional sentence type. Thereby, we contribute more evidence to the inconclusive research literature on context selection, which is frequently evaluated only on the upstream question realization task but not individually (see Section 2.4).

In the following, we will outline the challenges involved in definition and causal sentence extraction. Next, we describe our corpus construction and model training procedures for the two German context selection classifiers. Finally, we discuss insights from an empirical study investigating the proposed context selection mechanism.

5.1 DEFINITION & CAUSAL SENTENCE SELECTION

We have motivated in the introduction of this chapter that we will aim to select definitions and causal sentences as contexts. Next, we address the extraction of these semantic sentence types from texts and the challenges involved therein.

5.1.1 *Extraction Challenges*

Context selection via definitions and causal sentences needs to detect the respective sentence types in text. In both cases, it becomes difficult to explicitly characterize the sentence type [129, 143]. For causality, there are competing theories on how to define causality, and humans mostly rely on their intuitive understanding of the concept [129]. Concerning definitions, different types exist [49]. However, these formal types are often relatively rigid and hardly support their automatic detection. For instance, an intensional adequate definition [49] for a concept has no counterexamples. Yet, it is impossible to check if the given conditions in the text are actually intensional without having world knowledge. For these reasons, the definition and causal extraction literature frequently relies on machine learning to determine what constitutes an intext definition or causal sentence based on linguistic patterns. That is, no formal specification for causal sentences or definitions is applied. Instead, sentences in a ground truth dataset determine the linguistic patterns expected in definitions or causal sentences.

However, even implicit, text-based characterization is challenging. Take the brief definition:

> *"TCP (Transmission Control Protocol) refers to the connection-oriented transport protocol used in the Internet suite as specified in RFC2012."*

It exemplifies definitions with an explicit marker phrase (*"refers to"*). However, these markers are not required and other definitions implicitly characterize a concept without relying on a particular marking phrase. Causal sentences frequently comprise explicit markers like *"because"*, as in the sentence:

> *"UDP seems fast because it sends data with no expectation that it is ever received."*

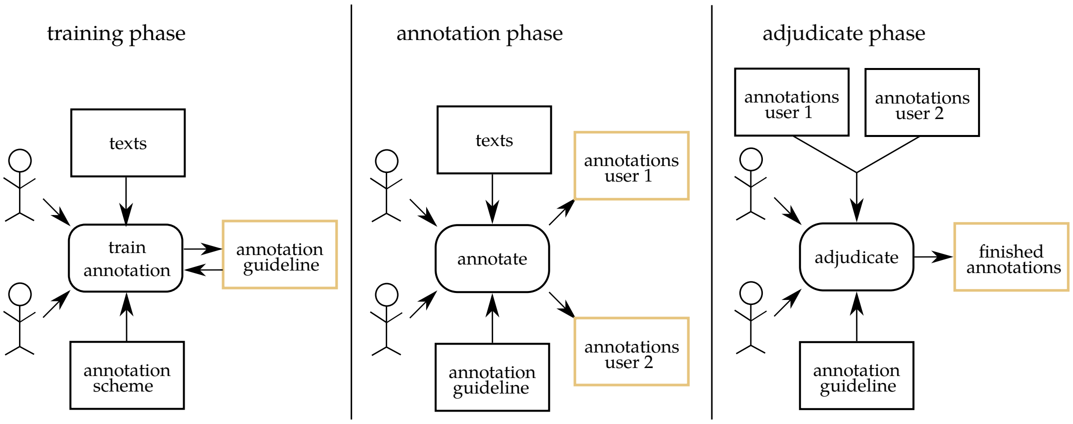


Figure 12: The corpus construction process. The rounded boxes indicate processes, the black boxes input artifacts, and the colored boxes indicate the output of the corresponding process step.

Yet, some causal sentences implicitly express the causal relation without markers. Moreover, classifying explicitly marked definitions and causal relations is challenging because the marking phrase used is not always an indicator of a definition or causal sentence. As a simple example, the *"refers to"* in the following sentence is not introducing a definition:

> *"In the lecture, Andrew Tanenbaum refers to the TCP/IP protocol."*

Thus, linguistic research indicates that pure rule-based detection is insufficient for both sentence types. Instead, contextualized embeddings perform better on the respective English sentence binary classification task for definitions [142] and the respective classification task for causal sentences [129, 181]. Therefore, we will rely on these for our causal and definition extraction in the proposed context selectors.

5.1.2 *Corpus Construction*

The previous section states that the extraction tasks are usually implicitly defined via a large corpus of annotated definitions and causal sentences and are addressed by transformer-based machine learning models. We transfer this approach to the educational domain and train those context selectors in German to alleviate language effects as cofounding factors during our context selection evaluation experiments. For this reason, we collect two corpora on German textbooks in the scientific domain utilizing the following annotation schemes and process.

Annotation Schemes

We utilize the following annotation schemes for definitions and causal sentences.

Definitions

We adapt the definition annotation scheme from the Definition Extraction from Texts (DEFT) corpus [143], which is an extensive English corpus for definition extraction. We use it because it describes the important categories of definitions and has produced sufficient Inter-Annotator Agreement (IAA) during previous use. The schema comprises token-level annotations covering multiple facets encountered in definitions like alias terms, multi-sentence definitions or qualifiers. In our study, we focus on single sentence definitions because they are easier to annotate and we aim to phrase the context selection task as sentence classification problem. Therefore, we simplify the scheme slightly (see Appendix Section A.2 for the schemes description). A sentence is treated as a definition for the binary classification task if it contains a term and a definition tag.

Causal Sentences

The causal sentence classification task has been addressed in various English works [43, 136, 180, 181]. Authors annotate linguistic markers, which often introduce cause and effect [45, 129], which they call trigger words. Only sentences comprising trigger words are considered as potential causal sentence candidates and other sentences are not considered for classification. Furthermore, a large-scale token-level German causal classification corpus based on the trigger word concept was recently introduced [129]. It is based on newspaper articles and political speeches. Although it is a German corpus, we decided against using it solely because its domain comprises legal sources that utilize language constructs foreign to science texts. However, we use a similar annotation scheme because Rehbein and Ruppenhofer [129] have shown that their annotation scheme allows the learning of causal sentence classifiers (see Appendix Section A.2 for the schemes description). Yet, science texts rely on different language constructs, leading us to collect science-specific trigger words during annotation. We start with the initial trigger word list from Rehbein and Ruppenhofer [129] and extend it iteratively based on frequently occurring novel trigger words in the annotation process. For a word to be added, all our annotators must agree that it frequently signals causality in the gathered examples. Consequently, the science-specific triggers allow us to detect more candidate sentences. Finally, during the binary classification task, every sentence containing a trigger, cause and effect is treated as causal.

Annotation Process

Given these annotation schemes, we annotated texts stemming from 20 German science textbooks for university students published by *Springer International Publishing*. We deliberately chose textbooks from various scientific domains ranging from psychology to glaciology to cover their multifaceted language. Four student assistant annotators with educational or linguistic expertise worked in the project and received monetary compensation. The annotation process followed common best practices [122] and comprised a training, an annotation, and an adjudication phase (see Figure 12). In the training phase, annotators were trained in the annotation schemes, and a more detailed annotation guideline document was derived describing how to handle difficult annotation cases. After the training, we measure the IAA via Krippendorff's α, which is a chance corrected IAA measure between zero and one, where one indicates perfect agreement. The annotators achieved a Krippendorff's $\alpha = 0.78$ on causal relations and $\alpha = 0.53$ for definitions. The annotation phase and the adjudication phase were iteratively executed. In the two-week annotation phase, annotators independently annotated text and every document was annotated twice. After that, annotators got together and resolved all annotation conflicts based on the provided annotation guidelines in a one-week adjudication phase. The overall annotation process took five months.

The resulting corpus yielded many causal sentences (N=1,460) but comparatively few definitions (N=375). We increase the training data size for the definition classification using data augmentation. Thereby, we rely on machine translation, for which we have shown that it is a valuable data augmentation technique for AQG [149]. Hence, we deemed it worthwhile to translate the DEFT corpus using the *WMT19* [112] translation model to German. The final corpus contains 5,641 German definitions and 1,460 German causal sentences. That is, the translation data represents a huge portion of the definition corpus. Yet, a qualitative review of the translated sentences indicated that their quality is sufficient to keep them in the final corpus. Finally, for model training, we sample equally many sentences not expressing a definition or causal relation from the textbooks as negative examples.

5.1.3 *Model Training And Testing*

We use the collected corpus to train one BERT-based [30] context selection model for each of the two sentence types. BERT-based models are selected because they are widespread, have shown excellent performance on various tasks [30] and come in a dedicated German variant [17]. The corpus is divided into 70% training data and 30% testing data for training. We do not perform hyperparameter tuning and train the models for five epochs because the initial BERT authors found this sufficient for many fine-tuning tasks [30] . We select the last model after five epochs for our evaluation study. The test set scores for the definition selection model are Precision=0.80, Recall=0.80 and F1=0.80 and Precision=0.78, Recall=0.78 and F1=0.78 for the causal selection model.

5.2 EMPIRICAL EVALUATION STUDY

Although the proposed context selection mechanisms are grounded in educational theory (see Section 3.4.1), their actual selection performance in learning scenarios is unclear. Therefore, this chapter investigates the proposed context selection mechanisms in an empirical evaluation study without proxy evaluation via the generated questions (RQ 1).

5.2.1 *Study Design*

An evaluation study with 37 participants is conducted. The study design is inspired by the study of Dee-Lucas and Larkin [27], where the factors learners perceive as relevant in texts were investigated. In our study, participants read five short texts between 30 and 50 lines stemming from biology, physics or psychology. Participants select the six sentences they perceive as most important in every text. We analyze the so collected data from two perspectives:

- A learner perspective
- A state-of-the-art comparision

On the one hand, the learner perspective investigates human performance on the task and if the results of the proposed context selection methods and the learner-selected sentences are associated. That is, if the proposed context selection methods help to recover learners' selected sentences.

On the other hand, the state-of-the-art comparison explores how the proposed context selection performs compared to selection criteria from related work on the collected study data. That is, it aims to evaluate if the proposed methods result in a competitive context selection mechanism.

5.2.2 *Data Collection*

The data of the evaluation study was sampled from the local university by voluntary sampling over all students with sufficient German language skills. Participants mean age is 26.8 years (SD=9.64), and the sample comprises 23 female and 14 male subjects. All except two participants were German native speakers, and the remaining two spoke it fluently on an expert level. Experiment participation was entirely online and required participants to work roughly 30 minutes under quiet working conditions. Participants received contact details for debriefing and inquiries about the collected data. All submissions were completely anonymous.

Before the experiment, detailed instruction outlined the learning scenario and the task. It was made clear that learning relevance is a subjective concept and that the selection decisions are not graded in any form. When the experiment started, reading comprehension materials contained introductory material from online learning resources and the German Wikipedia. The texts were deliberately selected such that anticipated readers from the sample would most likely comprehend them but were unlikely to have detailed knowledge of the subject matter. We chose this design to imitate a typical learning scenario where novices learn from scientific reading materials. In total, 37 participants took part in the study and selected six sentences in each of the five texts. The reading materials had 182 sentences in total. Thus the final dataset encompasses 182*37=6,734 sentence selection decisions and 1,110 importance ratings.

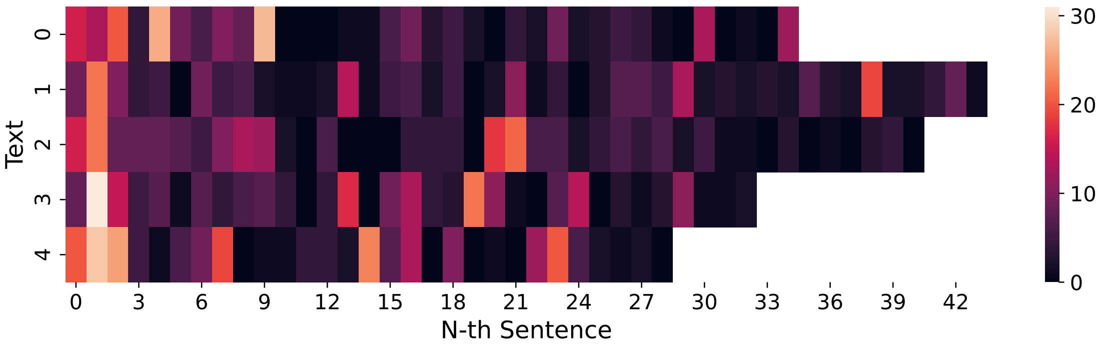


Figure 13: A heatmap showing the sentence selection count over the five texts. The more brightly colored a sentence is, the more often it was selected by participants. White tiles indicate that the text has ended, e.g., the text with identifier four only has 28 sentences.

5.3 RESULTS

We report results from a learner perspective and a state-of-the-art comparison perspective, starting with the learner perspective.

5.3.1 *Learner Perspective Results*

First, we discuss human task performance and then the proposed context classifiers' task performance.

Human Task Performance

First, we look at the participants' sentence selection characteristics from a learners' perspective because it depicts how difficult respectively subjective the task for humans is. Participants do not always agree on what constitutes a learning-relevant sentence. Figure 13 illustrates that the different participants spread their selection over large parts of the texts. They usually do not unanimously agree on what constitutes a given text's relevant or irrelevant information. However, a few key sentences are selected frequently in every text. If the percentiles of the selections are assessed, a similar picture emerges. There are 15% of sentences that were never selected as relevant given the 37 participants and the 182 possible selectable sentences. There is no single sentence that is selected as relevant by all participants and sentences are in median selected four times. In total, 25% of the sentences were selected once or less. Additionally, 75% of sentences received between zero and eight selections. What this indicates is that the task already causes considerable disagreement between human participants. Thus, we cannot expect any context selection algorithm to successfully recover every single selection prediction. Instead, only a weak association between human selection and context selection decisions should be expected. Nevertheless, all texts contain a few frequently selected sentences, which are perceived as relevant by many readers and some rarely selected sentences, which are perceived as irrelevant by most readers.

Type	Learning-relevant [%]	In All Sentences [%]	Ratio
Cause	54	43	1.26
Definition	55	36	1.53

Table 5: The relative ratio of definitions and causal sentences in the learning-relevant sentences and in all sentences.

Proposed Context Selector Performance

Next, we discuss the overlap between the proposed context selection methods and the sentences deemed learning-relevant in the study. According to our classification models (see Section 5.1.3), the 182 sentences in the text comprise 65 classified definitions and 79 causal sentences. The overlap between the sentence types is 32 sentences which comprise a definition and a causal statement. Consequently, 112 sentences have been classified at least once and the remaining 70 sentences did not receive a classification. Descriptively, definitions and causal sentences appear disproportionately often in the 1,110 learning-relevant selections of the participants (see Table 5). The effect is more pronounced for definitions which appear 1.53 times as often in all selected learning-relevant sentences than in all sentences.

Given these data points, initial descriptive evidence supports that classified definitions and causal sentences are valid context selection targets, perceived as learning-relevant. Moreover, the low agreement between readers shows that the overall task is subjective. Individual preferences beyond the plain text characteristics influence selection decisions. Thus, we cannot expect highly accurate selections from textual characteristics alone.

5.3.2 *State-of-the-Art Comparision*

Next, we describe the results from the state-of-the-art comparison, in which we compare different context selection methods' performances to recover participant selections. Thereby, we focus on the sentences having a higher agreement between the annotators. We believe correctly classifying them is particularly important, as human annotators agree on their relevance or irrelevance. Thus, a context selector with a strong performance on these sentences will more likely appeal to multiple learners. Hence, the raw participant data cannot be used directly as ground truth for comparison, and a ground truth dataset needs to be constructed.

Ground Truth Dataset

The upper and lower 25% sentence selection percentiles have less than two, respectively, more than eight selections. Therefore, we assume that sentences selected only once or less can be regarded as negative examples and sentences selected eight or more times can be regarded as positive selections with sufficient agreement. We transform the corpus to only contain these sentences for the context selectors' comparative evaluation. The resulting ground truth contains 101 data points, with 51 learning-relevant and 50 irrelevant data points.

Note that this dataset transformation changes the classification task considerably to allow us to measure the context selectors' performance on the high agreement selection cases. In real-world learning scenarios, selections are likely imbalanced and as nuanced as in the raw study data. Hence, the ground truth evaluation represents a sort of best-case scenario for the context selectors, which we are interested in because the high-agreement cases will likely yield the most value for the AQG task. Consequently, the reported values should be used exclusively comparing the detection of high agreement cases. On the general data, all of the compared context selectors likely perform worse as the data is more complex and even humans frequently disagree with their selections.

Given this ground truth as a corpus, we compute precision, recall, and F1 score for the evaluated context selection algorithms on the ground truth. Although we only regard the ground-truth sentences for our evaluation, we apply the context selectors to the complete text because some algorithms require the full-text structure during computation. Note that this leads to counter-intuitive selection sizes for fixed-size extraction algorithms. Algorithms extracting a fixed number n of the five text's sentences will not necessarily yield $n \cdot 5$ classifications on the ground truth because we discard all classifications on sentences not present in the ground truth. To measure this effect, we are interested in how many high agreement sentences a context selector yielded compared to its overall selections. We report its *high agreement prediction ratio*. The value is the fraction of all ground truth predictions divided by all classifier predictions. The closer to one, the more the context selector focuses on ground truth sentences.

Compared Context Selectors

In total, we compare context selectors from three different algorithmic categories. First, we include two unsupervised summarization algorithms in the comparison: *LexRank* [41] and *SumBasic* [111]. They are included because related works have used them as context selectors in AQG [6, 19, 134]. Particularly, LexRank has been proposed as a valid context selector for multiple corpora [19]. Second, we include two text heuristics: a *length heuristic* and a *position heuristic*. They have achieved surprisingly good results in related works [19]. Moreover, both are intuitively a proxy measure for learning relevance. Authors frequently structure their paragraphs with the key ideas stated at the beginning [70]. Furthermore, one may argue the longer a sentence, the more information it comprises and the more relevant it becomes. Third, we include the two proposed context selectors based on definitions and causal sentences in the comparison for the reasons stated in the introduction of this chapter.

Next, we will briefly describe the seven concrete context selectors and their configuration. Before that, we would like to point out that this comparison intentionally ignores supervised machine learning algorithms as context selectors, although they are frequently used in related works. However they are unavailable in the German language (see also Section 5.4).

SumBasic

The SumBasic [111] algorithm is an unsupervised text summarization algorithm based on word co-occurrence. It has been applied for learning-relevant context selection for automatic question generation in other studies but has only been evaluated in a combined task [6]. We apply German word stemming and stop word removal before using the algorithm on the texts as it is common practice. For every text, the algorithm is set to extract 10 sentences.

LexRank

The LexRank [41] algorithm is a graph-based unsupervised text summarization algorithm relying on word co-occurrence matrices. Multiple related works have used it with varying degrees of success [19, 134, 153]. In order to work properly, the algorithm requires a document set from which it can generate TF-IDF matrices. We, therefore, use the 10kGNAD[1] dataset to initialize the matrices. It comprises a good variety of German language spread across various news domains. In our experiment, the algorithm is set to extract 10 or 20 sentences.

Length Heuristic

We include a simple length heuristic because it is intuitive for sentence selection: the longer a sentence, the more information is comprised. We normalize the sentence length in characters relative to the longest sentence in a text and then select sentences longer than 30%, 40%, or 50% of the longest sentence.

Position Heuristic

The relative position heuristic in the text was an influential factor in other studies [19, 70]. Additionally, Figure 13 indicates that sentences in the texts' beginning have been selected frequently. As a result, we apply position-based selection by selecting all sentences in the initial 20%, 25% or 30% of a given text.

Definition and Causal Context Selection

Finally, we apply the already established definition and causal context selectors to the ground truth. Furthermťore, both selectors are combined into a single selector by selecting all sentences classified as a definition or causal sentence. The underlying assumption is that upstream process steps ideally generate questions about all relevant sentence types and not just one. Thus, they will not ask either about definitions or causal sentences but will employ multiple context selectors and generate a question if a sentence belongs to one of the target types.

[1] https://tblock.github.io/10kGNAD/

Classifier	P	R	F1	#Pred.	High Agreement Prediction Ratio
Cause	0.65	0.55	0.60	43	0.54
Definition	0.79	0.67	0.72	43	0.63
Combined	0.69	0.86	**0.77**	64	0.57
SumBasic[10]	0.48	0.27	0.35	29	0.67
LexRank[10]	0.74	0.45	0.56	31	0.62
LexRank[20]	0.60	0.71	0.65	60	0.60
Position[20%]	**0.88**	0.45	0.60	26	0.65
Position[25%]	0.84	0.53	0.65	32	0.65
Position[30%]	0.78	0.57	0.66	37	0.65
Length[30%]	0.65	**0.90**	0.75	71	0.54
Length[40%]	0.76	0.73	0.74	49	0.56
Length[50%]	0.87	0.53	0.66	31	0.66

Table 6: Classification results for the ground truth dataset (P=Precision, R=Recall). The best scores of every column are bold-faced. The column *#Pred.* shows how many sentences were extracted by every approach. The column *High Agreement Prediction Ratio* shows how many of all predictions of the context selector targeted high agreement sentences. This table has been adapted from Steuer et al. [156].

Comparision Results

The results of each of the seven different context selection approaches are depicted in Table 6. We can group the seven approaches into three classes. First, the unsupervised approaches score between F1=0.35 for SumBasic and F1=0.65 for LexRank with 20 sentences selected. They perform rather weakly on the given texts. Even if they are allowed to select 20 sentences, almost half of the given texts, they still score worse than the other approaches. Second, the heuristics perform well for the given texts. The 20% position heuristic has the best precision of all approaches. The 30% length heuristic has the best recall of all approaches. However, it selects the most sentences total, thus, to a degree, trading precision for recall. Additionally, it has a low high agreement prediction ratio, indicating how many of its overall predictions were on the high agreement sentences. Thus, it has a comparatively high likelihood of detecting sentences that were not clearly perceived as learning-relevant in the raw data. Third, the proposed causal and definition context selectors score F1=0.60 for causal and F1=0.72 for definition. The causal selector performs considerably worse than the definition selector not only for F1 but also in precision and recall. The combined approach achieves the best F1 score for the given texts (F1=0.77). It performs slightly better than the best length heuristic and selects slightly more high agreement sentences than the heuristic.

5.4 VALIDITY CONSIDERATIONS

Before discussing the main insights gathered in the context selectors' empirical evaluation, we would like to stress noteworthy validity considerations.

First, the results are derived from five science texts and 37 readers. Consequently, we regard the results as part of a case study, yielding task characteristics and rough performance estimates. These task characteristics and performance estimates are valuable because related work usually evaluated context selection only jointly in an AQG pipeline. Moreover, the case study demonstrates the competetiveness of the causal and definition context selectors from which we previously only knew that they should be helpful in theory. Nevertheless, it is unclear how these results generalize to a larger readership and other reading material. Varying text properties, such as length or domain and varying reader properties, such as prior knowledge or learning goal, will greatly influence transferability. We assume that the results are likely to transfer to scientific reading comprehension with a readership with little prior knowledge because it is the readership we mainly were sampling in the given experiment.

Second, the experiment only included unsupervised summarization algorithms as baselines and did not address supervised context selection. It is possible that supervised context selectors transfer better to the given texts, establishing a stronger baseline than the given unsupervised selectors. However, to the best of our knowledge, no large-scale German summarization corpora exist, and the English context selectors do not directly transfer to the German language. This highlights an evident weakness of supervised context selectors as they usually need extensive training corpora. For this reason, we initially considered conducting the study in English. However, initial plausibility tests with state-of-the-art English supervised summarization methods like Pegasus [183] trained on XSUM [108], were not very promising, often only selecting the first sentence of a given text. Therefore we prioritized having native speakers in the study over evaluating additional supervised context selectors.

Third, we select sentences that are perceived as learning-relevant by readers. However, it is unclear if this judgement is always sound. Readers may not thoroughly understand what information will help them to learn best. We used this approximation due to a lack of access to a larger group of domain experts able to annotate texts. However, we have evidence from the physics domain that learners often can identify the same relevant information as experts [27]. Still, future work ideally investigates context selectors with larger corpora annotated by multiple domain experts.

5.5 DISCUSSION OF THE MAIN INSIGHTS

The proposed context selectors' study reveals insights into their applicability to educational Automatic Question Generation.

5.5.1 *Performance of the Educationally Motivated Context Selectors*

The study finds evidence that the proposed context selectors provide competitive selections compared to other multi-domain unsupervised context selectors. The learner-selected sentences contain disproportionally more definition and causal sentences than the sentences not selected by learners. The effect is more pronounced for definitions than causal sentences. Moreover, both context selectors in combination slightly outperform all other approaches on the corpus of clear annotation cases. Their selections overlap moderately, complementing each other to some degree on the given corpus. The individual context selector performed worse, and the causal selector was outperformed by the best LexRank approach. These findings are consistent with the analysis in Chapter 3, indicating that definition and causal sentences are relevant for learning. Classification approaches selecting these sentence types provide competitive results without clearly outperforming the other approaches.

5.5.2 *Performance of Heuristics and Overall Task Subjectivity*

Nonetheless, the results also show that simple heuristics like a sentence's position or length perform well in the clear annotation cases. This is in line with findings of previous works on shorter texts, indicating that these are strong predictors for relevance on some corpora [19, 70]. However, we suspect that the position and length heuristics will yield worse results in different learning scenarios with longer texts because they do not capture sentences' semantics. Thus, although they worked for the short texts shown in the study, we suspect they become less indicative of relevance if we look at a complete textbook chapter. However, this is only an assumption, and future work should investigate context selection in longer texts. Furthermore, the results already suggest that although some heuristics perform well on the high agreement sentences, they contain the trade-off of also selecting more debated sentences.

Finally, the results suggest that the general task is to some degree subjective because the annotators often disagree with each other. Yet these individual selection nuances are currently not covered by context selectors, relying only on text characteristics and neglecting learner characteristics. Thus, investigating mechanisms to incorporate more learner information into the selection is a promising direction for future work.

5.5.3 *Learning-relevant Contexts via Educationally Motivated Context Selectors*

In summary, the first research question concerns how text characteristics help extract learning-relevant contexts. The described findings suggest that the proposed methods are able to exploit text characteristics for finding learning-relevant contexts. Moreover, the results suggest that educationally motivated selectors are easily combinable to cover different aspects of the texts. Context selection based on definitions and causal sentences is, thus, theoretically well-informed, and can be justified with empirical evidence. Particularly, author-facing approaches profit from the findings because they often have no signals besides text characteristics during the authoring process.

Yet, in the study, the text characteristics are only moderately associated with perceived learning relevance. Therefore, the extent to which the extracted information can be used in a fully automated educational Automatic Question Generation process is limited. Instead, based on the results, human-in-the-loop context selectors are more appropriate. Such approaches could, for instance, support an author as question recommenders. Additionally, further support of educationally motivated context selectors arises from transparency and acceptability considerations. It can be argued that the selection criteria are easier understandable for users than those of unsupervised summarization approaches or implicit learning approaches. In Chapter 8 of the thesis, the proposed definition-based context selection approach will be integrated into an educational Automatic Question Generation system for which we evaluate its overall question quality and affordances in a reading comprehension scenario.

6 TRANSFERRING EDUCATIONAL ANSWER SELECTION FROM NONEDUCATIONAL TEXTS

Context selection and answer selection are jointly responsible for the generation of learning-relevant questions in the educational Automatic Question Generation (AQG) process. We proposed a novel educationally motivated context selection mechanism in the previous chapter and established its competitiveness against other unsupervised methods (see Chapter 5). Next, this thesis addresses two open research challenges regarding answer selection, directly relevant to the first and second research question. First, this chapter investigates how machine learning-based answer selection approaches trained on noneducational corpora transfer to the educational domain and analyzes the strengths and weaknesses of such a transfer. Second, the following chapter aims to automatically construct educational corpora for training machine learning approaches on the answer selection task to alleviate the lack of manually created educational corpora.

We begin by investigating model transfer between noneducational and educational datasets. Transfer of noneducational models to the educational domain is an overall content selection issue discussed previously (see Chapter 2). It is unknown if the implicit selection criteria learned on noneducational data generalize or if they are distinct from the actual educational context or answers. While we chose not to verify this generalization premise for context selection in the absence of education training corpora, we will now investigate it for answer selection by constructing a new educational corpus via corpus transformation. Although the investigation only considers implicit answer selection, results likely also have implications for context selection approaches, as they frequently rely on the same machine learning models.

The analysis aims to detect if the textual patterns in noneducational and education texts are sufficiently similar for answer selection model transfer, which is relevant for the first research question according to the following rationale. Suppose they are dissimilar and the models do not transfer. In that case, implicitly learning educational answer selection from noneducational corpora is relatively limited as answer selection approaches trained on noneducational corpora will select different answers. Thus, novel educational training corpora would be needed. Consequently, with data from the experiments, we gain insights if the textual patterns in noneducational and educational patterns associated with the extracted answers are alike (RQ 1). Moreover, we can justify whether or not we use models trained on noneducational datasets for building educational AQG systems when addressing the second research question.

Next, we briefly derive why we suspect a difference between noneducational and educational answer selection. We then describe our evaluation setup, our results, and discuss our main findings.

6.1 IMPLICITLY LEARNT ANSWER SELECTION TRANSFER

We investigate the model transfer of noneducational models to educational texts. We know that most research in AQG is driven by noneducational datasets, stemming from various sources such as Wikipedia (e.g. [32, 34, 123]). Datasets such as the Stanford Question Answering Dataset (SQuAD) [128] and Natural Questions (NQ) [78] became frequently and successfully used for AQG because they comprise a large variety of paragraphs, questions and answers. Initially, these datasets were collected to train question-answering systems and have not been explicitly intended for educational usage. Therefore, the comprised contexts, answers and questions are not necessarily of educational nature. For instance, take the following SQuAD sentence:

> *"With modern insights [...] particle physics has devised a Standard Model to describe forces between particles smaller than atoms."*

While the source sentence might help train educational AQG systems, the crowdworkers annotating the corpus opted to mainly extract short noun-based answers [128]. For instance, they posed a question with the answer *"Standard Model"* but not for the answer *"describe forces between particles smaller than atoms"* in the given example sentence. This is not surprising because the annotators had no educational intent in mind and strived to generate diverse questions instead of questions improving reading comprehension. Therefore, we believe such datasets lack important answer options. Hence, it is unclear if answer selection models trained on these datasets yield sufficiently many valid answers for educational questions useful in science reading comprehension scenarios. Nevertheless, in the field of AQG it has been argued that machine learning models trained on the data may be used as educational means [36, 171, 185].

In consequence, we seek to understand the educational capabilities of models trained on the noneducational data. In order to assess the model transfer potentials, we create a new medium-sized educational dataset called Textbook Question Answering with Answer Spans (TQA-A) for the answer selection task (see Section 6.3.3). In contrast to the noneducational datasets under study, the TQA-A dataset comprises annotated answers to questions posed by educational experts to foster text comprehension. That is, the answers in TQA-A are a priori selected to support learners in their reading comprehension. The dataset permits the analysis of the transfer issues from two angles. First, it enables us to identify direct differences between the dataset distributions and their contents. Second, we train and evaluate six different BERT-based architectures on the sentence-based answer selection task using the noneducational datasets (SQuAD, NQ) and the educational dataset (TQA-A). Suppose that the noneducational and educational answer selection tasks are similar, then models trained on the noneducational dataset should achieve strong performance on the educational dataset, pointing towards the helpfulness of implicit machine learning-based noneducational models for educational AQG.

6.2 EVALUATION SETUP

We hypothesize that models trained on noneducational datasets learn different answer selection criteria. That is, the answer selection task defined by noneducational and educational data differs. We operationalize the hypothesis in the following way.

First, we compare TQA-A with the noneducational datasets SQuAD and NQ. We consider the answers' length in words, the answers' named entity ratio and the answers' verb ratio. The verb and named entity ratio indicate how many answers contain at least one verb or one named entity. For instance, every second answer in a dataset with a verb ratio of 0.5 contains at least one verb. Verb and named entity detection are done automatically using the *spaCy* tagger [107]. We expect the noneducational datasets to contain fewer verbs and more named entities in their answers. Thereby we built on our observation that the noneducational datasets focus on nouns, whereas educational answers may more frequently focus on actions, effects, or concept characteristics. Furthermore, we map the texts of the three corpora into an embedding space using *spaCy* sentence vectors and plot them in two dimensions using t-SNE data visualization [166]. We expect that dataset clusters emerge in the t-SNE plot indicating that the corpora can be separated in the embedding space and therefore define different tasks.

Second, we compare model performance of six widely used machine learning models. We evaluate the noneducational models' performance on their own datasets' test split and the test split of TQA-A. Furthermore, we train the models directly on TQA-A and evaluate their performance. Thus, we will gather 30 evaluation measurements because we train a total of 18 models on SQuAD, NQ and TQA-A, as well as transfer six models from SQuAD to TQA-A and six models from NQ to TQA-A. We expect the models to perform better on data from their training distribution than on the other data. Nevertheless, the noneducational models have seen lots of training data, and therefore we are curious to what extent they transfer their selection capabilities to TQA-A.

The primary performance metrics will be the phrase-level precision, recall and F1 scores. They only count an answer as matching if all words in the ground truth and the selected answer are equal. Besides these metrics, we employ character-level, word-level and embedding similarity-based scores to account for non-exact matches.

The Levenshtein distance [109] is used on the character level to measure the edit distance between two strings. We use it to define phrases that are *near misses* which means they are almost correct because they differ from the ground truth by less than five edits, chosen by the average word length in English [1]. Hence, given equal phrase-level results, a system with more near misses performs better than a system with fewer near misses. Additionally, we compute the relative Levenshtein distance dL between the ground truth answer a_{gold} and the predicted answer a_{pred}, with $|a_{pred}|$ being the answer length in characters, as:

$$dL = \frac{levenshtein(a_{pred}, a_{gold})}{max(|a_{pred}|, |a_{gold}|)} \tag{1}$$

[1] Based on the WolframAlpha knowledge base (10.07.2022): https://www.wolframalpha.com/input?i=mean+character+word+length+english

The value of dL estimates how many characters of the source answer need to be edited to derive the ground truth answer.

Next, the Jaccard distance is used on the word-level. The distance dJ measures the overlap between the words in the predicted answer and the correct answer and complements the character-level Levenshtein distance. Given the respective set of words w_{pred} and w_{gold} and the Jaccard coefficient J, the Jaccard distance dJ it is computed as:

$$J(w_{pred}, w_{gold}) = \frac{|w_{pred} \cap w_{gold}|}{|w_{pred} \cup w_{gold}|} \tag{2}$$

$$dJ(w_{pred}, w_{gold}) = 1 - J(w_{pred}, w_{gold}) \tag{3}$$

Finally, we use an embedding-based distance metric to compute the semantic similarity between different answers. Whereas the word-level and character-level metrics do not account for a model selecting *"dog"* as an answer, if the correct answer is *"hound"*, the embedding metric will recognize them as similar. Given the function $e(\cdot)$ that maps a word into the embedding space and S_{cos} as the well-known cosine distance, the embedding distance dE is computed as:

$$v_{pred} = \frac{\sum_{w \in w_{pred}} e(w)}{|w_{pred}|}, v_{gold} = \frac{\sum_{w \in w_{gold}} e(w)}{|w_{gold}|} \tag{4}$$

$$dE(w_{pred}, w_{gold}) = 1 - S_{cos}(v_{pred}, v_{gold}) \tag{5}$$

In this metric, the sums v_{pred} and v_{gold} represent the corresponding sentences continous-bag-of-word embedding vectors [46]. The embedding function $e(\cdot)$ is chosen a priori, and produces vectors that all have the same dimension.

6.3 DATASETS

We investigate the model transfer for answer selection approaches on three different datasets. The first dataset is SQuAD, and the second dataset is NQ. The last dataset is TQA-A, a novel educational dataset collected explicitly for the experiments.

6.3.1 *Stanford Question Answering Dataset*

The noneducational Stanford Question Answering Dataset (SQuAD) dataset [128], is a collection of questions and answers from Wikipedia articles, that we already used in Chapter 4. It was gathered initially for question-answering tasks but can also be used for question generation. We transformed the SQuAD training set into an answer selection corpus for this experiment. We removed all unanswerable questions and kept only the sentences containing answers in our sentence-level corpus. The final corpus contained 58,341 sentences with their corresponding answers.

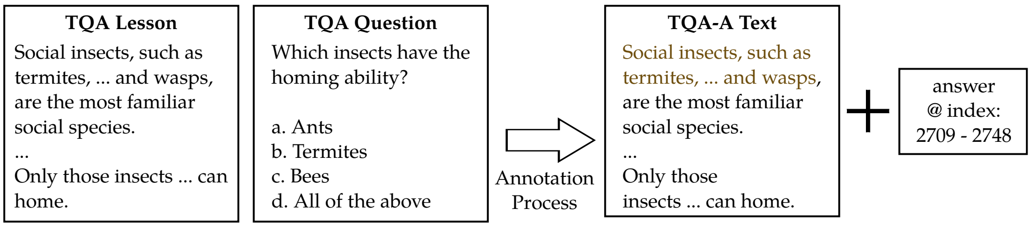

TQA Lesson
Social insects, such as termites, ... and wasps, are the most familiar social species.
...
Only those insects ... can home.
TQA Question
Which insects have the homing ability?
a. Ants
b. Termites
c. Bees
d. All of the above
Annotation Process
TQA-A Text
Social insects, such as termites, ... and wasps, are the most familiar social species.
...
Only those insects ... can home.
answer @ index: 2709 - 2748

Figure 14: In TQA the multiple-choice option differs structurally from the answer span and may not be easily found via string matching. Hence, human annotation is required. This figure has been adapted from Steuer, Filighera, and Tregel [154].

6.3.2 *Natural Questions*

The noneducational Natural Questions (NQ) dataset [78], consists of questions about Wikipedia articles. The questions are designed to explain paragraph information or entities mentioned in the paragraph. Per construction, annotators added therfore frequently short answers referring to named entities. The data is condensed to sentence-level for the experiments, resulting in a dataset of 79,452 sentences.

6.3.3 *Textbook Question Answering with Answer Spans*

We created a novel educational dataset called Textbook Question Answering with Answer Spans (TQA-A). It enriches the educational Textbook Question Answering (TQA) [68] dataset with answer span annotations enabling the answer selection task. To the best of our knowledge, this is the first educational question-answering dataset that contains the sentence, question, and answer span information for 794 middle school level lessons from life science, earth science, and physical science textbooks. Although it is smaller than SQuAD and NQ, it is derived from data explicitly designed for educational use and thus holds promise for researching AQG in education. We constructed the dataset from the source TQA dataset using the following procedure.

First, we extracted 794 of the 1,076 lessons from the TQA dataset. These lessons were chosen because they contain texts with fewer than 1,000 words, reducing the annotators' workload and are likely to be within the maximum sequence length of current machine learning architectures. Second, the 794 lessons contain multiple-choice questions for which we need to manually annotate the correct answer span within the lesson's text because the multiple-choice answer options may be paraphrases of the lesson's content. As the correct answer option is provided in the dataset, the annotation procedure is reasonably straightforward because one can read the correct answer first and then match the corresponding text span (see Figure 14). Accordingly, we decided to work with a single annotator for the task, which is a student assistant with prior experience in other annotation projects, receiving monetary compensation. Third, the resulting TQA-A dataset has annotated answer spans for 794 lessons and 2,906 questions. We transform it into a sentence-level corpus dropping all sentences not containing an answer span, arriving at a corpus of 2,770 sentences for our experiments.

6.4 ANSWER SELECTION MODELS

We compare six machine learning models on the three datasets. All compared models are built on the BERT-based transformer architecture family. We choose to use this architectural family for two reasons. First, BERT-based architectures recently received much attention in the educational technology community, and they have been applied successfully in the educational domain (e.g. [15, 118]). Second, most newer transformer architectures are built mainly on the same architectural premises but operate with more pretraining data and a larger parameter space. Consequently, they require sizeable cloud-based infrastructure for training as their memory footprint increases with their paramaters, making their allocation on single GPUs impossible. For instance, even for a batch size of one, the three billion parameters T5 model requires 40 GB of accelerator (GPU or TPU) memory [40]. Additionally, experiments would frequently explore multiple batch sizes. Thus, our available computational resources inhibit us from using the models, and renting the cloud-based infrastructure would be too expensive. However, we expect some results to transfer to these models because of their close similarity to the BERT model architectures.

6.4.1 *Model Architectures*

The used architectures in the experiments are BERT [30], RoBERTa [89], DEBERTa [54], DistilBERT [135], ALBERT [80], and SpanBERT [66]. The baseline model in the answer selection experiments is a token classification BERT model, which is a *cased*, meaning that it operates with lower and uppercase tokens. The model's architecture is extended with the classification head, and no other model aspects are changed. Furthermore, we include a RoBERTa model, the result of a replication study of the initial BERT research that outperforms the BERT model on the important GLUE benchmark due to enhanced hyperparameters and pretraining. Next, we include the DEBERTa architecture, a recent improvement over RoBERTa. The main changes in the model come from a new attention mechanism as well as from improved masking during the decoding process. This has resulted in the model outperforming RoBERTa on various tasks while using considerably fewer parameters. Next, we include two models with lower computational requirements in our experiments. We use DistilBERT, which has a fraction of the parameters of other BERT architectures, but has still achieved similar results as the original BERT on various tasks. Similarly, we will also be including the ALBERT architecture in our comparison. The model shares parameters across layers, resulting in less memory consumption than the full BERT architectures. Last, we also compare the SpanBERT model. The parameter size is similar to BERT's, but it is designed specifically for predicting text spans. The answer selection task is a span prediction problem, and therefore we suppose it may also perform well on the task in the educational domain.

	ALBERT	BERT	DEBERTa	DistilBERT	RoBERTa	SpanBERT
SQuAD	0.58 (4)	0.59 (4)	0.59 (4)	0.58 (4)	0.59 (4)	0.58 (4)
NQ	0.66 (4)	0.67 (3)	0.68 (4)	0.66 (4)	0.68 (4)	0.68 (3)
TQA-A	0.56 (4)	0.56 (3)	0.56 (5)	0.54 (4)	0.54 (3)	0.57 (3)

Table 7: Token-level F1 scores of the chosen models on the validation split of the different datasets. The number in brackets indicates after how many epochs a model achieved the best score. This table has been adapted from Steuer, Filighera, and Tregel [154].

6.4.2 *Model Training*

The six models are fine-tuned for six epochs on all three datasets on four NVIDIA RTX2080 GPUs. We choose six epochs as the initial BERT authors found this to be an effective training schedule for various fine-tuning tasks [30]. The training relies on the Hugging Face transformer library [178] and the noneducational datasets are split 70%/15%/15% whereas the TQA-A is split into 70% training data, 12% validation data and 18% test data due to its smaller size, yielding more testing data. Furthermore, the batch size is regarded as the only hyperparameter, and we vary it between 4, 8, 16, and 32, exploring the recommended choices by Devlin et al. [30] and additional smaller values fitting on the available GPUs. All models achieved their best performance on the validation dataset before the sixth epoch and showed little change in performance after that. Thus, the observation by the BERT authors is confirmed in the given experiments. Finally, every architecture's best-fine-tuned model is selected via token-level F1 score on the validation set (see Table 7).

6.5 RESULTS

Next, we discuss structural dataset differences. It follows an in-depth analysis of the experiments examining the models' performance differences. We conclude with a a discussion of models' dataset performance changes, and the educational transfer results.

6.5.1 *Structural Dataset Analysis*

We analyze the datasets and their structural differences to investigate whether the different datasets constitute different noneducational and educational answer selection tasks. The dataset statistics are shown in Table 8 and can be distinguished based on the descriptive statistics. All datasets contain mostly short answers with a median length of two words. The answer length frequently changes, with the mean answer length being considerably higher in all datasets. It does not look like the TQA-A answers are considerably longer than the answers in SQuAD or NQ. Yet, TQA-A contains slightly fewer answers per sentence than NQ. Furthermore, TQA-A answers contain considerably more verbs and less named entities in their annotated answers, as we expected

Dataset	#Sentence	#Answer	$\frac{\#\text{Ans.}}{\#\text{Sent.}}$	$M_{\lvert\text{Ans.}\rvert}$	$Mdn_{\lvert\text{Ans.}\rvert}$	$SD_{\lvert\text{Ans.}\rvert}$	N.E.	Verb
TQA-A	2,770	2,846	1.03	4.43	2	4.60	0.17	0.33
NQ	79,452	91,466	1.15	4.54	2	5.98	0.70	0.10
SQuAD	58,341	83,189	1.43	3.67	2	3.95	0.60	0.12

Table 8: Sentence-level statistics of the three datasets. The mean (M), median (Mdn) and standard deviation (SD) of the answer length (|Ans.|) are measured in words. The *N.E.* and *Verb* columns show the respective named entity and verb ratios in the answers(0..1). This table has been adapted from Steuer, Filighera, and Tregel [154].

based on the dataset construction processes. Although not too dissimilar, these dataset statistics already hint toward different answer selection tasks. Especially, the verb focus of the TQA-A dataset provides initial evidence that the educational data focuses more on actions or characteristics described by the sentences than the other datasets.

The t-SNE plot of the three datasets' embedding vectors supports this view (see Figure 15). One has to be careful with t-SNE parameterization because different hyperparameters might lead to vastly different plots [175]. Hence, we construct multiple plots with 5,000 iterations and varied perplexity between 10 and 50. We ensured that all different parameterizations resulted in similar plots, thus indicating that the emerging clusters were not a parameterization artifact. Visually, similar clusters emerged independent of the parameterization. Figure 15 contains roughly four distinct clusters for the three different datasets: one for TQA-A, one for SQuAD and two for NQ. Visually, the SQuAD and primary NQ cluster overlapped most, whereas TQA-A overlapped less with the other datasets. The second NQ cluster is completely separated from all other data. Finding this additional separation of NQ was unexpected, but we did not investigate it further because we believed it would not improve the understanding of educational and noneducational model characteristics. In summary, the t-SNE plot further establishes that the three datasets define three distinct answer selection tasks which may result in trained models utilizing different selection criteria. We will test this hypothesis further by examining the trained models' performances.

6.5.2 *Performance of Different Model Architectures*

Per Dataset Performance

We measured the exact match performance of the six model architectures trained directly on SQuAD, NQ and TQA-A. The results can be seen in Table 9. Overall, the worst and best model architectures differ by five F1 points on SQuAD, five F1 points on NQ and seven F1 points on TQA-A as seen Table 9's dataset columns.

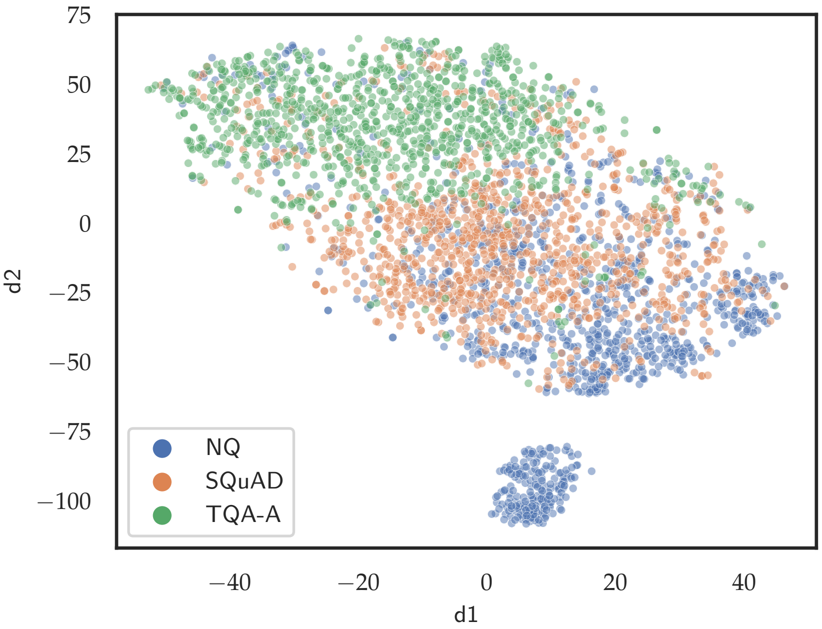


Figure 15: The t-SNE plot of the three datasets using a perplexity of 40. The data points cluster into four groups. Visually, there is more overlap between SQuAD and NQ than between TQA-A and the rest.

Large Architectures

The four larger models achieved similiar results differing at most three F1 points on the SQuAD and NQ datasets (see Table 9; rows 3-6). On the SQuAD dataset, the DEBERTa model performs best, whereas on the NQ dataset SpanBERT performance best, closely followed by DEBERTa. On the TQA-A dataset SpanBERT also performs best, outperforming DEBERTa considerably by a large margin of seven F1 points. Consequently, we consider SpanBERT the best model for the task based on the F1 score as exact match metric, because we are interested in educational task performance.

Moreover, The precision and recall values of the models usually do not differ much. Only BERT and DEBERTa have a considerably higher recall than precision on the NQ dataset. On the noneducational datasets, we see that the more recent large architectures RoBERTa, DEBERTa and SpanBERT perform slightly better than the original BERT (see Table 9; rows 3-6; columns SQuAD and NQ). However, we cannot see that the model improvements affect performance on the TQA-A dataset except for the SpanBERT model (see Table 9; rows 3-6; columns TQA-A).

Resource-efficient Architectures

The resource-efficient models perform weaker on the noneducational datasets (see Table 9; rows 1-2; column TQA-A). Particularly the smallest model's performance (DistilBERT) is considerably worse than the best architecture on any given dataset.

#		SQuAD			NQ			TQA-A		
		P	R	F1	P	R	F1	P	R	F1
1	ALBERT	0.29	0.28	0.29	0.40	0.41	0.41	0.19	0.21	0.20
2	DistilBERT	0.28	0.28	0.28	0.38	0.42	0.40	0.15	0.21	0.18
3	BERT	0.30	0.30	0.30	0.40	0.44	0.42	0.18	0.24	0.20
4	DEBERTa	0.33	0.32	**0.33**	0.43	0.44	0.44	0.15	0.18	0.16
5	RoBERTa	0.33	0.31	0.32	0.43	0.44	0.43	0.17	0.24	0.20
6	SpanBERT	0.31	0.30	0.30	0.44	0.45	**0.45**	0.21	0.27	**0.23**

Table 9: Phrase-level evaluation results for the six model architectures on the three datasets' test splits (P=Precision, R=Recall). We group the models into resource-efficient (row 1 and 2) and larger models. The best achieving model on a dataset is highlighted in bold. This table has been adapted from Steuer, Filighera, and Tregel [154].

Interestingly, only on TQA-A, the resource-efficient models are not the weakest because DEBERTa's performance deteriorates on the dataset, thus becoming performance-wise the worst classifier.

In summary, based on exact match metric, the SpanBERT model architecture performed best on the TQA-A and NQ, making it the best model. Moreover, using resource-efficient architectures lead to substantial performance loss in terms of F1 score on the SQuAD and NQ and slight performance loss on TQA-A.

Non-exact Match Metrics

However, the exact match analysis based on the F1 scores may not provide a complete view of the results as they might hide a good model that frequently predicts almost correct answer spans. Hence, character-level, word-level and semantic measures were analyzed to see if the models differ considerably in their non-exact match behaviour (see Section 6.2). All metrics indicated no significant differences between architectures on the same dataset, with a standard deviation of 0.01 on all three non-exact match metrics. Moreover, the non-exact match metrics vary only slightly between the datasets. Figure 16 illustrates this by plotting the distance metrics averaged over all models by dataset. The plot ignores all data points of exact matches, resulting in better visualization of the variation in non-exact matches. Hence, it does not depict the same model performance differences as the exact match metrics in Table 9. We can see in the plot that if the models fail, they frequently do so by a large margin and only rarely generate near misses. Thus, the non-exact match results validate the results obtained by the exact match F1 score analysis.

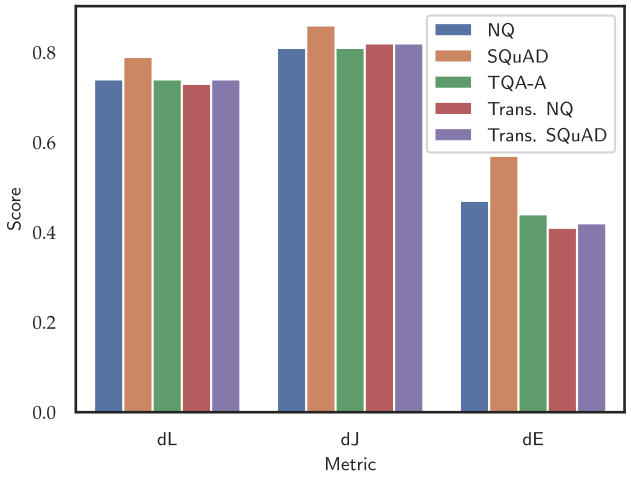


Figure 16: The non-exact match metrics. A single bar represents the average of all models on a given dataset. The standard deviation for every bar is 0.01 or smaller. The dL column reports Levenshtein, the dJ column Jaccard, and the dE column embedding space distance. This figure has been adapted from Steuer, Filighera, and Tregel [154].

6.5.3 *Model Performance and Dataset Size*

The models' performance on the three datasets varies widely. The respective best model achieves F1=0.33 on SQuAD (DEBERTa), F1=0.45 on NQ (SpanBERT) and only F1=0.23 on TQA-A (SpanBERT; see Table 9). One apparent reason for these large differences is that the datasets differ in sizes, with NQ being the largest and TQA-A being the smallest. Thus, the models see different amounts of training data to learn answer selection patterns. Ideally, the more data points a model encounters, the better it learns to distinguish between different selection patterns. In consequence, the performance difference may be explained by training dataset size. We conduct a downsampling experiment to gather more evidence for this explanation. In the experiment, we reduce the training size for the SQuAD and NQ models by randomly sampling data points from the original training set until reaching the size of TQA-A (2,770 sentences). Next, we train a SpanBERT model on the downsampled data. The downsampled model achieves Precision=0.23, Recall=0.27 and F1=0.25 on the SQuAD data and Precision=0.26, Recall =0.25 and F1=0.25 on the NQ data. That is, models trained on the downsampled noneducational corpora perform almost equivalent on their respective evaluation corpus to the models trained and evaluated on TQA-A. Accordingly, the dataset size is vital for model performance and the experiment yields evidence that TQA-A is too tiny to train models directly. Yet even for the larger corpora, there is a considerable performance gap with models performing at F1=0.33 on SQuAD but F1=0.45 on NQ. Although SQuAD is 26% smaller than NQ, it is still a large-scale dataset with plenty of examples to learn answer selection differences. Hence, whether the performance differences stem purely from the dataset sizes is uncertain. We consider it more likely that they partly result from the structural differences in the datasets (see Section 6.5.1).

#		TQA-A			Trans. SQuAD			Trans. NQ		
		P	R	F1	P	R	F1	P	R	F1
1	ALBERT	0.19	0.21	0.20	0.17	0.19	0.18	0.18	0.20	0.19
2	DistilBERT	0.15	0.21	0.18	0.18	0.21	0.19	0.17	0.21	0.19
3	BERT	0.18	0.24	0.20	0.18	0.22	0.20	0.16	0.21	0.18
4	DEBERTa	0.15	0.18	0.16	0.20	0.22	0.21	0.18	0.20	0.19
5	RoBERTa	0.17	0.24	0.20	0.20	0.22	0.21	0.18	0.20	0.19
6	SpanBERT	0.21	0.27	**0.23**	0.21	0.24	**0.22**	0.20	0.25	**0.22**

Table 10: Phrase-level evaluation results for the six models and their transfer to the TQA-A test split (P=Precision, R=Recall). We group the models into resource-efficient (row 1 and 2) and larger models. The *Trans.* columns show values for models trained on the respective dataset being transferred to TQA-A. The best achieving model on a dataset is highlighted in bold (adapted from Steuer, Filighera, and Tregel [154]).

6.5.4 *Educational Transfer Analysis*

Next, we analyze model transfer from the noneducational corpora to TQA-A.

Analysis via Model Comparision

Models trained on TQA-A frequently receive higher F1 scores than noneducational models applied to TQA-A but trained on an noneducational dataset, although their training dataset was a magnitude smaller (see Table 10). The only exception are the DEBERTa and DistilBERT architectures, where the transferred models perform better than the model trained directly on TQA-A (see Table 10; row 2 and 4). Furthermore, the performance on the source dataset is not associated with the performance on the TQA-A dataset. Even though models trained on NQ receive a higher maximum F1 score on the NQ dataset than the SQuAD models on the SQuAD dataset (see Table 9), the model families have approximately the same transfer performance on TQA-A in terms of F1 score. The SQuAD models lose on average 33% (SD=3.71%) and the NQ on average 55% (SD=2.81%) of their performance on TQA-A (see Table 10 *Trans.* columns vs. Table 9 dataset columns). Moreover, model transfer affects the resource-efficient models as well as the larger models (see Table 10; rows 1 and 2). In summary, the model transfer is challenging for all models independent of the architecture or source dataset.

Selected Answer Characterization

A potential reason for the challenging transfer to the educational domain and the models' drop in performance is that the datasets implicitly define different answer selection tasks. Thus, the answer selection criteria learned on SQuAD and NQ do not match those of TQA-A. This idea is already supported to some extent by the

	Dataset	Models		
	TQA-A	TQA-A	SQuAD	NQ
Relative named entity count	0.17	0.12	0.18	0.15
Relative verb count	0.33	0.26	0.21	0.21
Mean answer length	4.32	2.76	3.73	3.42
SD answer length	4.48	2.67	3.26	4.00

Table 11: Distribution of the ratio of named entities, verbs and the answer length in words as predicted by the best SpanBERT models trained on the different datasets on the TQA-A dataset. The column *TQA-A dataset* shows the actual distribution in the dataset.

structural dataset analysis. If this reasoning is valid, we should see different answer patterns selected by the noneducational and educational models. Hence, we look at the characteristics of the selected answers by different models. Respectively, the best SpanBERT models trained on SQuAD and TQA-A share 39% of predicted answers and 32% of predicted answers between NQ and TQA-A. Consequently, although they achieve similar performance, they do so by selecting different answers. Table 11 illustrates this with descriptive statistics of the model outputs. The trained TQA-A model is more verb-oriented and less named entity-focused in its outputs, whereas the trained SQuAD and NQ models prefer named entities over verbs. Furthermore, the TQA-A model prefers shorter answers more frequently than the other models. This is interesting because the average answer length in TQA-A is slightly shorter but still longer than the model's average selection length. Therefore, we suspect that the shorter predictions stem from insufficient training data for the TQA-A model, priming the model towards the median answer length.

6.6 VALIDITY CONSIDERATIONS

The analysis aimed to identify the extent to which models trained on noneducational data transfer to educational data and to measure models' overall performance on the novel educational TQA-A dataset. Before reviewing the results more closely in the discussion, we would like to present some validity considerations.

First, TQA-A is a dataset constructed from factual multiple-choice questions. It is not representative for every answer selection tasks but defines a task with answers one may find helpful in education. The results indicate that directly transferring noneducational answer selection models to this specific educational answer selection task is challenging. This finding is important because we think that the TQA-A answer selection task is challenging but still relatively close to many noneducational tasks. After all, the answers in TQA-A are mainly facts suitable for generating multiple-choice questions. We would assume more complex educational answer selection tasks would result in a even worse transfer. Yet, this assumption should be validated for other answer selection tasks, not targeting factual answers. Moreover, because TQA-A is small, we would like to point out a methodological consideration one should keep in mind in future studies. In the current experiments, we have an explicit train/validation/test split for TQA-A because we aim to keep the evaluation steps for the large and small corpora equivalent. Yet, if working solely with TQA-A, we assume refraining from a distinct validation split and instead relying on k-fold cross validation will result in more stable performance estimates due to the dataset size.

Second, one has to be cautious in interpreting the model analysis. Large neural networks such as the evaluated BERT-based architectures are inherently intransparent, and we cannot easily deduce their selection criteria. Thus, the reported analysis tries to gather insights into their selection criteria based on model outputs. In the output statistics, the models differ, and we suspect this is due to different selection criteria applied by the models. Moreover, other works have shown that these models learn to rely on spurious correlation in their training data [102, 114]. Hence although we assume that the selection criteria differ in their prioritization of verbs and named entities due to the different distribution in the datasets and the model outputs, we do not know for certain that this is the actual difference in the selection criteria. The models may apply a completely different selection criterion (e.g. based on the text's domain or length) that leads to more verbs being selected from the TQA-A trained models as a side effect. Therefore, based on the data, we conclude that the models' selection criteria are likely different. However, if the difference is associated with the verb and named entity count in the corpora is a reasonable hypothesis but not a proven fact.

Finally, the analysis focuses on BERT-based architectures. BERT and other transformer networks have become widespread and changed Natural Language Understanding (NLU) in recent years. Many state-of-the-art models in NLU and Natural Language Generation (NLG) currently have an underlying transformer network (e.g. [11, 30, 112, 123, 167]). However, there are other token classification, and answer extraction approaches. It cannot be excluded that, for example, neuro-symbolic approaches, which are currently becoming more widespread and promise better out-of-sample predictions [52], achieve better performance or are more likely to detect commonalities between the noneducational and educational tasks.

6.7 DISCUSSION OF THE MAIN INSIGHTS

The following main insights can be distilled from the results.

6.7.1 *Overall Model Performance and Model Architectures*

First, the model performance depended clearly on the source dataset in the conducted experiments. Even on the noneducational datasets, a wide performance gap emerged. We identified dataset size as one potential reason for this gap in the reported downsampling experiment. The downsampling experiment also demonstrated that TQA-A alone is probably not large enough for sole model training, and therefore, the dataset should be primarily used as an evaluation set for educational answer selection.

However, the reported experiments also established that dataset size is not the only factor influencing model performance. The datasets ′ descriptive analysis, the t-SNE plots and the analysis of the model predictions all implied that the datasets define different answer selection tasks. Accordingly, the three tasks are likely to have different difficulties, leading the models to perform differently depending on the task. Besides, the architectures had a much smaller impact on the overall result than the characteristics of the datasets. In most cases, all large BERT models solved the task with similar performance. Only the resource-efficient models performed considerably worse in comparison. Consequently, we conclude that the datasets' characteristics are vital for training educational answer selection approaches. The analysis clearly shows that one cannot simply assume that different datasets define similar tasks or that models transfer between these tasks.

6.7.2 *Transfer of Noneducational Models into the Educational Domain*

There were substantial variations in the performance of the models when transferred to educational data compared to evaluating the same models on their original test data. In most cases, the models achieved approximately the similar F1 scores as those trained directly on the educational data. On the one hand, we could therefore argue that the model transfer to education has succeeded and that the noneducational models approximate the educational dataset just as well as a model trained directly on it. However, this disregards that the noneducational models have seen an order of magnitude more training data. In contrast, using this additional training data, the models were supposed to perform noticeably better on the task than the sole TQA-A model. Moreover, according to the experimental data, the direct TQA-A model likely selects data based on other criteria than the noneducational models. Based on the output statistics it focuses more on verbs and less on named entities.

6.7.3 *Implicit Answer Selection Using Text Characteristics*

Combining this transfer analysis and the analysis of general model performance, we find novel insights related to the first research question. Namely, the extent to which noneducational answer selection methods and likely content selection methods, in general, can be transferred to education is rather limited. They cannot leverage their advantage in training data in our experiments and perform only equally as educational models trained on a magnitude fewer data. We also observed proxy measures of how the answer selection might differ, seeing that noneducational data results in a model focusing more on noun phrases. In contrast, educational data incorporates more verbs into the selection. It follows that while implicit answer selection works to an extent on the source dataset, it does not transfer its quality to out-of-distribution data. Thus, in terms of the first research question, the extent to which implicit answer selectors can be used to extract relevant information is limited by the availability of appropriate training data. For educational science texts, we lack appropriate training datasets and relying on noneducational training data likely results in subpar selections.

Therefore, we recommend collecting larger educational corpora if one aims to rely on implicitly learned answer selection or to rely on educationally motivated selectors similiar to the one proposed in Chapter 5.

Moreover, the analysis exposes a second shortcoming of implicitly learned models. Although we conducted a complex and in-depth analysis of the underlying models, we cannot be certain what the actual decision criteria of the trained models were. Concerning the second research question, this introduces an obstacle for any educational Automatic Question Generation system solely relying upon implicitly learned content selectors. In the educational context, it would be vital for users, such as teachers or authors, to understand why they should ask a question about a piece of information. After all, they are the experts deciding whether a question benefits the learner. However, with the implicitly learned models, it is hardly possible to provide such an explanation.

7 ANSWER SELECTION FROM AUTOMATICALLY CONSTRUCTED TEXTBOOK CORPORA

We established in Chapter 6 that transferring noneducational models to the educational domain has limited value. Hence, we will now investigate to which extent automatically constructed educational corpora may be used to train answer selectors. We thereby focus on learning the selection of learning-relevant concepts from texts.

Conceptually, successful text comprehension relies on understanding and remembering learning-relevant key concepts, and asking learners questions about these has been shown to foster their text comprehension (see Chapter 3). Therefore, if we know which sentence part entails a learning-relevant concept, we may inquire about this part or, depending on the connecting verb, about the remaining sentence parts. For instance, given the relevant concept *"Jitter"* and the sentence:

> *"Jitter causes network congestion and poor hardware performance."*

we may select *"Jitter"* as the answer, or, as the verb is *"cause"*, we select *"network congestion and poor hardware performance"* as the answer. Consequently, automatic learning-relevant concept selection is a building block for successful answer selection.

However, detecting learning-relevant concepts is complex for several reasons. First, The number of candidate concepts is vast. For example, any compound may be regarded as relevant in chemical texts. Second, context plays a crucial role for concept relevancy. In the chemical domain, *"water"* may be mentioned as an everyday concept in one text but be deemed relevant to another due to the surrounding context. Third, other challenges to solve include homonym expressions (like the word *"ring"* in mathematics) or the fact that concepts are often inflectable multi-word expressions.

Due to these associated challenges, the task appears similar to the Named Entity Recognition (NER) task, which aims to detect real-world objects in a text. Recently, BERT-based machine learning models have shown state-of-the-art performance on NER tasks [30]. Thus, we aim to transfer these machine learning methods to relevant concept selection, theorizing that these methods can learn to select the relevant concepts from the surrounding text characteristics. However, this requires large-scale annotated training corpora, comprising learning-relevant concepts used in texts.

In education, textbooks are ubiquitous and span almost any domain. They frequently encompass manually-curated back-of-the-book indices listing the relevant concepts. Moreover, they also contain text paragraphs using these concepts. Together, these two pieces of information allow the construction of a corpus of *(text, concept)* tuples that can be used to train machine learning models. Hence, they are an excellent building block to construct the required training corpora. However, textbooks usually come in the form of Portable Document Format (PDF) files, which are meant for printing and not parsing. Thus, the files convolute structure and layout information, and extracting the necessary data from the books requires sophisticated parsing techniques.

Consequently, in this chapter, we propose a novel PDF index extraction method. Given this extraction method, we automatically construct large-scale corpora for learning-relevant concept extraction, which we then use to train and evaluate different machine learning models for answer selection. We find that the quality of concept extraction depends on how much memorization the networks can apply. However, even if the possibility of memorization is eliminated, the approaches can extract various learning-relevant concepts by relying on the text characteristics. The resulting answer selector may be used to select relevant concept-based answers to guide the question realization process in texts where no back-of-the-book index is available.

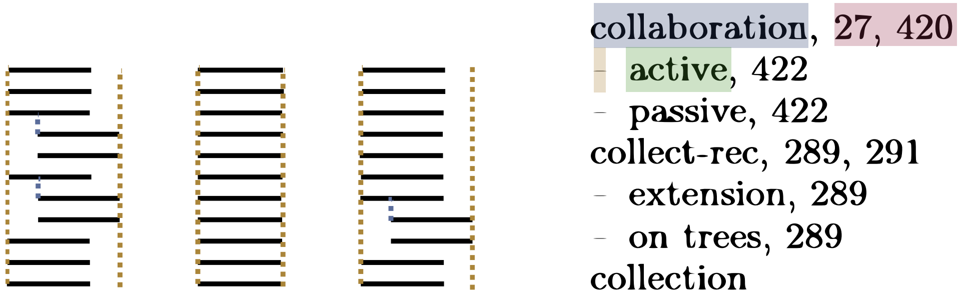
7.1 AUTOMATIC INDEX EXTRACTION

The corpora required to learn answer selection comprise *(text, concept)* tuples. They can be harvested from textbooks, usually available in PDF format. Although the idea sounds straightforward, PDF complicates its implementation because of the format's complicated parsing resulting in a lack of index extraction tools. To the best of our knowledge, only Alpizar-Chacon and Sosnovsky [1] tackled PDF index extraction from textbooks so far. The authors propose an information extraction tool, which transforms a PDF file into a knowledge source in a multistage algorithm. One step of the algorithm extracts the index concepts and stores them in a database. The authors report that their extractor achieves a recall of over 95% on their evaluation set of 40 books from three domains. However, the system is not lightweight and requires considerable setup effort and computational resources to run, as the index extraction is only a single step in the book transformation pipeline. Therefore, in the absence of a lightweight alternative, we design and implement a novel open-source program for PDF index extraction. We aim for a program with the following key features:

- highly accurate index detection for many of the typical index structures (e.g. two-column vs. three columns)
- an output format easily machine-readable for different kinds of upstream tasks (e.g. Automatic Question Generation (AQG) or information extraction)
- requires no complicated setup or external databases. Instead, it has a single purpose and is easy to use and fast.

The finished program *pdf-index-extractor* is based on the PDFAct library [5] and processes PDFs at the glyph and text level to extract indexes from them. The extraction algorithm thereby follows a three-step approach.

1. filter the book pages to the pages comprising the back-of-the-book index
2. compute relevant bounding boxes (e.g. column format)
3. extract concepts and subconcepts from the detected bounding boxes



(a) Visualizing column bounding boxes: the yellow lines divide the index columns. In a column, text lines always overlap on the X-axis (blue).

(b) A back-of-the-book-index: We extract the concept (blue), the page numbers (red), remove filling information (yellow) and map subconcepts to their parent concepts (green).

Figure 17: Index extraction requires multiple steps, such as graphically determining information that belongs together (a) and extracting the different parts of a given concept (b).

The initial step is based on location, character distribution and heading heuristics. The index detection only applies to the last 20% of a given PDF thereby ignoring many false-positives such as table of contents or tables, which have a similar structure but typically occur earlier in books. Moreover, index lines frequently start with a letter and end with a number. Hence, we use the distribution of character positions on the pages to estimate if a page belongs to an index. Finally, the index is often preceded by characteristic headings like *Index* or *Stichwortverzeichnis* that we also use for detection.

In the second step, we calculate the page layout and structure of the detected index pages. This includes determining the footer and header areas and removing them. Next, we compute the bounding boxes for the column layout through the PDFAct detected text lines. We start with the bounding boxes of all text lines. A recursive algorithm merges the two bounding boxes whenever they overlap on the X-axis, resulting in a larger bounding box spanning the area of the initial two bounding boxes (see Figure 17). The geometric invariant is that text lines in different index columns can never overlap horizontally. The algorithm terminates when there are no bounding boxes left for merging. Hence, the recursion results in the bounding rectangles of the index columns. Last, the columns are broken down into concept blocks and their subconcept blocks utilizing the indentation information of each column.

The last step extracts the different information pieces from the detected concept and subconcept blocks and generates the hierarchical concept tree for XML serialization. For that, it uses various regular expressions to detect the actual concept, the page numbers, or filler information in the concept blocks (see Figure 17). Moreover, potential offsets between the page numbering and the PDF file page numbering are calculated and attached to the corresponding concept, and page number ranges such as *"231-235"* are unrolled.

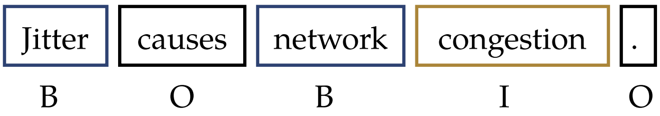


Figure 18: An illustrative example of the IOB sequence tagging scheme. A concept's beginning is marked with a *B*, and continuation is marked with an *I*. All tokens not contained in a concept are marked by *O*.

We evaluate the extraction quality of the proposed *pdf-index-extractor* by comparing it with ground truth. To form the ground truth, we crawl 150 textbooks from the *SpringerLink* publishing platform, which can be downloaded in PDF and EPUB format. The EPUB format is less widespread but more accessible and allows us to automatically extract a ground truth of index concepts for these books. Consequently, we can evaluate our system by comparing the PDF extracted concepts to the EPUB extracted concepts for evaluation. We have almost no restriction on the domains of the crawled textbooks, and the books span numerous topics from biology to economics. The only exception is that we manually filter out law books due to their different indexing nature, relying on law paragraphs. Besides, we exclude textbooks without an index from the evaluation. Given this ground truth, the extraction quality is measured using phrase-level precision, recall and F1 score. A true-positive exactly matches the ground truth concept's spelling, and we only trim leading and trailing whitespace. False-positives comprise either wrongly spelled concepts or phrases not part of a concept, such as page footers or headers. Similarly, false-negatives encompass concepts that only exist in the ground truth. We average the scores by the book, ensuring indices with few concepts and many concepts are weighted the same. In this evaluation setup, the *pdf-index-extractor* achieves Precision=0.91, Recall=0.93 and F1=0.92.

7.2 AUTOMATIC CORPUS CONSTRUCTION

Extracting the back-of-the-book indices is a first step in training machine learning models to become learning-relevant concept taggers. However, the training data requires concepts and the corresponding text paragraphs, using these concepts in a learning-relevant setting because concepts may be considered relevant in one text but not in another. Consequently, we utilize the index page offsets and PDFAct's text block extraction capabilities to extract accompanying text paragraphs for the various index concepts. We base the constructed corpora on these *(text, concept)* tuples. The actual machine learning task is modeled as sequence tagging. The concepts' in-text mentions are Inside–Outside–Beginning (IOB) encoded (see Figure 18), ensuring we can also handle multiword concepts (see Table 12). If the text block contains multiple index concepts, all are marked inside the block.

German and English corpora are constructed to evaluate the usefulness of the automatic corpus construction procedure described above. We rely on two languages because we expect the respective textbooks to vary in language and style. For instance, German authors may use more nouns and fewer verbs, and the language's grammar is more complex than in English. Moreover, most Natural Language Understanding

Corpus	Books	Text Blocks	M_{chapter}	M_{blocks}	M_{concepts}
German	766	159,613	12	211	544
English	280	39,673	13	155	296

Table 12: Corpora statistics for the German and English evaluation corpus. The M_{chapter}, M_{blocks} and M_{concepts} columns are the respective means, averaged over the books and rounded to the next integer. This table has been adapted from Steuer et al. [151].

(NLU) research is focused on English, with fewer resources spent on German models. Thus it is interesting to see how models trained in the two languages perform on this novel task. The corpora are constructed by crawling the *SpringerLink* publishing platform and downloading books accessible by TU Darmstadt's service agreement. The final dataset statistics can be seen in Table 12.

The corpora are divided into a training set and three test sets: a *Random* split, a *Unseen Books* split and a *Unseen Concepts* split. Neither of these splits contains data points already seen during training. The *Random* split retains 15% of the text blocks containing a concept (11,975 DE / 2,317 EN), sampled from the dataset after partitioning the other two test splits. It is the easiest evaluation set because concepts and different text blocks from the same book may have already been seen during training. The *Unseen Books* spilt comprises data points of 30 books (6,982 DE / 3,808 EN) not seen during training. It is more challenging because models need to generalize more on this data. Yet, the split still lends itself to a degree of memorization because some concepts may occur in multiple books. One of these books could be in the training set and one in the evaluation set. Therefore, the final *Unseen Concepts* split ensures that remembering concepts is impossible by sampling text blocks for 1,000 concepts not present in the training data (3,643 DE / 875 EN). In consequence, it becomes the most challenging evaluation split for the models. Besides these evaluation splits, a validation set is sampled before training containing 5% of the training data.

7.3 ANSWER SELECTION MODEL COMPARISION

Five approaches are compared on the English and German dataset sequence labeling task: BERT [30], Frozen BERT [44], a Lookup table, YAKE [14], and the *spaCy* NER tagger [107]. Each approach aims to correctly predict if a given token is part of a relevant concept. The baseline is the lookup table approach. The approach stores all concepts seen during the training and then marks the corresponding words on the evaluation set. We expect the approach to perform well on the *Random* split, worse on the *Unseen Books* split, and not better than majority vote performance on the *Unseen Concepts* split.

Next, we investigate the two machine learning models' performance. The first model is a BERT model fine-tuned on the task, which we regard as a proxy for the prototypical large pre-trained transformer model. BERT has shown strong NER performance; thus, we expect it to learn regularities in the given task. Moreover, we include an Frozen BERT approach. Research suggests that freezing the first half-layers of BERT

	BERT (BS=4)	Frozen-BERT (BS=4)
DE	0.79	0.77
EN	0.76	0.74

Table 13: Validation set token-level F1 scores used for model selection on each corpus.

before fine-tuning improves NER task performance [44]. As we presume a certain task similarity between NER and learning-relevant concept extraction, we seek to verify if the frozen layer model performs better in our experiments as well. We initialize both models with their respective language snapshot and fine-tune using the batch size as a hyperparameter (4, 8, 16, 32, 64) for five epochs as recommended by Devlin et al. [30]. The best model performances on the validation set can be seen in Table 13.

Last, we apply two unsupervised approaches. The YAKE keyphrase extraction algorithm [14] is investigated. It is a general-purpose algorithm for finding central concepts in arbitrary texts and has been used by related work for educational answer selection [152]. Hence, we wonder how it performs given the ground truth from the back-of-the-book indices. Finally, we are interested to what extent *spaCy* NER tagger can address the task. Related work has proposed that named entities can be used as a baseline for educational answer selection [35, 177]. However, we expect it to underperform on the given task because learning-relevant concepts from arbitrary domains do usually not belong to the restricted set of real-world objects detected by named entity taggers.

7.4 EVALUATION RESULTS

We report the primary evaluation results in Table 14. All entries in the table are on the token level. Thus, we have three classes based on the IOB tagging scheme: *'I'*, *'O'* and *'B'* (see Figure 18). We calculate the corresponding metrics for each class and average them. Moreover, we have a heavy class imbalance, as most tokens are labeled *'O'*. Consequently, a metric of 0.33 is not better than the majority vote (voting *'O'* always). We use the F1 score as the summary metric for model comparison.

The performance on the German dataset shows that memorization is a valid technique for the *Random* split as the lookup table approach outperforms all other methods. It achieves a precision of 1.00 indicating that concepts are not as contextualized as assumed in the sampled data. That is, a concept that appeared in the index was rarely also used as an everyday word. Besides, the two machine learning models have also learned to recognize learning-relevant tokens on the *Random* split. The BERT model performs eight F1 points weaker than the lookup table on the *Random* split, still achieving an F1=0.8 and the frozen BERT approach results in a performance of F1=0.78. Although worse than the lookup table approach, both models outperform the unsupervised YAKE and NER systems, scoring almost a doubled F1 score. However, both unsupervised approaches vastly underperform. The NER tagging approach scores only F1=0.40, and the YAKE algorithm scores only F1=0.33, similar to majority vote performance.

		Random Split			Unseen Books			Unseen Concepts		
		P	R	F1	P	R	F1	P	R	F1
DE	BERT	0.80	0.80	0.80	0.73	0.64	0.68	0.51	0.52	0.51
	Frozen-BERT	0.78	0.77	0.78	0.72	0.62	0.67	0.54	0.53	**0.53**
	NER	0.39	0.42	0.40	0.39	0.40	0.39	0.36	0.37	0.36
	YAKE	0.39	0.34	0.33	0.37	0.34	0.33	0.36	0.34	0.33
	Lookup	1.00	0.80	**0.88**	0.99	0.6	**0.71**	0.33	0.33	0.33
EN	BERT	0.76	0.76	0.76	0.66	0.60	**0.62**	0.58	0.53	0.55
	Frozen-BERT	0.75	0.66	0.70	0.67	0.55	0.59	0.64	0.54	**0.58**
	NER	0.37	0.43	0.38	0.37	0.41	0.37	0.36	0.40	0.37
	YAKE	0.38	0.34	0.34	0.37	0.34	0.34	0.37	0.34	0.34
	Lookup	0.99	0.68	**0.77**	0.98	0.51	0.59	0.33	0.33	0.33

Table 14: Token Level F1-score evaluation results (P=Precision, R=Recall). Best results on each split are bold-faced. This table has been adapted from Steuer et al. [151].

On the *Unseen Books* split, the lookup table approach still performs best. However, the performance gap between it and machine learning approaches shrinks. While the precision is still high, the system's recall becomes considerably worse as the unseen books hinder concept memorization. The BERT model has a slight edge over the frozen BERT model, yet both models perform relatively similarly. The YAKE and NER approaches are not dependent on the training data and perform on the same level as on the other split. They still underperform the machine learning algorithms considerably.

The most challenging split is the *Unseen Concepts* split. The lookup table exhibits majority vote performance because memorization is impossible on the split. Nonetheless, the machine learning models partly generalize their predictive performances to this split. Although they perform poorer compared to the previous splits, their classification quality is superior to the unsupervised baselines. This indicates that the learned textual patterns provide a signal for concepts not seen in training. That is, these learned textual patterns generalize to unseen concepts. Besides, in a direct model comparison, the frozen BERT model is slightly ahead of the normal BERT model on this split.

The model differences on the German corpus emerge also on the English corpus. Yet overall, most approaches achieve slightly lower F1 scores on the *Random* and *Unseen Books* splits. Moreover, the BERT model outperforms the lookup table already on *Unseen Books*, probably due to less concept reuse between books in the English dataset. Although the models perform worse on the first two splits, the BERT and frozen BERT models perform better than their German counterparts on the *Unseen Concepts* split. It is unclear why, particularly because the German fine-tuning corpus is larger than the English. One may speculate that this stems from the larger pre-training corpus for the English BERT, allowing the models to generalize better to unseen data.

	Random Split			Unseen Books			Unseen Concepts		
	P	R	F1	P	R	F1	P	R	F1
BERT	0.70	0.72	**0.71**	0.62	0.50	**0.56**	0.31	0.22	**0.26**
Frozen-BERT	0.67	0.68	0.67	0.60	0.47	0.53	0.34	0.21	**0.26**

Table 15: Phrase-level F1-score evaluation results for the two German models (P=Precision, R=Recall). Best results on each split are bold-faced.

Finally, we also report phrase-level, exact match, precision, recall and F1 scores for the two German machine learning approaches (see Table 15). These results aim to give an impression of the extent to which the models recover concepts in an application scenario and are complementary to the token-level model comparision. Suppose the application scenario contains texts similar to the training data, for instance, because one designs an educational AQG for a couple of fixed domains. In this case, memorizing concepts during training will help the models as in the *Random* split condition. In the condition, the best model has a precision of 0.7, indicating that approximately 70% of the recovered concepts are learning-relevant. Moreover, the approach has a recall of 0.72. Hence, it will also recover approximately 72% of all learning-relevant concepts. However, as already indicated in the token-level results, the performance of both models worsens if they cannot rely on memorization. Suppose you built an educational AQG system relying on learning-relevant concept extraction that is entirely domain-independent. In consequence, your performance will be closer to the *Unseen Concepts* case. In this case, the best model can only recover roughly 20% of the relevant concepts. Moreover, it will misclassify irrelevant concepts as relevant rather frequently (P=0.31).

7.5 VALIDITY CONSIDERATIONS

We want to point out the following two important validity constraints.

First, the index extraction evaluation is conducted by automatically constructing the ground truth from online EPUBs. This introduces biases in the evaluation corpus. For instance, the crawled textbooks are all from one publishing house, and we could only access books that were also in the collection of TU Darmstadt's library. Hence, we may miss some back-of-the-book index formats from other publishers or other domains. However, we believe the biases have only a modest effect on the proposed index-extractors usefulness in practice. For one, the publisher's textbooks already come with various back-of-the-book index formats for which the proposed extractors were adapted. Thus, we believe that we already cover the most frequent index formats used in textbooks. We tested this with smaller trial extractions on books from other publishers (e.g., OpenStax [146]) where we did not have a ground truth. In these trials, the extracted XMLs we reviewed were largely correct and contained only minor errors, such as incorrectly adding a page number to a concept.

Second, we would like to emphasize that the model training is conducted on noisy automatic corpora where the ground truth is not 100% accurate. The automatic construction process introduces errors during PDF text block extraction or when matching the concepts to the text blocks. For instance, when extracting the text blocks, we did not try to detect and annotate hyphenated concepts, like "jitt-er" because we assumed that when evaluating the general usefulness of the annotated corpora for machine learning, such preprocessing was not required. However, this hyphenation sometimes occurs if a word is written at the text line's end. Hence, the reported evaluation results will likely differ from possible results on manually annotated corpora. Yet, manual inspection of the corpus showed that such errors only happen infrequently, and the corpus comprises correctly annotated data for most data points. As long as the noise is limited, machine learning models should be able to learn to ignore it and learn meaningful concept extraction from the corpus, which is backed by our experiments. Still, future work may want to look at how to remove this mislabeled data from the corpus.

7.6 DISCUSSION OF THE MAIN INSIGHTS

We infer the following main insights from the evaluation results.

7.6.1 *Automatic Corpora Construction*

It has been uncertain whether automatically constructed corpora from textbooks are suitable for training educational content selectors. We have shown that the proposed automatic corpus construction approach yields large-scale educational concept extraction corpora for English and German. The *pdf-index-extractor* had high accuracy and permitted the annotation of learning-relevant concept ground truth data. The data was sufficient to train machine learning algorithms on the task of learning-relevant concept extraction. The proposed splits allowed us to compare the actual learning on the corpora with different scenarios in mind. The splits simulated a fixed domain strong memorization scenario (*Random* split) and an open domain no memorization scenario (*Unseen Concepts* split). Therefore, we could better understand the different approaches' memorization and generalization capabilities. For this reason, we conclude that the proposed automatic corpus construction is a productive approach for researching the answer selection tasks.

Concerning the first research question it increases the understanding of whether textbooks' textual characteristics can be leveraged to build educational content selectors. This research approach is constructive if we like to generate questions about about relevant domain concepts. However, we have seen in Chapter 6 that educational answer selection corpora also comprise more complex answers. There the approach is still limited, as it does not enable a system to learn to extract these answer types. However, in Chapter 8, we present a combined context and answer selection approach that aims to find more complex answers in sentences that can be supported by the methods presented in this chapter.

7.6.2 *Performance of Answer Selectors on Automatically Constructed Corpora*

We gathered the following insights from the results obtained on the automatically constructed corpora, concerning to what extent textual textbook features are suited for answer selection (RQ 1).

The data indicate that Named Entity Recognition and general-purpose keyphrase extraction do not extract relevant concepts on the constructed corpora. They perform slightly over a majority vote in our model comparison results. It shows that the educational answer selection tasks and the Named Entity Recognition task are discordant. Similiar to the findings in Chapter 6 this provides more evidence, that content selection in the educational Automatic Question Generation setting must be learned on educational data and that transferring models from other tasks is not sufficient.

Second, the lookup table results indicate that simple concept memorization is a competitive strategy if the amount of memorized concepts is sufficiently large and the evaluation scenario contains concepts seen in training. This is somewhat contrary to our initial assumptions, as we expected concepts to be more context-dependent, changing their relevance more frequently depending on the surrounding texts. Therefore, it is a valuable insight with respect to the first research question showing us to what extent textual features in varying textbooks influence concept relevance and to what degree relevance remains stable over different text passages.

Because concepts mostly remain relevant or irrelevant in the given data for different texts, one may investigate the following task variation in future work. Suppose the concepts keep, as the data suggests, their relevance across all text passages and different books. Consequently, classifying multiple text passages jointly and determining an overall relevance of the concept for the text across all text blocks may be a better task representation than sequence tagging. As a result, one would have significantly more textual features and thus more data to decide a concept's relevance. In turn, this may lead to more accurate concept extractors, which then can be used in the answer selection of educational Automatic Question Generation systems.

Third, even though memorization is a good strategy, the machine learning approaches also perform well and also learn to generalize unseen concepts. Hence, while the lookup table is a valid strategy, learning educational concept extraction is reasonable, particularly if only a few relevant concepts are known in advance. In consequence, the experiments provide evidence that textual features support the detection of learning-relevant information in educational texts.

7.6.3 *Learning-relevant Answers from Automatically Constructed Corpora*

The results allow us to estimate how these systems can already be applied to extract learning-relevant concepts for upstream systems such as educational Automatic Question Generation pipelines. We gather two main insights based on the token-level and phrase-level results. For one, the token-level results suggest that if the learning scenario deals with fixed domains that are known in advance, a lookup table is a straightforward approach that can hardly be outperformed by machine learning-based approaches (*Random Split* case). Hence, we would argue that answer selectors for learning-relevant concepts based on memorization are sufficient when addressing this scenario. That is, for fixed domain settings, the first research question likely not requires machine learning-based concept extractors.

Yet, the more we move the application scenario into an open domain setting, not knowing all concepts in advance, the better the machine learning models perform compared to the lookup table. Hence, we assume that such answer selectors are particularly helpful in an open domain setting, for instance, if we aim to build an educational Automatic Question Generation for most science texts. In turn, the more open domain the application scenario of an educational Automatic Question Generation is, the more likely it profits from machine learning-based concept extraction.

However, although the phrase-level results imply that the machine learning approaches perform better than the other approaches in the open domain setting, they still struggle with accurate concept selection. Specifically, in the worst-case scenario (*Unseen Concepts* split), they achieve F1 scores which are not sufficient for fully-automated content selectors, in our opinion. The worst-case scenario is contrived, and a degree of memorization is likely possible in real-world use cases. Still, based on the results, we argue that from an answer selection point of view, an open domain educational Automatic Question Generation system is likely only viable by utilizing a human-in-the-loop approach or additional data sources such as back-of-the-book indices (see Chapter 8).

8 QUESTION GENERATION IN READING COMPREHENSION SCENARIOS

In the following chapter, the gathered insights of the previous chapters are combined as we design and implement an educational Automatic Question Generation (AQG) approach with educationally driven content selection usable for science textbooks. We seek to investigate the second research question, to which extent educational AQGs can support reading comprehension, with the proposed approach by conducting two empirical studies. First, an intrinsic annotation study examines the generated questions' quality regarding linguistic and educational dimensions. Second, an extrinsic reading comprehension case study investigates to what extent the proposed approach scaffolds learners' text comprehension while reading scientific texts.

The following key insights from the previous chapters are utilized. We analyzed linguistic AQG capabilities and discussed crucial content selection challenges and potential solutions for building educational AQG approaches, supporting learners in a text comprehension scenario. In Chapter 4, the study data gathered attested neural question realization approaches a high linguistic quality. We concluded that those question realization algorithms are likely useful for educational AQG as soon as the content selection directs their outputs to learning-relevant information. Chapter 5 showed for context selection that causal sentences and definitions are weakly but reliably associated with perceived learning-relevant sentences. That is, they perform slightly better than other unsupervised methods from related work when detecting clear cases of learning-relevant sentences. With respect to answer selection, we established in Chapter 6 that machine learning approaches trained on noneducational corpora do not accurately transfer to educational datasets. In search of a solution to this issue, Chapter 7 introduced the idea of learning answers from educational textbook corpora constructed from back-of-the-book indices, which we regard as a curated list of learning-relevant concepts. We find that the machine learning approach works, yielding models generalizing to novel concepts. However, the classification quality is insufficient for a fully automated AQG approach without human verification.

8.1 THE EDUCATIONAL AUTOMATIC QUESTION GENERATOR

We design and implement an educational AQG approach for scientific textbooks in the English language. The input for the approach is a single chapter of a textbook comprising multiple paragraphs. The educational AQG generates $k \in \mathbb{N}$ questions per text paragraph. The number of generated questions k ranges from zero to dozens of questions depending on the length and concept density of the input paragraph. The approach follows the educational AQG process encompassing context selection, answer selection, and question realization (see Figure 19).

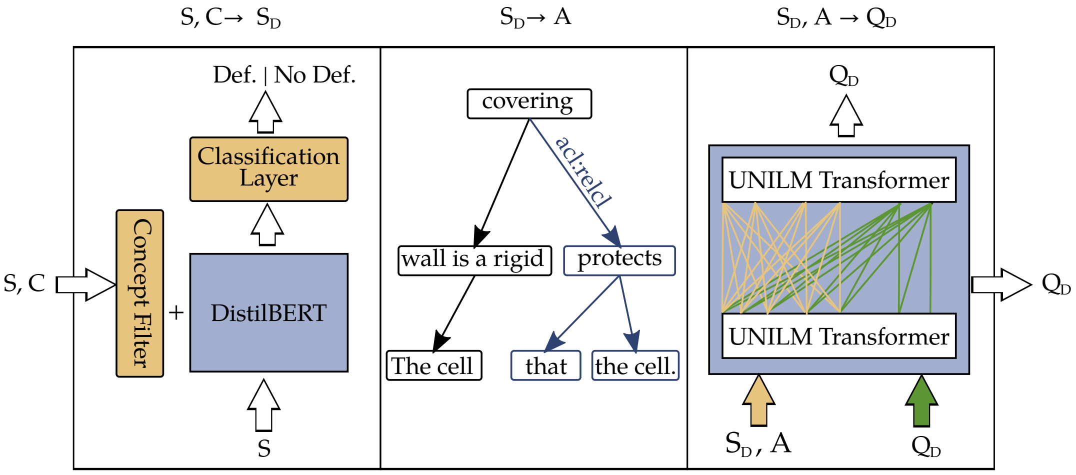


Figure 19: The overall architecture of our approach. Left, the context selection step for all sentences S and key concepts C using a keyword filter and a DistilBERT classifier [135]. Every sentence S is classified as definitory $S \in S_D$ or not. Center, the answer selection step based on semantic graph matching [16] on the dependency graph of a given definition. In the example, the words in the purple boxes are extracted as answer A because they constitute a relative clause (other relations are omitted for readability). Right, the UNILM transformer [32], which given answer A and definition S_D, generates a literal question Q_D. It applies a bidirectional mask to the given definition and answer and a left-to-right mask to the question under generation. The inputs A and S_D are represented as sequences of words. The sequence Q_D is autoregressively generated. This figure has been adapted from Steuer et al. [150].

8.1.1 *Context Selection*

The proposed context selection mechanism extracts definitions mentioning learning-relevant concepts stated in the back-of-the-book index. We focus on these due to multiple reasons.

Conceptual Design

First, we have seen in our analysis that text comprehension is concept-based, grounding this context selection approach in educational theory (see Chapter 3). Second, the experiments in Chapter 5 provided empirical data supporting the theoretical grounding. They showed that such definition-based context selectors are competitive systems extracting clearly learning-relevant sentences with high precision. The automatically detected definitions appeared disproportionally often in the perceived learning-relevant sentences and selecting them in the approach's evaluation experiments performed well. Third, we established that back-of-the-book index concepts are likely learning-relevant and can be automatically extracted from textbooks with high accuracy, allowing us to implement a concise learning-relevant concept filter for textbook inputs.

Approach

We address the context selection on the sentence level. Every chapter is sentence-segmented, and the context selector operates on all sentences S and the back-of-the-book index concepts C. It first filters all sentences not comprising any index concept from S, and then binary classifies whether each of the remaining sentences is of definitory nature. Consequently, the remaining sentences S_D are definitions which mention a learning-relevant concept from the textbooks' index. The back-of-the-book index was thereby manually extracted for the experiments. It could, however, also be automatically extracted using the *pdf-index-extractor* proposed in Chapter 7. The utilized definition classifier is analogous to the classifier described in Chapter 5. Yet, in contrast to the previous chapter, the classifier operates in English as our input data encompasses English textbooks. Thus, we do not use the custom collected German corpus for training but rely solely on the Definition Extraction from Texts (DEFT) corpus [143] because it is sufficiently large and contains a wide variety of definitions. The trained model is a DistilBERT [135] architecture. It achieves a phrase-level F1=0.78, Precision=0.78, Recall=0.78 on the DEFT corpus test set. These evaluation metrics show that the model is already quite accurate in binary definition detection for in-distribution data on the DEFT corpus. However, particularly on data not stemming from the DEFT corpus distribution, the model will sometimes still produce false predictions.

Learning Scenario Specific Adaptations

We assume that the different model errors are not equally problematic for learners. We built on results from human-computer interaction research investigating how imperfect artificial intelligence systems should be added to users' workflows. If possible, users prefer systems to fail with the error that causes the least harm to the workflow [72]. For instance, Kocielnik, Amershi, and Bennett [72] show in their work, if an intelligent calendar assistant highlights important appointments, users prefer the system fails by highlighting unimportant appointments. After all, these highlightings are easy to remove, whereas missing an important appointment due to lacking highlighting comes with higher costs. We transfer the line of thought to the context selection scenario. We argue that, on the one hand, low-quality questions will likely interrupt readers' text comprehension because they target irrelevant concepts, pointing readers' attention to unimportant information. On the other hand, missing questions are also problematic, but manually authored textbook questions also frequently do not entirely cover the text's important information. Consequently, readers should be more accustomed to missing questions, being aware that the available questions will not necessarily test their understanding of every relevant fact. Hence, we assume that a few high-quality questions are more important to users than generating more lower-quality questions and prefer precision over recall.

We already know from the study in Chapter 5 that extracting definitions as contexts yields learning-relevant contexts with comparably high precision. To improve upon this finding, we aim to adapt the trained classifiers' classification threshold to favor pre-

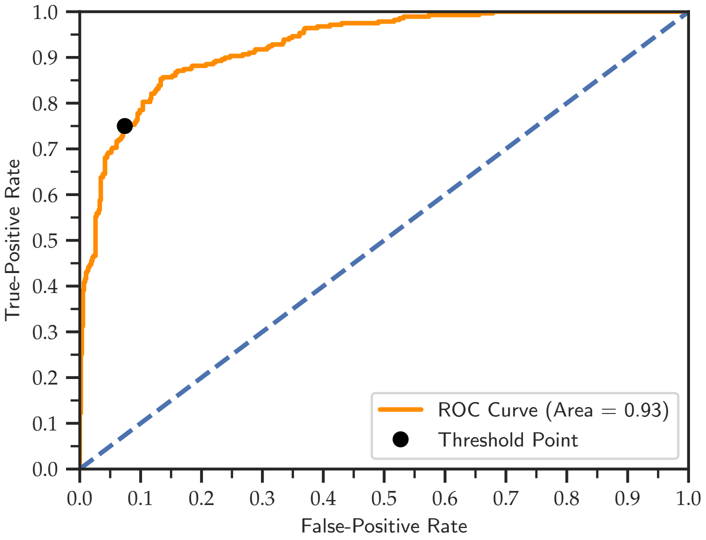


Figure 20: ROC curve for the DistilBERT-based classifier on the DEFT validation set. The threshold point is set to 0.7. This figure has been adapted from Steuer et al. [150].

cision over recall. A suitable threshold can be determined by observing the classifiers Receiver Operating Characteristic (ROC) plot as it depicts the classifier's true-positive and false-positive rates for different decision thresholds (see Figure 20). That is, it visualizes the trade-off between true-positives and false-positives. We visually choose 0.7 as the classification threshold. It yields a high true-positive rate of 0.75 with only a false-positive rate of 0.08. Shortly after, the ROC curve begins to flatten, and relatively speaking, fewer true-positives are detected by allowing more false-positives.

Furthermore, the classification threshold not only influences the true-positive and false-positive rates. In addition, it affects how many sentences are selected by the context selection and thus how many questions are generated. Consequently, we seek to understand how different classification thresholds influence the generated question count. To analyze this influence, we selected the first three chapters of six textbooks from varying domains and compared the amount of context selected sentences' given different classification thresholds (see Figure 21). The initial threshold is set to 0.5 and is increased in 0.05 steps. We can see in Figure 21 that the generated question count decreases slowly with increasing classification threshold. The context selection still selected many sentences per book at the chosen threshold of 0.7, not dropping below 50 questions per book. This supported the assumption that the chosen classification threshold will not hinder the AQG process by providing too few relevant sentences to the subsequent process steps. That is, from a context selection point of view, the data suggests sufficiently many relevant and question-worthy sentences will be selected.

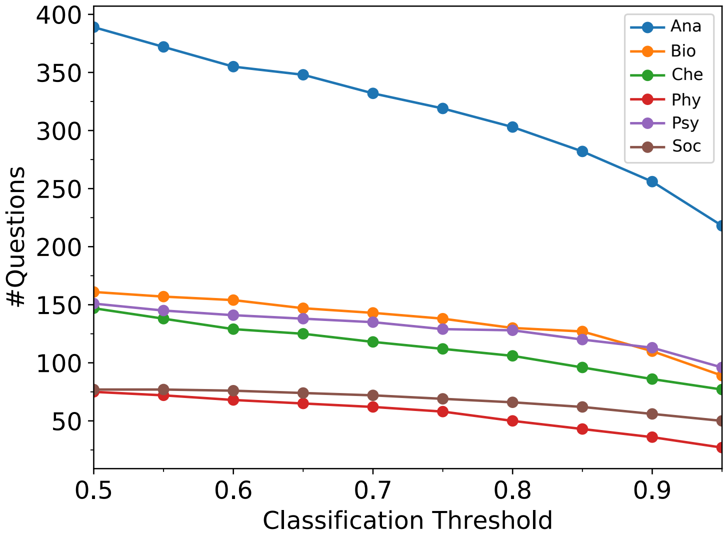


Figure 21: The question count decreases slowly with increasing classification threshold on the observed books in various domains: anatomy, biology, chemistry, physics, psychology and sociology. This figure has been adapted from Steuer et al. [150].

8.1.2 *Answer Selection*

Answer selection aims to extract an expected answer from a given sentence to guide the question realization process.

Conceptual Design

The inputs to the answer selection step are definitory sentences to concepts from the back-of-the-book index. Next, the answer selection aims to extract relevant sentence parts that constitute a meaningful answer to the question being generated for every sentence. We have established in Chapter 6 that a sole selection of noun phrases as answers is sometimes insufficient for generating educational questions. Therefore, the proposed educational AQG generates questions not solely about the noun phrases encompassed in the selected sentences. Instead, the answer selection aims to extract phrases describing the characteristics of selected definitions' concepts.

Approach

Due to the approach's design, the input sentences for the answer selection step are definitory in nature and comprise the concept to be defined and its characteristics. Recalling these characteristics reactivates propositions in the readers' mental model, thereby fostering the concepts' memorization (see Chapter 3). We heuristically designed the approach around linguistic constructs used to further describe sentences' subject.

Precisely, we extract dependency relations of relative clauses, adjectival clauses, and clausal complements from the context sentences (see Figure 19 middle). Dependency relations have already been used sucessfully to add important information to the sentence [101, 117]. Moreover, we observed heuristically that relations often cover the characterization of concepts in definitions. We implement this answer selection strategy with Semgrex matching (see Chapter 4) and extract the longest subgraph of the dependency tree. Additionally, we implement a fallback condition of direct objects whenever we cannot find any other fitting tag.

Pattern Coverage

We compute the coverage of the proposed patterns on the DEFT corpus. After all, the answer selection pattern must frequently match even for syntactically different definitions. If not, we will miss relevant sentences in the upstream question realization phase because we can only generate questions for sentences where we have a selected answer. We count every pattern match on the corpus validation split, and if multiple patterns match a sentence, we only count the first match. In total, the respective patterns cover 79% of all sentences, and the fallback condition accounts for 16% of the 79% of sentences covered. We deem this coverage sufficiently high to continue with our experiments.

8.1.3 *Question Realization*

After the content selection phases, we have a selected context sentence and an expected answer therein. This information is used to guide answer-aware neural question realization to derive a valid question given these inputs.

Conceptual Design

We learned from the study in Chapter 4 that the UNILM neural question generator [32] has good linguistic capabilities. In our study, annotators judged most of its outputs as linguistically acceptable and frequently answerable. Thus, we employ UNILM as the question realization approach in the proposed educational AQG approach, relying on the English snapshot provided by the UNILM authors.

Approach

The inputs to the neural question generator are the concatenated context sentence and the expected answer separated by the *[SEP]* token. As a decoding procedure during inference, we rely on greedy decoding with a beam size of 1 and a max sentence length of 48 tokens. That is, during each autoregressive generation step, we use the most likely token as the next token for the prediction. While other more advanced decoding strategies based on sampling exist (e.g. [58]), they are often nondeterministic, and we, therefore, opted against them, aiming for better reproducibility and traceability of our approach. The resulting question is only accepted if it ends with a question mark, and other generated sequences are discarded. The final output of the proposed educational AQG contains a set of triplets of the form (s, q, a) for every paragraph, where s is the context sentence, q is the generated question and a is the selected answer. In consequence, the proposed AQG allows enhancing a textbook on its paragraph level with generated text comprehension questions. It not only provides the questions but also anchors these questions in the paragraph via the context sentence and provides an answer for every question generated.

In a potential usage scenario, the questions are next shown to the textbook authors during editorial phases such that they may select the most appropriate exemplars for inclusion in the book. Moreover, depending on the question quality and learners' ability to identify nonfunctional questions, these questions may be directly shown to the readers for self-study. In consequence, we next evaluate the affordance of the proposed AQG approach and which usage scenarios arise from the generated questions' quality characteristics.

8.2 evaluation studies

We conduct an intrinsic annotation study and an extrinsic learner-based evaluation study of the proposed educational AQG.

The intrinsic study gathers insights into the generated questions' linguistic and educational quality characteristics via expert annotation of the generated questions. It is essential to have a human evaluation, as we have seen in our analysis in Chapter 2 and in Chapter 4 where we found that automatic measures are unreliable for estimating the quality characteristics of the generated questions. In contrast to an extrinsic study, the intrinsic study seeks to estimate the questions' inherent quality characteristics detached from a specific learning scenario. The question-inherent quality criteria are thereby helpful because they allow drawing informed inferences about the proposed AQG improvement potential and its application scenarios. For example, if the generated questions were linguistically sound but occasionally off-topic, an author-facing recommender system might be more appropriate than a user interface directly appealing to learners. Moreover, such data would point towards an improvement in content selection and not question realization.

However, intrinsic evaluation alone is often insufficient, and the actual application scenario needs to be studied to see whether the inferences drawn from the intrinsic data were correct (see Section 2.2.5). Thus, a subsequent learner-based extrinsic evaluation study complements the intrinsic study results. It aims to understand the viability of posing generated questions in an actual reading comprehension learning scenario. For that, it assesses the impact of the generated questions on the learning outcome during a reading comprehension session. It is complementary to the intrinsic study because it is unclear which intrinsic question criteria are necessary to support learners' reading comprehension. Moreover, it provides a more objective view on the generated questions' educational qualities as it is not based on the learners' subjective question judgments and generally operates with more participants.

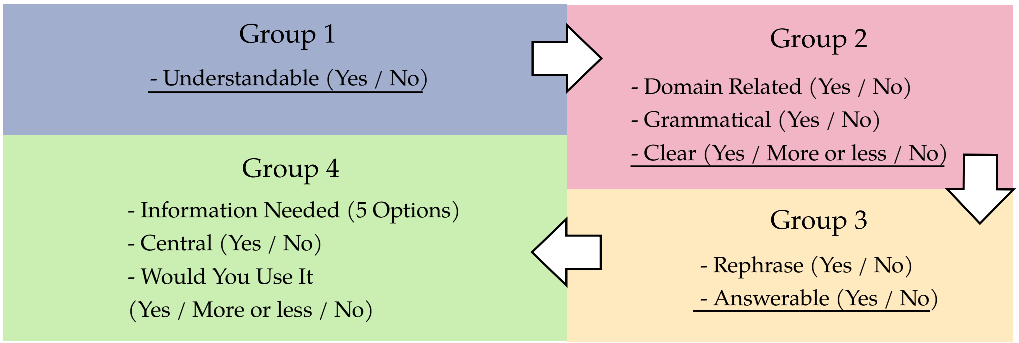

Group 1
- Understandable (Yes / No)
Group 2
- Domain Related (Yes / No)
- Grammatical (Yes / No)
- Clear (Yes / More or less / No)
Group 3
- Rephrase (Yes / No)
- Answerable (Yes / No)
Group 4
- Information Needed (5 Options)
- Central (Yes / No)
- Would You Use It
(Yes / More or less / No)

Figure 22: The hierarchical annotation scheme stops whenever an underlined evaluation item is answered with "no".

8.3 INTRINSIC EVALUATION STUDY

In the following, we introduce the intrinsic evaluation study and describe the used question dataset statistics.

8.3.1 *Material*

We utilize the freely available educational textbooks published by OpenStax for the study [146]. The textbooks have been written by subject matter and educational experts and cover introductory lessons at high school and university levels. The study is conducted on six different domain textbooks representative of various science domains from which we use three chapters per textbook for question generation. The domains comprise: anatomy, biology, chemistry, physics, psychology, and sociology. Although the textbooks are available as Portable Document Format (PDF) version, we decided to extract the texts for each generation run manually from the OpenStax web interface to prevent extraction artifacts such as faulty line breaks or paragraphs interfering with the educational AQG pipeline. Additionally, we only focus on the textbooks' main contents and ignore any figures, tables, captions, exercises or textual inserts. The proposed educational AQG is independently applied to the first three chapters of each of the six books. We apply stratified sampling to its results, randomly selecting 25 questions from each of the six domains to represent every book equally in the annotation data. Hence, in total, 150 questions are annotated in the conducted evaluation study.

8.3.2 *Annotation Scheme and Annotators*

After reviewing different annotation practices, we opted to apply the scheme of Horbach et al. [59] in the study because it overlaps well with our design analysis (see Chapter 3 and Figure 22). It measures the generated questions' linguistic and educational qualities on nine items that accurately cover our design analysis's linguistic and educational requirements. Six items are binary (yes/no), and the remaining items are categorical.

The items are structured as a hierarchy of four groups: *general understanding* (group 1), *linguistic appropriateness* (group 2), *answerability* (group 3), and *educational appropriateness* (group 4). The *general understanding* group is congruent to the concept of semantic understandability in Section 3.3.1. Moreover, the *linguistic appropriateness* group measures the spelling and grammar requirement defined in Section 3.3 and captures the initial text relatedness of the questions. Finally, the *answerability* group covers the corresponding requirements from Section 3.3.2, and the *educational appropriateness* encompasses items we deemed important in our educational requirement analysis in Section 3.2.

On the group level, the annotation may stop if, after annotating a hierarchical level, its answers indicate that the question quality is too low to annotate the remaining levels. For instance, a question judged as not understandable on the lowest level will not be annotated for *linguistic appropriateness*, *answerability*, or *educational appropriateness*. That is, if a question is incomprehensible, then there is no point in trying to estimate other quality characteristics. We opted for this schema due to the following reasons. In contrast to the pure linguistic Likert items frequently used in other AQG research, the schema is not only measuring linguistic quality aspects but also educational aspects. Moreover, it measures them on a fine-grained level on nine items. Besides, the hierarchical structure reduces the annotation efforts and the noise in the annotation process. The hierarchical construction helps to understand better which dimensions lead to the errors since questions on higher hierarchy levels are only annotated if they have sufficient quality on the lower levels.

We recruited four annotators for the study, working independently on the annotation. However, one annotator was excluded post-study due to a systematic violation of the annotation scheme. After observing strong negative effects on the Inter-Annotator Agreement (IAA) when including his or her data, we qualitatively reviewed the annotations. We determined a systematic misinterpretation of the clarity and answerability evaluation categories. Hence, we report the annotator statistics for the three remaining annotators. Each annotator has been studying in educational sciences or a field closely related at the university level for at least two years. Their mean age was 29 years, and they all speak English on the C1 level of the Common European Framework of Reference for Languages (CEFR) according to their formal qualification. They annotated the 150 questions with the schema independently of each other over approximately two weeks. Before the annotation started, annotators received detailed guidelines describing the different annotation categories and providing examples for every category. They were allowed to pose questions about the scheme and the guidelines if they had any. Moreover, the annotation guidelines could be consulted at any point in the annotation process. While annotating, they saw the textbook chapter from which the questions were generated jointly with the context sentence and the generated questions for this chapter. Every generated question was judged on all of the annotation scheme's dimensions before annotating the next question.

Item	%-Agreement	Krippendorff's α	Confidence Interval
Understandable	0.81	0.35	[0.27, 0.42]
Domain Related	0.74	0.28	[0.21, 0.34]
Grammatical	0.70	0.30	[0.23, 0.35]
Clear	0.60	0.25	[0.20, 0.30]
Rephrase	0.53	0.19	[0.16, 0.23]
Answerable	0.67	0.22	[0.15, 0.24]
Information Needed	0.42	0.18	[0.14, 0.22]
Central	0.57	0.26	[0.21, 0.31]
Would You Use It	0.41	0.16	[0.12, 0.20]

Table 16: The inter-annotator agreement of the three experts on the 150 questions of our evaluation study. The Krippendorff's α's are calculated on all 450 observations. The confidence interval of Krippendorff's α is estimated via the bootstrap.

8.3.3 *Inter-Annotator Agreement*

We report the three annotators' IAA in Table 16 comprising measures for percentage agreement as well as Krippendorff's α. The IAA computation handles the hierarchical nature of the annotation scheme by introducing an answer choice "not applicable" which is used when one annotator stopped the annotation before the respective item. This accounts for annotators deciding that a question should not be annotated at this hierarchical level in the first place. Otherwise, we would lose this information in the IAA measure. Moreover, we report the 0.95 confidence interval for Krippendorff's α, computed using the bootstrap method similar to Zapf et al. [182] using 10,000 iterations.

The annotators achieved only slight agreement on the overall task, and the annotation schemes' lower hierarchical levels have better IAA than the higher levels. Thus, we must be careful when generalizing the study's result to larger readerships as the IAA suggests that generated questions' perceived quality differs for most of our annotators. This restriction due to low IAA is not unexpected and is relatively common in Natural Language Generation (NLG) studies [165]. For instance, Horbach et al. [59] reports that the scheme was difficult to apply in their study as well, yielding only limited agreement between experts and almost no agreement between crowd-workers. They conclude that their scheme requires expert annotation, as in our study, because crowd-workers do not possess the necessary skills to perform the annotation task. Moreover, some authors argue that low IAA has to be expected when evaluating NLG approaches. After all, language is multi-faceted and is therefore assessed differently by evaluators based on their prior knowledge and understanding of language. For this reason, we believe that although IAA is essential for interpreting the generalization capabilities of the results, IAA should not be regarded as the sole evaluation metric for NLG.

Book	#Sections	#Paragraphs	#Questions	M_q	SD_q	Mdn_q
Overall	325	1,989	865	2.66	3.61	2
Anatomy	96	451	335	3.49	4.14	2
Biology	60	285	143	2.38	3.16	2
Chemistry	44	312	118	2.68	4.75	1
Physics	43	348	62	1.44	2.44	0
Psychology	53	325	135	2.55	3.09	2
Sociology	29	268	72	2.48	2.43	2

Table 17: Generation statistics on different books. The mean (M_q), standard deviation (SD_q) and the median (Mdn_q) indicate the respective statistical value per section (q=question). For instance, on average, we generated 2.38 questions for each of the 60 sections in the biology textbook. This table has been adapted from Steuer et al. [150].

8.3.4 *Generated Questions Dataset*

The proposed educational AQG should work on textbooks from various domains generating multiple questions per chapter. This is not self-evident, as texts from different domains and authors have different structures and writing styles. Thus, we automatically analyze the generated questions' count by textbook. Every book contains titled sections consisting of multiple paragraphs. The generation statistics on the section level can be found in Table 17.

In total, 325 sections with 1,989 paragraphs have been extracted from the first three chapters of the six textbooks. The paragraph lengths in words vary with an average length of 505.28 words ($Mdn_{para} = 332\,words$). Sociology has the longest paragraphs with on average 781.55 words, and anatomy has the shortest paragraphs with on average 369.52 words. Per section on average 2.66 ($Mdn = 2$) questions have been generated overall, with a standard deviation of 3.61. Two domains stand out, physics and anatomy. On the one hand, in physics, we generated comparatively few questions per section ($M = 1.44, Mdn = 0$). We attribute this to the frequent use of mathematical notation and definitions in the book, which is not covered by the proposed AQG. On the other hand, in anatomy, we generated comparatively many questions per section ($M = 3.49, Mdn = 2$). In the book, body parts are frequently connected with their names, which fits well with our definitory AQG approach.

Aside from the plain question number, we analyzed the question wh-word distribution to roughly estimate the type of questions we generate. Similar to our results in the piloting study, the questions remain primarily factual (see Chapter 4). In total, 90% of the questions are *"What"* questions. These questions start to 53% with *"What is"* and another 13% with *"What are"* and 12% with *"What does"*.

8.3.5 *Annotation Study Results*

The averaged intrinsic evaluation studies results of all three annotators are depicted in Figure 23. It is important to note that the annotated question count decreases with increasing the annotation scheme level. As already explained, a question is not further annotated if it is considered unsatisfactory at a lower level. Thus, we report two percentages in text in the form *(relative % / absolute %)*. The relative percentage is computed using the remaining number of questions at the hierarchy level, whereas the absolute percentage is calculated based on the total number of questions. Next, we describe the annotation results in detail based on the averaged results. Hence, if we state that some percentage of questions is linguistically sound, this does not necessarily imply that all annotators regard the same questions as linguistically sound.

General Understanding and Linguistic Appropriateness

The generated questions' linguistic quality is measured primarily on the first and second levels of the annotation scheme. Most of the questions are comprehensible according to the annotators (83% total). Still, 17% of questions are incomprehensible and sorted out at the lowest level of the annotation hierarchy. The remaining 83% of questions are frequently free of grammatical errors (88% / 73%) and are usually perceived as *clear* (78% / 64%) or *more or less clear* (20% / 17%). The generated questions' linguistic variance is also illustrated in the examples seen in Table 18. Many of the questions follow relatively simple questioning schemes such as *"What is X?"*. Therefore, these questions read similar to questions generated by template-based methods. Yet, in contrast to templates, the proposed approach operates without hand-crafted rules on texts from different domains, even when the sentences are complex (see Example 2-3 in Table 18). Furthermore, the approach is surprisingly fault-tolerant. In the Example 3 in Table 18 the selected context is false-positive and lacks definitory content. The answer selection then selects the answer *"increased accessibility and accommodation for people with physical handicaps"* which is only one of the crucial roles described in the context sentence. Nevertheless, the question realization step infers a meaningful question regarding the content selection inputs. It even detects that the characteristic provided as an answer is not exhaustive and phrases the question accordingly. The annotators judged the resulting question valuable, although the context and answer selection did not provide the best inputs. This result further reinforces the fluency of neural AQGs as we have discussed in Chapter 2 and in Chapter 4, which is advantageous in open domain settings. However, not all generated questions are linguistically flawless. One typical mistake we observed in the generated questions' data is the repetition of a specific phrase, as seen in Example 4 in Table 18 , which is a typical artifact of neural text degeneration [58]. Besides, the overall approach sometimes leads to grammatical plausible but semantically meaningless sentences such as Example 5 in Table 18.

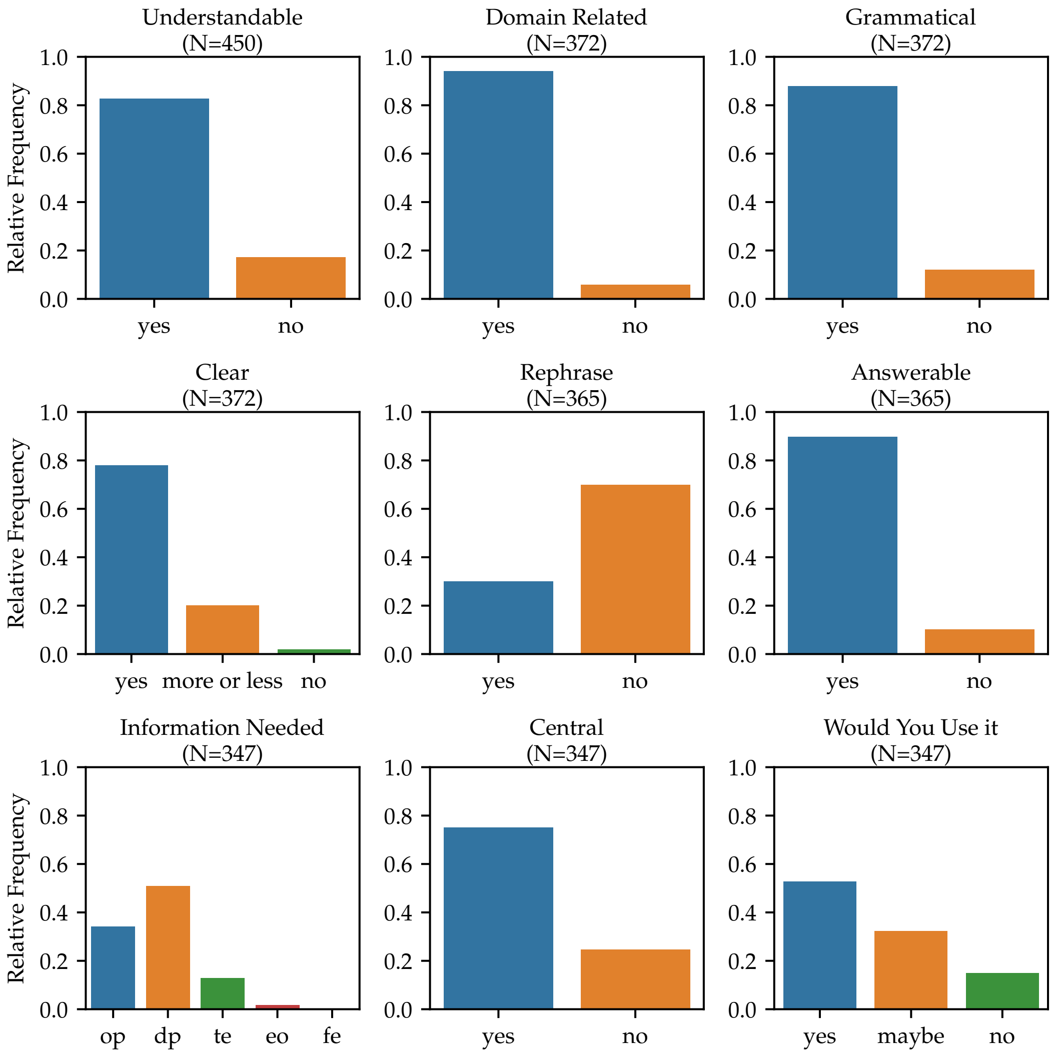


Figure 23: Answer distributions in the evaluation study based on the three experts' annotations. The N indicates how often the given evaluation category was applicable. Hence, N decreases four times from category one of the annotation scheme (upper left corner) to category four of the annotation scheme (lower right corner). For the information need frequencies the following abbreviations are used: one position in the text (op), different positions in the text (dp), text + external knowledge (te), external knowledge only (eo) and feelings and beliefs of the readers (fe). This figure has been adapted from Steuer et al. [150].

#	Rating	Context Sentence	Generated Question
1	useful	*Metabolism* is the sum of all anabolic and catabolic reactions that take place in the body (Figure 1.6).	What is Metabolism the sum of?
2	useful	Sociologists often study culture using the *sociological imagination*, which pioneer sociologist C. Wright Mills described as an awareness of the relationship between a person's behavior and experience and the wider culture that shaped the person's choices and perceptions.	How do sociologists study culture?
3	useful	Besides desegregation, sociology has played a crucial role in many important social reforms, such as equal opportunity for women in the workplace, improved treatment for individuals with mental disabilities or learning disabilities, increased accessibility and accommodation for people with physical handicaps, the right of native populations to preserve their land and culture, and prison system reforms.	What is one way sociology has played a role in social reforms?
4	incomprehensible	The cell stores energy in the synthesis (*anabolism*) of ATP, then moves the ATP molecules to the location where energy is needed to fuel cellular activities.	Where do ATP molecules move the ATP molecules to?
5	incomprehensible	Less blood circulating means reduced blood pressure and reduced perfusion (penetration of blood) to the brain and other vital organs.	What is another way to describe less blood?

Table 18: Examples of questions rated useful for teaching (example 1,2,3) or incomprehensible (example 4,5). In case the definition classification correctly identified a definition, the definiendum is marked in italic. The good questions where annotated as useful for teaching, whereas the bad questions were filtered out as incomprehensible. This table has been adapted from Steuer et al. [150].

Text Relatedness and Answerability

The proposed educational AQG approach frequently generated domain related questions (94% / 78%) that were answerable by the expert annotators (89% / 72%). Therefore, we find evidence that the proposed answer selection by extracting specific dependency subgraphs provides enough information for the question realization step to generate meaningful questions related to the text. Although UNILM's training data often encompasses shorter noun phrases connected to different dependency subgraphs, the question realization step still generalizes to the longer phrases from the subgraphs extracted by the proposed approach. However, these generalization abilities are restricted, and faulty answer selection can cause inadequate question generation, as Example 5 in Table 18 illustrates. In the example, parsing the sentence into a dependency graph malfunctioned, so an incomplete part of the sentence was selected as the answer, resulting in a semantically meaningless question.

Besides the general text relatedness, the annotation scheme captures where the information for answering the generated questions can be found. According to the proposed conceptual design, all answers should be found directly in the text, as we only incorporate text knowledge into the generation pipeline. However, the pretraining of large neural networks allows them to generate sentences stating information not mentioned in the text [99]. Hence, if the network struggles with the selected contexts or answers, it may generate a fallback question that cannot be answered from the text. The expert ratings indicate that this is usually not the case, and answers were found in precisely one text place (51% / 39%) or at multiple places (34% / 26%). The rest of the questions may be trivial to answer, or are plausibly phrased but without referencing the text and thus require only text external knowledge or ask for the reader's beliefs. That the answer to many questions can be found in multiple places can be explained by the general structure of the underlying textbooks. They frequently comprise coreference chains after introducing a concept to explain it further. While these additional explanations are not directly considered by the proposed educational AQG, they are often still helpful for answering definitory questions about the underlying concept. Accordingly, answers to the generated questions can be found in the initial definition but also in the auxiliary sentences surrounding the definition.

Educational Appropriateness

Determining the generated questions' educational quality characteristics is the most intricate and subjective according to IAA. Nevertheless, the results of the respective evaluation items in the *educational appropriateness* group provide a rough estimate of the questions' helpfulness in educational scenarios. In the intersection of text relatedness and educational utility, we can observe that more than half of all questions (75% / 57%) target central information in the text that is relevant for learning. These results imply educational value since the proposed educational AQG approach aims to improve text comprehension of core concepts and their characteristics and many of the generated questions target such core concepts.

Apart from the centrality, annotators state that they may use approximately 66% of all generated questions in an educational reading comprehension scenario. Particularly 53% / 41% of questions have been marked as usable, and another 32% / 25% of questions have been marked as may be usable. While we know that these results are subjective, we would like to point out that all annotators possess experience in educational sciences. Therefore, we suspect that they seek high-quality in their reading comprehension questions and that they sort out faulty questions rather quickly. Under this assumption, the disagreement more likely comes from potentially useful questions marked as too poor by some experts. Consequently, we consider these results mildly positive. The proposed question generator generated educationally useful questions in more than half of the cases according to the quality standards of educational expert annotators. Yet, we are aware that this annotation study provides only an initial approximation of the generated questions' quality, and learner-based studies are needed to obtain a more refined estimate of their usefulness in a reading comprehension scenario.

8.3.6 *Study Limitations and Validity Considerations*

The reported intrinsic evaluation study yields an initial characterization of the generated questions and their quality. However, as with most empirical studies, it has a limited scope and comprises some experimental tradeoffs we would like to point out.

First, it has to be noted that the study is explorative and focused on the quality of the proposed AQG approach without evaluating other approaches. This was a necessary tradeoff because the annotators had only limited resources, and annotating another approach's output was infeasible. Therefore, comparing the study results with results from the related work becomes difficult. Yet, even if we had the resources for annotating questions generated by a related approach, a fair comparison is challenging. We intentionally chose another path to educational AQG than related work. Instead of considering every sentence a potential question target or learning to select contents implicitly, we only focus on a limited subset of the texts' sentences because we aimed to have a content selection grounded in educational theory. A fair comparison must account for this conceptual difference because the amount of generated questions and their intended use for learning may differ. Besides, we deliberately don't report automated metrics such as Bilingual Evaluation Understudy (BLEU) scores. Thus, one cannot simply compare, for instance, this study's BLEU score with another study's BLEU score. Additionally, we previously observed that while these metrics seem objective, they provide little insight into the generated questions' actual merits (see Chapter 4).

Second, we annotated 150 questions from the first three chapters of six different domain textbooks. Although better than a study on a single domain or a single text, this covers only a tiny fraction of potential textbooks. In consequence, the collected results are not easily transferable across all science domains and textbook styles. Instead, we view them as evidence of the generated questions' quality on natural science texts used particularly in the first university semesters. Whether or not the results transfer to more advanced study material, dissimilar domains such as economics or even other languages should be investigated in future work.

Third, the annotation study was conducted with educational experts instead of linguistic experts. The IAA was relatively low, and the provided judgments were somewhat subjective, similiar to the results we gathered in the annotation study in Chapter 4. Nevertheless, we are convinced that relying on educational experts instead of linguistic experts for annotation was reasonable because they are the potential users of an educational AQG approach. Thus, we believe that we traded a higher IAA for a better ecological validity [33, p. 106] of the experimental results. Nonetheless, the low IAA poses a challenge to the result's interpretation often encountered in NLG research [165]. Therefore, future work should take the reported numbers as initial quality estimates and conduct extrinsic reader-based studies before applying the approach in learning scenarios. Next, we conduct such an extrinsic study for science reading comprehension to see how the generated questions affect learners and to complement the findings from this intrinsic evaluation.

8.4 extrinsic evaluation study

The intrinsic evaluation study has provided us with initial quality estimates for the proposed educational AQG's generated questions. The results suggested a high linguistic quality of the generated questions. However, the experts also expressed rather subjective assessments, especially concerning the questions' educational qualities. Although they attested a certain educational quality to the questions, they often disagreed about their centrality and usability in reading comprehension scenarios. Therefore, we conduct an extrinsic study to measure the questions' educational utility in a reading comprehension study from the learner's point of view. The study assesses the impact of the generated questions on the learning outcome during a reading comprehension session. For that, we recruit 48 college students for an in-between-groups reading comprehension case study on two texts. We test two hypotheses concerning the questions' potential positive and negative effects on text comprehension. Moreover, we explore how the questions influenced the learners descriptively. This approach complements the intrinsic expert study by studying fewer texts and domains while collecting data from actual learners in a reading comprehension learning scenario.

8.4.1 *Study Design*

The study follows an in-between subject, pre/posttest, treatment/control design. In accordance with the reading comprehension literature, we call the generated questions shown during the reading comprehension *adjunct questions* for the remainder of the chapter. They are adjunct as they are added to support readers' text comprehension.

The study comprises two groups. A control group participates in the reading comprehension without adjunct questions, and a treatment group answers six adjunct questions during the reading comprehension. The study's primary goal is to measure the adjunct questions' effect on learning outcomes. We operationalize learning

Type	Variable
independent	adjunct question condition (yes / no)
dependent	related learning outcome
	unrelated learning outcome
other	time-on-task
	language skill
	prior knowledge
	student answers to the adjunct questions
	adjunct questions' perceived author (computer / human)

Table 19: Variables elicited in the experiment. This table has been adapted from Steuer et al. [155].

outcomes into two distinct variables and measure additional variables for descriptive analysis and control (see Table 19).

The first kind of learning outcome measured is *related learning outcome*. It assesses to what extent subjects remember information about concepts for which they have seen an adjunct question during their reading comprehension. The second kind of learning outcome measured is *unrelated learning outcome*. It assesses to what extent subjects remember information about concepts not targeted by any adjunct question. Related works find that manually authored adjunct questions usually increase the related but not the unrelated learning outcome [51]. Hence, we would expect an increase in related learning outcomes in our study with automatically generated questions. Moreover, related works suggest that adjunct questions may decrease learning outcomes under specific circumstances [51]. For instance, adjunct questions before the reading material may prime learners' attention to specific concepts, leading to impaired learning of concepts not targeted by the questions. We cannot exclude that automatically generated questions may also introduce adverse learning effects due to their imperfections. For instance, one could argue that linguistically problematic questions distract learners and divert their attention from the learning material. Two hypotheses emerge from this reasoning:

H1 The treatment group will have a higher related learning outcome compared to the control group

H2 There is no difference between the treatment and control group in the unrelated learning outcome

To control for group differences, we record the following variables. First, the participants' language proficiency is recorded. The study is conducted in English, but we sample a large portion of the subjects in a German university. Thus, the participants' English language skills may differ between the control and treatment groups which could affect how they perform in reading comprehension. Second, the participants' prior knowledge is collected. Suppose there were differences between the prior knowledge of the treatment and control group. In that case, it becomes difficult to determine whether this difference in prior knowledge confounds potential question effects. Third, the participants' time-on-task is registered. The variable is complex. On the one hand, it could be a confounding factor since one could assume that the longer someone works with the text, the higher their score on the posttest. On the other hand, it is affected by the treatment because answering the adjunct questions requires time that is not needed in the control group. Therefore, we record it mainly to see if a time difference occurs. Determining the extent to which the difference is causally related to the questions is not part of the analysis.

Besides the hypotheses, we collect two additional variables to descriptively analyze the generated questions' effects. The generated questions' answerability is gathered during the study. Participants may remark for any question they encounter that it is incomprehensible or unanswerable from the text. The resulting variable measures the minimal linguistic properties of the generated questions. After all, only questions perceived as answerable by participants can lead to a learning effect. Note that we do only measure if a question was answered. We do not analyze if the provided answer was correct. Furthermore, it is captured whether the subjects perceived the questions as being authored by a human or computer. The variable captures aspects of questions' perceived linguistic and educational quality because readers expect linguistically sound and educationally meaningful questions from educators. It allows us to explore whether there is an association between the perceived questions' author and their usefulness for reading comprehension.

Participants

The study participants are sampled from three primary sources.

Some participants are recruited via survey circles. A survey circle is a social network group in which scientists seek experimental subjects for their studies. The social network relies thereby on mutual study participation. That is, research team A or one of their subjects participates in the study of researcher B, which leads research team B or one of their subjects to participate in the study of research team A. We rely on monetary compensated student assistants to participate in the respective studies of other researchers to gather participants. Additionally, roughly a third of the participants are recruited via email lists of the local university. The list comprises the department's students from current and previous semesters. Participation was incentivized via a lottery, offering the chance of winning 5x20 Euro Amazon gift cards. Participation in the lottery was voluntary, and some participants chose only to participate in the experiment but not the lottery. Last, we recruited participants directly in a single university lecture. In the lecture, potential participants received two incentives. First, they were allowed to participate in the lottery of gift cards. Second, participation counted as an exercise submission for the course, and participants who earned sufficient points in exercises received a bonus on their exam score. Taking part in the incentive program was again voluntary. Moreover, if students chose not to participate, we provided an alternative exercise, ensuring no negative effects arose from not participating.

The external rewards for participants may lead to them not being diligent in their work and only seeking to complete the experiment quickly to receive compensation. Thus, the experiment included three control questions designed to detect cheating attempts. In case of failing a control question, the participant's data was not considered for the experiment, and no reward was given to the participant. The recruitment process resulted in 57 participants, of which 48 remained after evaluating these control questions, randomly assigned to the control group (N=24) and treatment group (N=24). The sample encompassed 17 females, 30 males, and one non-binary subject with an average age of 24 years (SD=3.6). Most participants were enrolled in their second university semester, whereas the rest was enrolled between one and fifteen semesters at university, except four participants who had already finished their studies. Participants studied computer science or related (N=34), electrical engineering or related (N=12), or psychology (N=2). The language skill was measured via self-report because we did not want every participant to take an additional language placement test, as this disproportionately prologues the experiment. Language proficiency was high, with 83% of all participants stating B2 (upper-intermediate) or better as their skill level in CEFR. The lowest reported level was A2 (elementary) which only one participant reported, and the highest was C2 (mastery), reported by seven participants.

Material

The material comprises two reading comprehension texts accompanied by two self-report pretests and two single-choice posttests.

The two reading comprehension texts originate from the biology and anatomy domain and were published in corresponding OpenStax [146] textbooks for undergraduate education. We revised them before the experiment to guarantee they function as self-contained reading lectures concerning their topics. The biological text covers the Eukaryotic Cells in approximately 1,400 words, whereas the anatomical text covers the Layers of the Skin in about 1,700 words. Both texts introduce the respective topic, teaching essential concepts and core ideas. Their comprehension does not require advanced knowledge in biology or anatomy from readers. Still, the texts and domains were deliberately chosen in such a manner that the contents of the texts are probably new to readers who are inexperienced in the domain. Hence, it is unlikely that participants in our sample of non-biology or anatomy students will be familiar with the texts' contents. Consequently, we expect that all participants will have almost no prior knowledge before starting the experiment.

In the treatment group, every text included six generated adjunct questions. They were generated using the proposed educational AQG where all bold-faced words in the texts served as index concepts for the generator. We divided each text into three almost equal-length passages and selected two generated adjunct questions per passage for the participants to answer. The selected questions were chosen randomly, and we deliberately did not consider properties such as syntactic correctness or answerability.

Yet, there was a single section where we regenerated two questions because the initial questions were semantically incomprehensible after inspection.

The self-report pretests were constructed for each of the two texts. Each pretest contains five five-point Likert items from one (know nothing) to four (expert), allowing participants to indicate the extent to which they are familiar with the text's key concepts. We have chosen self-reporting over an actual knowledge test due to the priming effects described in related work. The meta-review by Hamaker [51] indicates that pretest questions guide the readers' attention towards inquired information. Thus, readers focus on the questioned concepts during reading comprehension and neglect other concepts. A knowledge test as a pretest would likely introduce such an effect in the participants, altering their reading patterns during the following experiment. Therefore, we opted for self-reporting, although self-reported information is likely less accurate than knowledge tests. We furthermore believe that less accurate reporting is less significant in the experiment because we deliberately selected reading material where most subjects should have no knowledge, not requiring a fine-granular prior knowledge assessment.

The posttests were constructed separately for the two reading comprehension texts. They comprised a total of 15 manually authored, single-choice items in three groups. Six items targeted related learning outcomes, and six items targeted unrelated learning outcomes. Besides, the three control items to detect cheating attempts formed a separate group and were randomly inserted between the twelve actual test items. The related posttest items were manually authored after the AQG generated the adjunct questions. Consequently, they are guaranteed to inquire about information related to the adjunct questions. However, no posttest item was a direct copy of an adjunct question. Instead, they either paraphrased an adjunct question or asked about a closely related idea or concept.

We revised the pretest items, posttest items, and reading materials three times, with 3-10 experts giving feedback each time. We clarified the instructions, exchanged unfitting posttest items and distractors, and made the reading materials easier to read. The revisions did not affect the generated questions in any way. They were generated initially and were never revised manually.

Procedure

We conducted the experiment over six weeks entirely online. The experimental procedure is depicted in Figure 24. Participants were given an initial two-week deadline for participation to ensure timely participation. They could extend the deadline if they required more time. Each participant was informed about the incentives, the timeframe (ca. 45 minutes), and the collected data before starting the experiment. They received a contact address for potential debriefing and inquiries about the collected data.

The experiment started with the collection of demographic information. Afterward, detailed instructions outlining the task and the experimental sequencing were given, including a call to work diligently without cheating. Furthermore, the instructions indicated that the reading materials were purposefully unfamiliar and challenging. Finally, two online translation services to translate unknown words were provided,

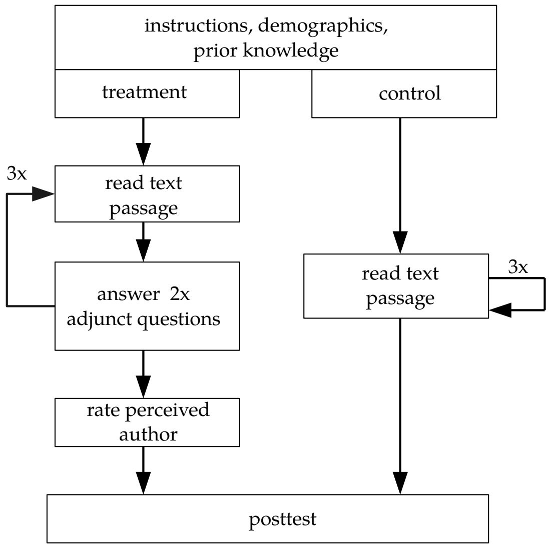


Figure 24: An overview of the experimental procedure followed. Each group read three text passages. Additionally, the treatment group answered two adjunct questions after every text passage. This figure has been adapted from Steuer et al. [155].

and the treatment group was reminded to answer every adjunct question in their own words.

Next, participants' prior knowledge was collected via self-report before the actual reading comprehension started. The respective readings (biology or anatomy) were shown during the experiment as three similar length passages. Every participant read only a single text, with participants randomly assigned to their text condition. The readings were the same for the treatment and control groups except for the two adjunct questions the treatment participants answered after every passage. Participants read one text passage before continuing to the next, and they could not navigate back after continuing to the next passage. No timeboxes were given, and participants read at their own pace.

After reading comprehension, the treatment group participants decided for every seen adjunct question if it was manually authored or computer-generated. They did not know that all questions were computer-generated but were explicitly instructed that this was possible. Next, all subjects participated in the posttest and answered the respective 15 items. The experiment concluded with an information page allowing participants to voluntarily state their email addresses to participate in the lottery. For the participants sampled in the lecture, an additional in-person debriefing was conducted where the experimental design and measures were described.

8.4.2 *Reading Comprehension Study Results*

We begin by reporting descriptive results and continue by evaluating the hypotheses. For better interpretation, every test score is normalized between 0 and 1. For instance, the posttest yields 12 points maximum (6 related + 6 unrelated learning outcomes), and a score of 3 out of 12 points is reported as 0.25.

Descriptive Results

The study yields the following descriptive results.

Sample Structure

The study comprises data of N=48 participants. We initially divided them according to two experimental factors, the reading material (biology or anatomy) and whether a subject participates in the control or treatment group. Participants read the biology text 21 times (treatment=10 and control=11). They scored between 0.16 and 0.80 points in the corresponding posttest. The anatomy text was read 27 times (treatment=14 and control=13), and the scores varied between 0.44 and 0.83. No floor or ceiling effects [33, p. 251] were observed in either posttest. However, unfortunately, the sampling did not provide enough data points for a text-specific analysis. Therefore, we collapse the text factor into a single variable for the rest of the analysis. The resulting control and treatment groups consist of 24 participants each.

Confounders

The potential confounders hardly differ per group. The control group has a mean age of 24.38 years (SD=3.15), and the treatment group has a mean age of 23.67 years (SD=4.08). Participants' language skills are M=0.71 (SD=0.20) in treatment and M=0.68 (SD=0.19) in the control group. Moreover, their prior knowledge is barely higher in the treatment group M=0.30 (SD=0.22) than in the control group M=0.28 (SD=0.21). Only the time-on-task differs considerably, with an average time on task of 34 minutes (SD=13 min) in the treatment group and 29 minutes (SD=14 min) in the control group. However, we expected a difference between groups in this variable because answering questions increases the time compared to only reading the text.

Posttest Scores

We analyze the post-scores in three quantities based on the related learning outcomes, the unrelated learning outcomes, and their total sum (see Figure 25). The total scores in the treatment and control groups differ. The treatment group scores on average 0.62 (SD=0.22), whereas the control group scores 0.56 (SD=0.21). If we look at the two learning outcomes directly, we see that there is not much difference in unrelated learning outcomes between the control (M=0.51, SD=0.25) and the treatment group (M=0.49, SD=0.30). In contrast, a more pronounced difference occurs in the related learning outcomes, scoring M=0.76 (SD=0.19) for the treatment group and M=0.62 (SD=0.25) for the control group.

Perceived Author Ratio & Answerability

Besides the learning outcomes, we are interested in how readers perceived the generated questions. Particularly, we measure how often they could provide an answer to

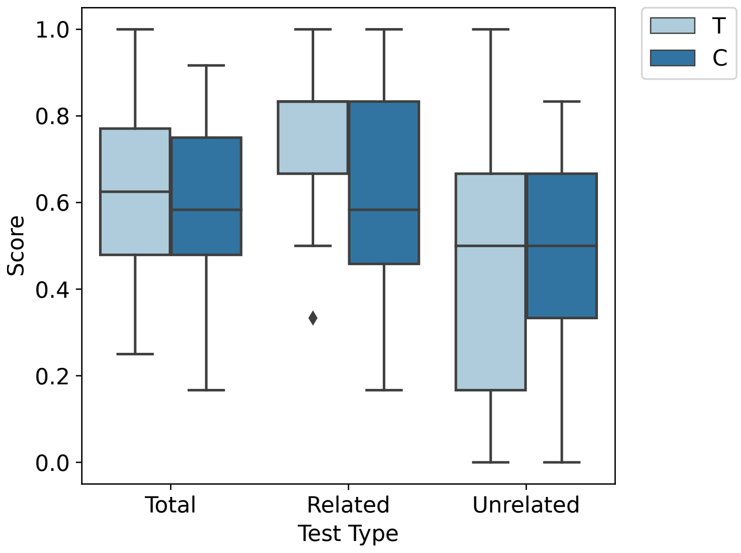


Figure 25: Boxplot of the different test scores of the treatment (T) and control (C) groups. Whiskers indicate 1.5 interquartile range and the bar inside the box indactes the median. The median of the *Related* group is at the top of the box. This figure has been adapted from Steuer et al. [155].

the generated questions and if they perceived them as manually authored or computer-generated. The information was only gathered in the treatment group. Participants tried to answer a total of 144 questions. Only in six cases did participants indicate that the question was incomprehensible, therefore not providing an answer. The incomprehensible rating was five times given to the same question (see Table 20; Example 3). Furthermore, all answer attempts seemed sincere and plausible to the experimenter during a brief qualitative review of the results. Answers often comprised single concepts or ideas but varied in wording and length. Concerning the perceived question author, Table 20 reports participants' average perception. It shows the ratio of participants perceiving a question as manually authored. On average, participants considered half of the six questions they saw manually authored (M=2.95, SD=1.12). Depending on the subject, one to six questions were perceived as manually authored. The best question fooled nine out of ten participants to rate it as manually authored (see Table 20; Example 8).

Effects of Perceived Author on Related Posttest Scores

Next, we investigate if questions that were perceived more frequently as manually authored were associated with higher learning outcomes. Such analysis yields insight into whether we need to generate questions that are as close to manually authored questions as possible before learners can profit from them.

The analysis builts on the fact that the control and the treatment group see identical posttests. The related posttest items are associated with corresponding adjunct questions in the treatment condition. Thus, under simplified assumptions, ignoring cofounders, we assume that the *control-treatment score difference* of a related postest

#	Question	Perceived Author Ratio
1	The skin and its accessory structures make up what system?	$\frac{6}{14} \approx 0.43$
2	What are cells in all of the layers except the stratum basale called?	$\frac{8}{14} \approx 0.57$
3	What happens to the growth of fingerprints in a growing fetus?	$\frac{6}{14} \approx 0.43$
4	How are keratinocytes formed?	$\frac{10}{14} \approx 0.71$
5	What does the hypodermis serve?	$\frac{5}{14} \approx 0.36$
6	What cells produce melanin?	$\frac{8}{14} \approx 0.57$
7	What does the plasma membrane control?	$\frac{3}{10} \approx 0.30$
8	What is the structure of the cytoplasm?	$\frac{9}{10} \approx 0.90$
9	Where is the nucleoplasm located?	$\frac{8}{10} \approx 0.80$
10	What does Chromatin describe?	$\frac{2}{10} \approx 0.20$
11	What do scientists often call mitochondria?	$\frac{2}{10} \approx 0.20$
12	What are mitochondria?	$\frac{4}{10} \approx 0.40$

Table 20: The 12 adjunct questions used in the experiment and their perceived author ratio. A high ratio indicates that the question is perceived more often as manually authored. We show the preceived author ratio as a fraction to clarify how many raters have seen each question. This table has been adapted from Steuer et al. [155].

item stems from the active engagement caused by the respective adjunct question (see Section 8.4.3 for the limitations). Moreover, every adjunct question has a *perceived author ratio* associated with it. The ratio quantifies how often a question was perceived as manually authored and how often as computer-generated. Suppose whether or not a question is perceived as manually authored is vital for the learning outcomes. Then, we would expect that the *control-treatment score difference* and the *perceived author ratio* are associated with each other. After all, questions more frequently perceived as manually authored should increase the learning outcome more.

Given this idea, Figure 26 visualizes the variables' association in the data. The Y-axis indicates the *contral-treatment score difference* for the related learning outcome questions. The higher the value, the better the treatment group performed on the respective question compared to the control group. The X-axis shows the *perceived author ratio*. The higher the value, the more often a question was perceived as manually authored. However, the plot does not indicate any relationship between the two variables for the 12 data points of the two related posttests (biology and anatomy). Instead, the data scatter fairly randomly. Thus, we conclude that there is no strong connection between the questions' perceived author and their effects on learning. Therefore, generating natural questions perceived as manually authored may not be necessary to improve learning in all cases.

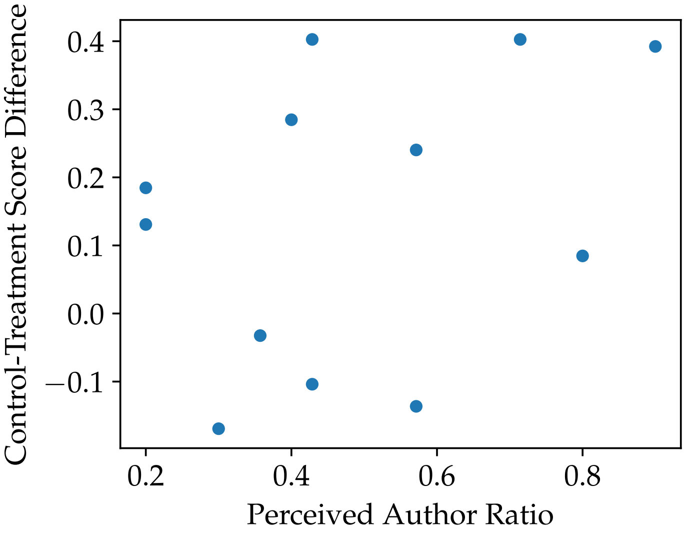


Figure 26: Scatterplot of the *control-treatment score difference* and the *perceived author ratio*. The y-axis indicates the mean difference in the correct answer ratio for the control and treatment groups of every posttest item. The x-axis shows their mean perceived author ratio. The higher the rating the more participants rated the question to be human-written. This figure has been adapted from Steuer et al. [155].

Hypotheses Tests

We test two hypotheses with a significance level of $\alpha = 0.05$.

First Hypothesis

The first hypothesis (H1) states that we have a higher related learning outcome in the treatment group. As the data is not normally distributed as visually confirmed by Figure 27, we used a Mann-Withney-U Test to test for statistically significant group differences. The test result indicated a significant higher posttest score for the treatment group (M=0.76, Mdn=0.83) compared to the control group (M=0.62, Mdn=0.58, U=194.5, p=0.049). The corresponding confidence interval for the related posttest mean difference ($M_{\mathtt{diff}} = 0.14$) estimated by the bootstrap with 10,000 iterations is $CI_{0.95} = [0.01, 0.24]$. According to Cohen's d, we have an effect size of d=0.63, implying a medium effect [137]. Similarly, it corresponds to an effect size of ES=0.14 posttest point differences in the meta study by Hamaker [51]. Given these results, we accept the first hypothesis.

Second Hypothesis

The second hypothesis (H2) states that the unrelated learning outcomes of both groups are equivalent. That is, the null hypothesis assumes that the groups are different, and the alternative hypothesis is an equivalence hypothesis. Therefore, we apply a two-one-sided-test procedure based on the Mann-Whitney-U test to test whether the two

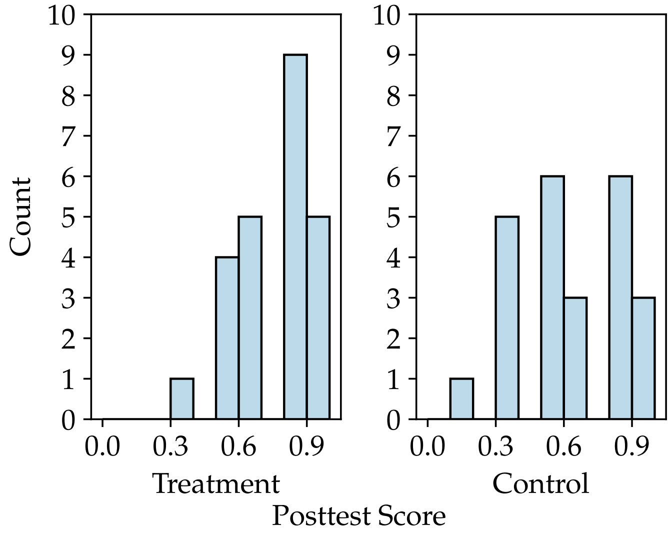


Figure 27: Histograms of the related posttest scores.

group scores are equal [176, pp. 126-136]. We follow the liberal equivalence tolerance of Wellek [176], allowing a maximum of $\epsilon = 0.2$ standardized mean difference on both sides. The test is insiginificant on the $\alpha = 0.05$ level for an equivalence of control group (M=0.51, Mdn=0.5) and treament group (M=0.49, Mdn=0.5) with $W_{+} = 0.45$, $\sigma[W_{+}] = 0.08$ and $C^{\alpha}_{MW} = 0.08$. For a technical discussion of these parameters see Wellek [176, pp. 126-136]. The corresponding unrelated posttest mean difference is $M_{\texttt{diff}} = -0.02$ with a 10,000 iteration bootstrap confidence interval of $CI_{0.95} = [-0.15, 0.10]$. Thus, we cannot accept the second hypothesis. However, the confidence interval for the mean difference is rather broad, and descriptively, both groups' mean and median values are relatively close. Consequently, we assume that our study might be underpowered (see Section 8.4.3), and the amount of data is insufficient to draw definitive conclusions regarding the hypothesis.

8.4.3 *Study Limitations and Validity Considerations*

We have carefully designed the study and selected the material. Furthermore we increased the study's validity through feedback from the three pilots. Nevertheless, there are validity constraints which we would like to discuss.

First, we are dealing with a case study with 48 participants. The sample size is fairly small, although similar group sizes are used in other studies in this area [13, 50, 130, 170]. Due to the sample size, the findings are certainly not as clear-cut as with larger samples because the study only has low statistical power [12]. Thus, the case study gives us initial evidence that the proposed educational AQG may work in reading comprehension scenarios, but further studies with a larger sample should validate the findings.

Second, the study employs two texts stemming from similar domains. In turn, it is not excluded that the findings do not generalize to other domains or texts. In particular, we chose texts containing multiple definitions, allowing the educational AQG to find targets in the text. The generator will probably work differently for advanced science texts focusing primarily on the interpretation of established concepts. Yet, definition-rich texts are common in university-level science curricula, particularly during the first semesters, when students often find it challenging to engage with new learning conditions. The study addresses these cases and shows the value of the educational AQGs.

Third, there is a confounding factor due to the participants' varying language skills. Although they frequently speak English well, they are still non-native speakers and probably perceive the texts and questions differently because they lack linguistic intuition. This difference in perception might influence their answering of questions and their rating of the perceived question author. However, we assume this was not a crucial confounding factor in the present study. Questions and texts encompassed a linguistic complexity that almost all participants should not struggle with as they spoke English for several years and used it frequently in their school and university careers.

Next, experimenting online due to the pandemic affected controlling for external confounders in the experiment. We informed participants to work deliberately and under calm conditions, but we could not control it. Furthermore, we used the control questions to find the most severe violations of the experimental procedure, resulting in the exclusion of nine participants. However, we cannot quantify to what extent other disturbing factors such as noise or interruptions played a role for some participants.

Finally, we report a potential bias in the perceived author item. None of the participants rated all questions as computer-generated. We assume they tended not to set all values equal in the binary scale, although they were informed that this is a valid possibility. Consequently, the values are probably somewhat biased towards the middle.

8.5 DISCUSSION OF THE MAIN INSIGHTS

In the following, we will first discuss the two studies' results (see Section 8.3 and Section 8.4) and connect them to the second research question.

8.5.1 *Intrinsic Study*

The study provided insights into the linguistic and educational quality of the generated questions.

Linguistic Quality

The evaluation reveals the fairly good linguistic quality of most questions and various generated questions per text, even though we entered texts from different scientific domains into the approach. This is promising, as a decent number of generated questions, is helpful in practice because multiple literal questions are often used in textbooks as quick warm-up questions before readers try more complex tasks.

Moreover, the linguistic quality ought to exceed a minimum because actively engaging learners with the text requires at least comprehensible questions. Given the data, we assume that questions are usually phrased in comprehensible language and can be interpreted by the person reading them in the context of the text. After all, most questions are annotated as understandable, free of language errors, and related to the text.

Nonetheless, it has to be noted one out of six questions was incomprehensible. Furthermore, the questions often were worded in a similar form, which might hinder the motivation of students working with them for longer. Overall, we nevertheless conclude that, based on the intrinsic data, the linguistic quality of the questions is sufficient for scaffolding readers' active text engagement. Consequently, depending on the learning situation and the ability of the learner to reflect, we could imagine that the questions could be used directly in a learning scenario. This would require making it clear to the learner that the questions are generated, and thus they may encounter misleading questions. We assume, learners with sufficient language competence and self-regulation can quickly get over incomprehensible questions. However, this requires that the good questions not only engage learners arbitrarily with the text but inquire about relevant information to foster text comprehension.

Educational Quality Characteristics

In terms of educational quality characteristics, the data provides the following insights. Most questions are related to the text, and answering them requires information from one or more places in the text. Furthermore, according to experts, the answers to the questions are often helpful for text comprehension as they cover crucial text aspects. That is, learners must actively recapitulate what they have read concerning the key concepts to construct an answer to the generated questions. Asking about the concepts activates them in the learners' mental representation, improving their retention. This speaks for the educational usefulness of the generated questions. Yet, it shall be recognized that the questions cover single concepts and their characteristics. Although this is good for fostering the core concepts in the reader's mental model, integration and reasoning questions are also required to improve the mental model's overall concept integration. The questions in the proposed educational Automatic Question Generation are unable to accomplish this. However, other question objectives, such as the causal sentences we extracted in Chapter 5, could complement the approach here, allowing the activation of multiple concepts together.

In total, the experts indicated for about two out of five questions that they would like to use them in a learning scenario, and for two out of three questions, they would maybe use them in a learning scenario. The most significant improvement opportunity lies in the questions' centrality and precise formulation. The experts remarked in feedback interviews that although the questions are linguistically sound and understandable, one would often not formulate a question exactly like this in a learning scenario. Based on all this data, we think we see a trend in the intrinsic evaluation data that points to a certain educational quality of the questions.

8.5.2 *Extrinsic Study*

Questions' Positive Effects

The first hypothesis stated that the generated questions' have a positive effects on the related learning outcome. According to the descriptive analysis confounding factors are distributed almost equally between the group and thus do not affect the group analysis considerably. Based on the statistical test, we accept the first hypothesis. Statistical evidence indicates improved related learning outcomes because of the generated questions. That is, generated questions improve learners' text comprehension in our reading comprehension scenario compared to the passive control group. Hence, the results attribute a similar effect to computer-generated questions as previous studies to manually authored literal questions [4, 51].

The effect size in terms of related posttest point difference is ES=0.14, which is comparatively large for a related question posttest. The meta study by Hamaker [51] measures on average ES=0.16 for repeated posttest items and ES=0.07 for related items. However, in the meta-study, related questions comprise multiple item types and consist, for example, of paraphrases, items targeting similar information, or higher-order items about the concepts. Consequently, we hypothesize that this study's related posttest items are more similar to repeated questions than some items included in the average estimation in the meta-study. This would explain the comparatively large effect size we measure in this study.

Besides, learners' time-on-task is higher in the treatment group. We suspect that answering the questions required them to engage more in-depth with the text resulting in more time spent. The additional time required is anticipated but may weaken the generated question effects. Subjects may experience a fatigue effect when the experiment takes too long, decreasing their posttest performance. For example, suppose that after 25 minutes, subjects put less effort into the experiment, worsening their results. Whether such an effect occurred is out of the scope of the study. We consider it unlikely due to the relatively short experimental time. However, investigating its existence would be an excellent opportunity for future work.

Questions' Negative Effects

We theorized that the generated questions in the worst case affect the unrelated learning outcomes negatively. Investigating this theory was difficult due to the study's sample size. The second hypothesis stated that the unrelated learning outcomes are equal in the treatment and control groups. However, we could not show their equivalence statistically, and thus we reject the second hypothesis based on the data. This is likely a limitation of our case study's small sample size due to the limited experimental resources. Future studies may look into detecting the effect statistically with large samples. For now, descriptively, the means and medians of the data from both groups are fairly similar, and the 95% confidence interval of the unrelated posttest score's mean difference is wide, including negative and positive values.

Thus, based on the case study's data were are optimistic that either the questions affected learning not negatively or that the negative effect was relatively small.

The Perceived Author and Its Effects On Learning

The author of the adjunct questions was frequently not easy to determine for the participants. The ratings are roughly split half whether the questions were generated or manually authored. We associate a sufficient linguistic quality with this data because the questions frequently deceived the participants, indicating their strong grammar, spelling, and text-relatedness. Yet, the data can not only be interpreted in this direction as it is confounded by factors such as participants' language skills. Take question three in Table 20 as an example. It was repeatedly rated as manually authored, although it is unnaturally phrased. Thus, we take the perceived author ratios as initial evidence for the questions' linguistic quality but acknowledge that it is not very nuanced and probably only indicates if the questions achieved at least minimum plausible fluency to trick readers. Moreover, combining the perceived author data with the questions' answerability posttest scores provides additional insights.

Participants answered almost all questions and were not discouraged when a question appeared computer generated. Furthermore, the answer attempts were all plausible and demonstrated that participants had taken another look at the text. That is, to actively engage learners, questions perceived as manually written were not necessary. These results are also supported by Figure 26. We can see there that whether a question was frequently identified as computer-generated had no direct effect on its effect on the learners' posttest performance. These results align with those of Van Campenhout et al. [163], who found that learners have no clear preference between computer generated and manually authored questions. Hence, based on the data, we theorize that even if questions' are perceived as computer-generated, they still actively engage learners and trigger learning-relevant cognitive effects. Consequently, generating questions of the highest linguistic and educational quality may not be required before they improve learners' text comprehension. Learners can probably connect even subpar questions with the surrounding text to foster their learning outcomes. However, the study is only a case study with relatively small sample size, and more extensive future work is required to validate this theory.

8.5.3 *The Support For Reading Comprehension Scenarios*

In conclusion, we infer the following findings regarding the second research question. The generated questions' linguistic quality remains sufficient over multiple science domains and in the expert and reader-based study. That is, the results of the piloting study carry over to the proposed Automatic Question Generation approach. So, strictly from a linguistic perspective, the proposed educational Automatic Question Generation approach is convincing and generates many syntactically correct and semantically interpretable questions. Consequently, the generated questions afford a general reflection of the inquired information, improving learners' active text engagement.

Regarding the afforded reflections' educational usefulness, experts find educational value in the questions measured by the educational annotation items. However, they often disagree on the questions' usefulness in the different evaluation items. Given the results, we are, optimistic that roughly half of the questions have the potential to facilitate text comprehension. The reader-based study provides further evidence for this interpretation and shows that the questions improved reading comprehension compared to a no-question control group. We, therefore, conclude that the questions not only cause random text engagement but stimulate the reader to reflect upon learning-relevant content. Thus they afford learning-relevant text engagement. Furthermore, there is first evidence that this is the case even if the questions were not as perfectly formulated as by a human author, which is an exciting opportunity for future work.

9 SUMMARY, CONCLUSIONS, AND OUTLOOK

This chapter summarizes our main contributions and draws conclusions based on our results. Finally, we discuss open issues and potential future work.

9.1 SUMMARY OF THE THESIS

The thesis investigated educational Automatic Question Generation (AQG) with respect to two research questions:

RQ1 To what extent can content selectors, based on textual features, extract learning-relevant information from educational science texts in different domains?

RQ2 To what extent can educational AQGs generate literal questions supporting reading comprehension of educational science texts in different domains?

We motivated the need for educational AQG and the corresponding research questions in Chapter 1. We defined their scope and highlighted the research challenges involved. Next, Chapter 2 conveyed the necessary technical background of AQG to the reader. Building on this background knowledge, we discussed relevant related works concerning the content selection and question realization phases and their strengths and weaknesses, yielding potential research gaps. In Chapter 3, we deduced our research approach based on conceptual works in learning-relevant content selection, linguistic question quality, and the technical research gap identified in Chapter 2. Next, we summarize the contributions we presented in this thesis's subsequent chapters.

Contributions to the First Research Question

We investigated the educational content selection for AQG on the context selection and answer selection level contributing to the first research question.

Our *first contribution* empirically shows that textual characteristics are associated with sentences' perceived learning relevance. We find the association for unsupervised summarization algorithms, text heuristics, and the proposed context selectors based on sentence type classification. Moreover, we observed that objectively selecting learning-relevant sentences was difficult even for humans, who disagreed considerably with their selection decisions. On sentences perceived as learning-relevant by many readers, approaches based on textual features successfully extracted numerous sentences.
Therefore, our first contribution provides evidence that in the absence of individual learner features, textual features provide a helpful signal for context selection of clearly learning relevant sentences. However, they should not be used in fully automated systems as their selections are likely too noisy. Our proposed context selectors have the

added benefits of being anchored in educational considerations and being flexibly extendable with other learning-relevant sentence types. Hence, we reckon their selection criteria are better understood by technical laymen such as many teachers or authors. This improves their benefits in human-in-the-loop scenarios where educators must understand why a question is being proposed.

Furthermore, our *second contribution* to the first research question demonstrates that text characteristics informing a machine learning approach's selection criteria differ significantly depending on the training corpora. Thus, if we design machine learning-based answer selectors purely on data without an educational assumption of relevant content, systems trained on different datasets will likely select answers according to different opaque criteria. Particularly, in our experiments, it appeared that models trained on noneducational data did frequently not find educational answers because the selection criteria differed from models trained on educational data. These results imply that implicit content selection based on text characteristics is dataset-dependent. We consider these dataset dependencies suboptimal since they worsen the systems' generalization and create non-transparency because the learned selection criterion is unclear. Therefore, we consider implicitly trained models without educational grounding insufficient as general purpose content selectors as sought by the first research question. To reduce such dataset dependencies, we argue that one must determine in advance which transferable selection criteria should be learned and construct datasets accordingly. The proposed context selection approach of our first contribution applies this reasoning to generate transferable and educationally grounded selectors.

Next, our *third contribution* indicates that if one wants to select educational concepts, the automatic construction of textbook concept corpora is a viable approach. Our constructed concept corpora entailed concepts from different domains and various books in two languages. They allowed machine learning approaches to learn educational concept selection beyond pure memorization. That is, with regard to the first research question, text characteristics provide information about whether a concept stated in texts should be selected as relevant by content selectors. Concerning educational AQG, this can be helpful for the context selection subtask and the answer selection subtask.

All in all, these contributions to the first research question indicate that text characteristics are helpful for content selection in educational AQG and establish vital cornerstones for their use.

Contributions to the Second Research Question

The thesis contributes the following insights regarding the second research question.

The *fourth contribution* comes in the form of our piloting study. We could show that the excellent linguistic quality characteristics of neural question generators transfer to educational texts. That is, it persists on out-of-distribution data. This is vital as a minimum linguistic quality is required for every educational question.

The *fifth contribution* encompasses the design, implementation, and expert evaluation of an educational AQG approach based on the insights from the content selection experiments. The expert evaluation of the resulting AQG painted a nuanced portrait of the generated questions and their linguistic and educational strengths and weaknesses. A considerable strength of the system is its high linguistic quality because we could generate high-quality questions on various textbook chapters across multiple domains. Thus, concerning the second research question, we would argue that the limiting factor for the practical use of educational AQG is not linguistic quality. Instead, the expert annotation results suggest that the questions' educational quality characteristics are occasionally suboptimal, yet these results are subject to significant disagreement between expert raters.

Finally, our *sixth contribution* comprises a learner-centric case study investigating generated questions' effects on readers' learning outcomes in a reading comprehension scenario. We gathered evidence that the questions increased learning outcomes compared to a no-question control group. These results provide more evidence that an educational AQG based on neural methods is not only generating linguistically pleasing but also educationally helpful questions.

All in all, these contributions to the second research question guide the application of educational AQG for reading comprehension, revealing their strengths and weaknesses in potential learning scenarios.

9.2 conclusions

The thesis examined automatically generating literal questions for science reading comprehension scenarios. We found that text characteristics facilitated the selection of question-relevant sentences and expected answers to guide the AQG process. According to our results, text characteristics which shall be considered relevant by the machine learning algorithms must be selected explicitly and in advance according to educational criteria. We found that relying on implicitly learned relevance criteria from data is suboptimal and results in opaque selection criteria not necessarily reflecting educational needs. We proposed educationally-grounded context selection and answer selection mechanisms and found that they are moderately associated with perceived learning relevance or annotated important concepts, making them suitable for human-in-the-loop systems.

Furthermore, we obtained encouraging results regarding the AQG's usefulness in education. Our reading comprehension study indicates that the questions contributed to increased learning outcomes. Moreover, experts attest that the generated questions have a high linguistic quality. However, the expert evaluation occasionally suggests not using a generated question for reading comprehension because it lacks linguistic or educational quality. In conclusion, we are optimistic that the proposed educational AQG can support textbook authors in human-in-the-loop scenarios. Still, there is a need for further research on the educational fit and the worst-case behavior of such educational AQG approaches.

9.3 OUTLOOK

The thesis investigated the potential of educational AQG systems. We infer from the results that the AQG quality is likely sufficient to support human-in-the-loop systems for authors and teachers. Future work should explore such human-in-the-loop systems thoroughly. Possible research directions include measuring the AQGs's effect on the authoring process and systematically exploring to what extent it reduces time and effort spent on authoring questions. Additionally, in human-in-the-loop systems, user acceptance of the system is not guaranteed [72]. Therefore, the factors influencing a successful integration of AQG in the authoring process are an excellent target for future work. Particularly, we argue in this thesis that transparent and explainable system decisions are helpful. Thus, they should be empirically validated in future work.

Additionally, we focused our research mainly on the English language. Hence, another exciting opportunity for future research is investigating the transfer of AQG systems into different languages. We explored this in Steuer et al. [149] by jointly using machine translation and an existing English AQG system. The next steps would be to address the problem with more sophisticated approaches such as multilingual models [124] or textual alignments [64]. After all, having easy transfer mechanisms for bringing AQG into other languages would open up novel target audiences as learners speak numerous languages apart from English.

Finally, we think educational AQG could be an interesting additional data source for learning analytics approaches. Learning analytics often uses rather noisy features such as click stream data or course material downloads to make predictions about learner behavior, such as student dropout or students' exam performance. In contrast, the answers to the generated questions constitute less noisy assessment data. They provide useful information about the student learning and can therefore be used to improve the efficiency and effectiveness of the learning process [18]. Hence, the answers to the generated questions of an educational AQG could help to receive better signals in many learning analytics approaches, which may result in more adaptive learning system designs [148, p. 182] due to more nuanced interventions. However, for this to scale, not only educational AQG must be investigated further but also more advanced automatic short answer grading methods that also provide feedback [42] as educators likely need help scoring all the generated questions.

Thus, our contributions toward the understanding and development of educationally sound AQG systems open up further research opportunities in the aforementioned directions.

ACKNOWLEDGMENTS

This research was funded by the *Bundesministerium für Bildung und Forschung* in the project: *Software Campus 2.0 Microproject: Textbook-AQG* and in the *Hessian State Chancellery* in the department of *Digital Strategy and Development* in the *Förderprogramm Distr@l*.

BIBLIOGRAPHY

[1] Isaac Alpizar-Chacon and Sergey Sosnovsky. "Order out of Chaos: Construction of Knowledge Models from PDF Textbooks." In: *Proc. 2020 ACM Symp. Document Engineering*. Virtual Event: Association for Computing Machinery, 2020, pp. 1–10. DOI: `10.1145/3395027.3419585`.

[2] Jacopo Amidei, Paul Piwek, and Alistair Willis. "Rethinking the Agreement in Human Evaluation Tasks." In: *Proc. 27th Int. Conf. Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 3318–3329.

[3] Richard C. Anderson. "How to Construct Achievement Tests to Assess Comprehension." In: *Review of Educational Research* 42.2 (1972), pp. 145–170. DOI: `10.3102/00346543042002145`.

[4] Richard C. Anderson and Barry W. Biddle. "On Asking People Questions about What They are Reading." In: vol. 9. Psychology of Learning and Motivation. Academic Press, 1975, pp. 89–132. DOI: `10.1016/S0079-7421(08)60269-8`.

[5] Hannah Bast and Claudius Korzen. "A Benchmark and Evaluation for Text Extraction from PDF." In: *Proc. 2017 ACM/IEEE Joint Conf. Digital Libraries*. Toronto, Canada: IEEE Press, 2017, pp. 1–10. DOI: `10.1109/JCDL.2017.7991564`.

[6] Lee Becker, Sumit Basu, and Lucy Vanderwende. "Mind the Gap: Learning to Choose Gaps for Question Generation." In: *Proc. 2012 Conf. North American Chapter Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, 2012, pp. 742–751.

[7] Anja Belz and Albert Gatt. "Intrinsic vs. Extrinsic Evaluation Measures for Referring Expression Generation." In: *Proc. ACL2008 Human Language Technologies, Short Papers*. Columbus, Ohio, USA: Association for Computational Linguistics, 2008, pp. 197–200.

[8] Anja Belz and Ehud Reiter. "Comparing automatic and human evaluation of NLG systems." In: *Proc. 11th Conf. European Chapter Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics, 2006, pp. 313–320.

[9] Rachel M. Best, Michael Rowe, Yasuhiro Ozuru, and Danielle S. McNamara. "Deep-level comprehension of science texts. The role of the reader and the text." In: *Topics in Language Disorders* 25.1 (2005), pp. 65–83. DOI: `10.1097/00011363-200501000-00007`.

[10] Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. "Taxonomy of educational objectives: The classification of educational goals." In: *Handbook I: Cognitive Domain* 1 (1956).

[11] Tom Brown et al. "Language Models are Few-Shot Learners." In: *Proc. 33rd Int. Conf. Neural Information Processing Systems*. Vancouver, Canada: Curran Associates, Inc., 2020, pp. 1877–1901.

[12] Katherine S. Button, John P. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. Robinson, and Marcus R. Munafò. "Power failure: why small sample size undermines the reliability of neuroscience." In: *Nature Reviews Neuroscience* 14.5 (2013), pp. 365–376. DOI: 10.1038/nrn3475.

[13] Aimee A. Callender and Mark A. McDaniel. "The benefits of embedded question adjuncts for low and high structure builders." In: *Journal of Educational Psychology* 99.2 (2007), pp. 339–348. DOI: 10.1037/0022-0663.99.2.339.

[14] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. "YAKE! Keyword extraction from single documents using multiple local features." In: *Information Sciences* 509 (2020), pp. 257–289. DOI: 10.1016/j.ins.2019.09.013.

[15] Leon Camus and Anna Filighera. "Investigating Transformers for Automatic Short Answer Grading." In: *Proc. 2020 Int. Conf. Artificial Intelligence in Education*. Virtual Event: Springer International Publishing, 2020, pp. 43–48. DOI: 10.1007/978-3-030-52240-7_8.

[16] Nathanael Chambers et al. "Learning alignments and leveraging natural logic." In: *Proc. 2007 ACL–PASCAL Workshop Textual Entailment and Paraphrasing*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 165–170.

[17] Branden Chan, Stefan Schweter, and Timo Möller. "German's Next Language Model." In: *Proc. 28th Int. Conf. Computational Linguistics*. Virtual Event: International Committee on Computational Linguistics, 2020, pp. 6788–6796. DOI: 10.18653/v1/2020.coling-main.598.

[18] Mohamed A. Chatti, Anna L. Dyckhoff, Ulrik Schroeder, and Hendrik Thüs. "A Reference Model for Learning Analytics." In: *International Journal of Technology Enhanced Learning* 4.5/6 (2012), pp. 318–331. DOI: 10.1504/IJTEL.2012.051815.

[19] Guanliang Chen, Jie Yang, and Dragan Gasevic. "A comparative study on question-worthy sentence selection strategies for educational question generation." In: *Proc. 2019 Int. Conf. Artificial Intelligence in Education*. Chicago, Illinois, USA: Springer International Publishing, 2019, pp. 59–70. DOI: 10.1007/978-3-030-23204-7_6.

[20] Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. "LearningQ: a large-scale dataset for educational question generation." In: *Proc. 12th Int. Conf. Web and Social Media*. Stanford, California, USA: Association for the Advancement of Artificial Intelligence (AAAI), 2018, pp. 481–490.

[21] Yi Cheng et al. "Guiding the Growth: Difficulty-Controllable Question Generation through Step-by-Step Rewriting." In: *Proc. 59th Annu. Meeting Association for Computational Linguistics and the 11th Int. Joint Conf. Natural Language Processing (Volume 1: Long Papers)*. Virtual Event: Association for Computational Linguistics, 2021, pp. 5968–5978. DOI: 10.18653/v1/2021.acl-long.465.

[22] Yuang Cheng et al. “WikiFlash: Generating Flashcards from Wikipedia Articles.” In: *Proc. 2021 Int. Conf. Neural Information Processing*. Sanur, Indonesia: Springer International Publishing, 2021, pp. 138–149. DOI: 10.1007/978-3-030-92273-3_12.

[23] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches.” In: *Proc. 8th Workshop Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 103–111. DOI: 10.3115/v1/W14-4012.

[24] Noam Chomsky. “Three models for the description of language.” In: *IRE Transactions on Information Theory* 2.3 (1956), pp. 113–124.

[25] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. “Don’t Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases.” In: *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 4069–4082. DOI: 10.18653/v1/D19-1418.

[26] Sérgio Curto, Ana Cristina Mendes, and Luisa Coheur. “Exploring linguistically-rich patterns for question generation.” In: *Proc. 2011 UCNLG+ Eval: Language Generation and Evaluation Workshop*. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 33–38.

[27] Diana Dee-Lucas and Jill H. Larkin. “Novice strategies for processing scientific texts.” In: *Discourse Processes* 9.3 (1986), pp. 329–354.

[28] Diana Dee-Lucas and Jill H. Larkin. “Attentional strategies for studying scientific texts.” In: *Memory & Cognition* 16.5 (1988), pp. 469–479. DOI: 10.3758/BF03214228.

[29] Takshak Desai, Parag Dakle, and Dan Moldovan. “Generating questions for reading comprehension using coherence relations.” In: *Proc. 5th Workshop Natural Language Processing Techniques for Educational Applications*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 1–10. DOI: 10.18653/v1/W18-3701.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In: *Proc. 2019 Conf. North American Chapter Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[31] Pedro Domingos. “A few useful things to know about machine learning.” In: *Communications of the ACM* 55.10 (2012), pp. 78–87. DOI: 10.1145/2347736.2347755.

[32] Li Dong et al. "Unified language model pre-training for natural language understanding and generation." In: *Proc. 32nd Int. Conf. Neural Information Processing Systems*. Vancouver, Canada: Curran Associates Inc., 2019, pp. 13063–13075. DOI: 10.5555/3454287.3455457.

[33] Nicola Döring and Jürgen Bortz. *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. 5th ed. Heidelberg, Germany: Springer-Verlag, 2016. DOI: 10.1007/978-3-642-41089-5.

[34] Xinya Du and Claire Cardie. "Identifying where to focus in reading comprehension for neural question generation." In: *Proc. 2017 Conf. Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 2067–2073. DOI: 10.18653/v1/D17-1219.

[35] Xinya Du and Claire Cardie. "Harvesting Paragraph-level Question-Answer Pairs from Wikipedia." In: *Proc. 56th Annu. Meeting Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 1907–1917. DOI: 10.18653/v1/P18-1177.

[36] Xinya Du, Junru Shao, and Claire Cardie. "Learning to Ask: Neural Question Generation for Reading Comprehension." In: *Proc. 55th Annu. Meeting Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1342–1352. DOI: 10.18653/v1/P17-1123.

[37] Nell K. Duke and David P. Pearson. "Effective practices for developing reading comprehension." In: *Journal of Education* 189.1-2 (2009), pp. 107–122. DOI: 10.1177/0022057409189001-208.

[38] Harold P. Edmundson. "New methods in automatic extracting." In: *Journal of the ACM* 16.2 (1969), pp. 264–285. DOI: 10.1145/321510.321519.

[39] Raquel E. Eisenkraemer, Antônio Jaeger, and Lilian M. Stein. "A systematic review of the testing effect in learning." In: *Paidéia* 23 (2013), pp. 397–406. DOI: 10.1590/1982-43272356201314.

[40] Saar Eliad, Ido Hakimi, Alon De Jagger, Mark Silberstein, and Assaf Schuster. "Fine-tuning giant neural networks on commodity hardware with automatic pipeline model parallelism." In: *Proc. 2021 USENIX Annu. Technical Conf.* Virtual Event: USENIX Association, 2021, pp. 381–396.

[41] Günes Erkan and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." In: *Journal of Artificial Intelligence Research* 22 (2004), pp. 457–479. DOI: 10.1613/jair.1523.

[42] Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. "Your Answer is Incorrect... Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset." In: *Proc. 60th Annu. Meeting Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 8577–8591. DOI: 10.18653/v1/2022.acl-long.587.

[43] Lei Gao, Prafulla K. Choubey, and Ruihong Huang. "Modeling Document-level Causal Structures for Event Causal Relation Identification." In: *Proc. 2019 Conf. North American Chapter Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 1808–1817. DOI: 10.18653/v1/N19-1179.

[44] Abbas Ghaddar, Philippe Langlais, Ahmad Rashid, and Mehdi Rezagholizadeh. "Context-aware adversarial training for name regularity bias in named entity recognition." In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 586–604. DOI: 10.1162/tacl_a_00386.

[45] Roxana Girju. "Automatic Detection of Causal Relations for Question Answering." In: *Proc. ACL2003 Workshop Multilingual Summarization and Question Answering*. Sapporo, Japan: Association for Computational Linguistics, 2003, pp. 76–83. DOI: 10.3115/1119312.1119322.

[46] Yoav Goldberg. "A primer on neural network models for natural language processing." In: *Journal of Artificial Intelligence Research* 57 (2016), pp. 345–420. DOI: 10.1613/jair.4992.

[47] Arthur C. Graesser, Zhiqiang Cai, Max M. Louwerse, and Frances Daniel. "Question Understanding Aid (QUAID) a web facility that tests question comprehensibility." In: *Public Opinion Quarterly* 70.1 (2006), pp. 3–22. DOI: 10.1093/poq/nfj012.

[48] Arthur C. Graesser, Vasile Rus, and Zhiqiang Cai. "Question classification schemes." In: *Proc. 2008 Workshop Question Generation*. Arlington, Virginia, USA, 2008, pp. 10–17.

[49] Anil Gupta. "Definitions." In: *The Stanford Encyclopedia of Philosophy*. Winter 2021. Metaphysics Research Lab, Stanford University, 2021.

[50] H. W. Gustafson and David L. Toole. "Effects of adjunct questions, pretesting, and degree of student supervision on learning from an instructional text." In: *The Journal of Experimental Education* 39.1 (1970), pp. 53–58. DOI: 10.1080/00220973.1970.11011231.

[51] Christiaan Hamaker. "The effects of adjunct questions on prose learning." In: *Review of Educational Research* 56.2 (1986), pp. 212–242. DOI: 10.3102/00346543056002212.

[52] Kyle Hamilton, Aparna Nayak, Bojan Bozic, and Luca Longo. *Is Neuro-Symbolic AI Meeting its Promise in Natural Language Processing? A Structured Review*. arXiv:2202.12205, preprint. May 2022. DOI: 10.48550/ARXIV.2202.12205.

[53] Richard J. Hamilton. "A framework for the evaluation of the effectiveness of adjunct questions and objectives." In: *Review of Educational Research* 55.1 (1985), pp. 47–85. DOI: 10.3102/00346543055001047.

[54] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. "DeBERTa: Decoding-Enhanced BERT with Disentangled Attention." In: *Proc. 2021 Int. Conf. Learning Representations*. Virtual Event, 2021.

[55] Michael Heilman and Noah A. Smith. "Extracting simplified statements for factual question generation." In: *Proc. QG2010: The 3rd Workshop Question Generation*. Pittsburgh, Pennsylvania, USA, 2010, pp. 11–20.

[56] Michael Heilman and Noah A. Smith. "Good question! Statistical ranking for question generation." In: *Proc. 2010 Conf. North American Chapter Association for Computational Linguistics*. Los Angeles, California, USA: Association for Computational Linguistics, 2010, pp. 609–617.

[57] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[58] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. "The Curious Case of Neural Text Degeneration." In: *Proc. 2019 Int. Conf. Learning Representations*. Virtual Event, 2019.

[59] Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. "Linguistic appropriateness and pedagogic usefulness of reading comprehension questions." In: *Proc. 12th Language Resources and Evaluation Conf.* Marseille, France: European Language Resources Association, 2020, pp. 1753–1762.

[60] Tom Hosking and Sebastian Riedel. "Evaluating Rewards for Question Generation Models." In: *Proc. 2019 Conf. North American Chapter Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 2278–2283. DOI: 10.18653/v1/N19-1237.

[61] Yan Huang and Lianzhen He. "Automatic generation of short answer questions for reading comprehension assessment." In: *Natural Language Engineering* 22.3 (2016), pp. 457–489. DOI: 10.1017/S1351324915000455.

[62] Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. "How to Ask Good Questions? Try to Leverage Paraphrases." In: *Proc. 58th Annu. Meeting Association for Computational Linguistics*. Virtual Event: Association for Computational Linguistics, 2020, pp. 6130–6140. DOI: 10.18653/v1/2020.acl-main.545.

[63] Yiping Jin and Phu Le. "Selecting Domain-Specific Concepts for Question Generation With Lightly-Supervised Methods." In: *Proc. 9th Int. Natural Language Generation Conf.* Edinburgh, United Kingdom: Association for Computational Linguistics, 2016, pp. 133–142. DOI: 10.18653/v1/W16-6623.

[64] Benny G. Johnson, Jeffrey S. Dittel, Rachel Van Campenhourt, Rodrigo Bistoflfi, Aida Maeda, and Bill Jerome. "Parallel Construction: A Parallel Corpus Approach for Automatic Question Generation in Non-English Languages." In: *Proc. 4th Int. Workshop Intelligent Textbooks*. Durham, United Kingdom, 2022.

[65] Karen S. Jones and Julia R. Galliers. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Heidelberg, Germany: Springer-Verlag, 1995. DOI: 10.1007/BFb0027470.

[66] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. "SpanBERT: Improving pre-training by representing and predicting spans." In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 64–77. DOI: 10.1162/tacl_a_003009.

[67] Sven Judel, Timo Bergerbusch, and Ulrik Schroeder. "Automatisierte Generierung von Automaten und automatenbasierten Aufgaben." In: *Proc. 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik - DELFI*. Virtual Event: Gesellschaft für Informatik e.V., 2020, pp. 133–144.

[68] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. "Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension." In: *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*. Honolulu, Hawaii, USA: IEEE, 2017, pp. 4999–5007. DOI: 10.1109/CVPR.2017.571.

[69] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed W. Zamir, Fahad S. Khan, and Mubarak Shah. "Transformers in vision: A survey." In: *ACM Computing Surveys* (2021), pp. 1–41. DOI: 10.1145/3505244.

[70] David E. Kieras. *Thematic processes in the comprehension of technical prose*. Technical Report No. (UARZ/DP/TR-82/ONR-10) Department of Psychology University of Arizona, Tucson, Arizona 85721. 1982.

[71] Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. "Improving Neural Question Generation Using Answer Separation." In: *Proc. 2019 AAAI Conf. Artificial Intelligence*. Honolulu, Hawaii, USA: AAAI Press, 2019, pp. 6602–6609. DOI: 10.1609/aaai.v33i01.33016602.

[72] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. "Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems." In: *Proc. 2019 CHI Conf. Human Factors in Computing Systems*. Glasgow, United Kingdom: Association for Computing Machinery, 2019, pp. 1–14. DOI: 10.1145/3290605.3300641.

[73] Kenneth R. Koedinger, Elizabeth A. McLaughlin, Julianna Zhuxin Jia, and Norman L. Bier. "Is the Doer Effect a Causal Relationship? How Can We Tell and Why It's Important." In: *Proc. 6th Int. Conf. Learning Analytics and Knowledge*. Edinburgh, United Kingdom: Association for Computing Machinery, 2016, pp. 388–397. DOI: 10.1145/2883851.2883957.

[74] Girish Kumar, Rafael E. Banchs, and Luis F. D'Haro. "RevUP: Automatic gap-fill question generation from educational texts." In: *Proc. 10th Workshop Innovative Use of NLP for Building Educational Applications*. Denver, Colorado, USA: Association for Computational Linguistics, 2015, pp. 154–161.

[75] Vishwajeet Kumar, Kireeti Boorla, Yogesh Meena, Ganesh Ramakrishnan, and Yuan-Fang Li. "Automating reading comprehension by generating question and answer pairs." In: *Proc. 2018 Pacific-Asia Conf. Knowledge Discovery and Data Mining*. Melbourne, Australia: Springer International Publishing, 2018, pp. 335–348. DOI: 10.1007/978-3-319-93040-4_27.

[76] Vishwajeet Kumar, Manish Joshi, Ganesh Ramakrishnan, and Yuan-Fang Li. "Vocabulary matters: A simple yet effective approach to paragraph-level question generation." In: *Proc. 1st Conf. Asia-Pacific Chapter Association for Computational Linguistics and the 10th Int. Joint Conf. Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 2020, pp. 781–785.

[77] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. "A systematic review of automatic question generation for educational purposes." In: *International Journal of Artificial Intelligence in Education* 30.1 (2020), pp. 121–204. DOI: 10.1007/s40593-019-00186-y.

[78] Tom Kwiatkowski et al. "Natural questions: a benchmark for question answering research." In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 453–466. DOI: 10.1162/tacl_a_00276.

[79] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. "RACE: Large-scale ReAding Comprehension Dataset From Examinations." In: *Proc. 2017 Conf. Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 785–794. DOI: 10.18653/v1/D17-1082.

[80] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." In: *Proc. 2020 Int. Conf. Learning Representations*. Virtual Event, 2020.

[81] Timo Lenzner, Lars Kaczmirek, and Alwine Lenzner. "Cognitive burden of survey questions and response times: A psycholinguistic experiment." In: *Applied Cognitive Psychology* 24.7 (2010), pp. 1003–1020. DOI: 10.1002/acp.1602.

[82] José A. León and Gala E. Peñalba. "Understanding causality and temporal sequence in scientific discourse." In: *The Psychology of Science Text Comprehension* (2002), pp. 155–178.

[83] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. "Gated Graph Sequence Neural Networks." In: *Proc. 2016 Int. Conf. Learning Representations*. San Juan, Puerto Rico, 2016.

[84] Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. "Asking questions the human way: Scalable question-answer generation from text corpus." In: *Proc. 2020 Web Conf.* Taipei, Taiwan: Association for Computing Machinery, 2020, pp. 2032–2043. DOI: 10.1145/3366423.3380270.

[85] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation." In: *Proc. 2016 Conf. Empirical Methods in Natural Language Processing*. Austin, Texas, USA: Association for Computational Linguistics, 2016, pp. 2122–2132. DOI: 10.18653/v1/D16-1230.

[86] Ming Liu, Rafael A. Calvo, and Vasile Rus. "Automatic question generation for literature review writing support." In: *Proc. 2010 Int. Conf. Intelligent Tutoring Systems*. Pittsburgh, Pensylvania, USA: Springer International Publishing, 2010, pp. 45–54. DOI: 10.1007/978-3-642-13388-6_9.

[87] Ming Liu, Rafael A. Calvo, and Vasile Rus. "G-Asks: An intelligent automatic question generation system for academic writing support." In: *Dialogue & Discourse* 3.2 (2012), pp. 101–124. DOI: 10.5087/dad.2012.205.

[88] Shusen Liu et al. "Visual exploration of semantic relationships in neural word embeddings." In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2017), pp. 553–562.

[89] Yinhan Liu et al. *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv.1907.11692, preprint. July 2019. DOI: 10.48550/ARXIV.1907.11692.

[90] Robert F. Lorch Jr. "Text-signaling devices and their effects on reading and memory processes." In: *Educational Psychology Review* 1.3 (1989), pp. 209–234. DOI: 10.1007/BF01320135.

[91] Robert F. Lorch Jr., Elizabeth P. Lorch, and Madeline A. Klusewitz. "Effects of typographical cues on reading and recall of text." In: *Contemporary Educational Psychology* 20.1 (1995), pp. 51–64. DOI: 10.1006/ceps.1995.1003.

[92] Robert F. Lorch Jr. and Paul Van den Broek. "Understanding reading comprehension: Current and future contributions of cognitive science." In: *Contemporary Educational Psychology* 22.2 (1997), pp. 213–246. DOI: 10.1006/ceps.1997.0931.

[93] Owen H. Lu, Anna Y. Huang, Danny C. Tsai, and Stephen J. Yang. "Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students Learning Performance." In: *Educational Technology & Society* 24.3 (2021), pp. 159–173.

[94] Sedigheh Mahdavi, Aijun An, Heidar Davoudi, Marjan Delpisheh, and Emad Gohari. "Question-Worthy Sentence Selection for Question Generation." In: *Proc. 2020 Canadian Conf. Artificial Intelligence*. Virtual Event: Springer International Publishing, 2020, pp. 388–400. DOI: 10.1007/978-3-030-47358-7_40.

[95] William C. Mann and Sandra A. Thompson. "Rhetorical Structure Theory: Toward a functional theory of text organization." In: *Text - Interdisciplinary Journal for the Study of Discourse* 8.3 (1988), pp. 243–281. DOI: 10.1515/text.1.1988.8.3.243.

[96] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. "Scoring, term weighting, and the vector space model." In: *Introduction to Information Retrieval*. Cambridge, United Kingdom: Cambridge University Press, 2008, pp. 100–123. DOI: 10.1017/CBO9780511809071.007.

[97] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit." In: *Proc. 52nd Annu. Meeting Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014, pp. 55–60. DOI: `10.3115/v1/P14-5010`.

[98] Patricia D. Mautone and Richard E. Mayer. "Signaling as a cognitive guide in multimedia learning." In: *Journal of Educational Psychology* 93.2 (2001), pp. 377–389. DOI: `10.1037/0022-0663.93.2.377`.

[99] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. "On Faithfulness and Factuality in Abstractive Summarization." In: *Proc. 58th Annu. Meeting Association for Computational Linguistics*. Virtual Event: Association for Computational Linguistics, 2020, pp. 1906–1919. DOI: `10.18653/v1/2020.acl-main.173`.

[100] Karen Mazidi and Rodney D. Nielsen. "Leveraging multiple views of text for automatic question generation." In: *Proc. 2015 Int. Conf. Artificial Intelligence in Education*. Madrid, Spain: Springer International Publishing, 2015, pp. 257–266. DOI: `10.1007/978-3-319-19773-9_26`.

[101] Karen Mazidi and Paul Tarau. "Infusing NLU into automatic question generation." In: *Proc. 9th Int. Natural Language Generation Conf.* Edinburgh, United Kingdom: Association for Computational Linguistics, 2016, pp. 51–60. DOI: `0.18653/v1/W16-6609`.

[102] Tom McCoy, Ellie Pavlick, and Tal Linzen. "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference." In: *Proc. 57th Annu. Meeting Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 3428–3448. DOI: `10.18653/v1/P19-1334`.

[103] Danielle S. McNamara and Joe Magliano. "Toward a comprehensive model of comprehension." In: *Psychology of Learning and Motivation* 51 (2009), pp. 297–384. DOI: `10.1016/S0079-7421(09)51009-2`.

[104] Bonnie J. Meyer. "Identification of the structure of prose and its implications for the study of reading and memory." In: *Journal of Literacy Research* 7.1 (1975), pp. 7–47. DOI: `10.1080/10862967509547120`.

[105] Timothee Mickus, Mathieu Constant, Denis Paperno, and Kees Van Deemter. "What do you mean, BERT? Assessing BERT as a Distributional Semantics Model." In: *Proceedings of the Society for Computation in Linguistics* 3 (2020), pp. 350–361. DOI: `10.7275/t778-ja71`.

[106] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." In: *Proc. 2013 Int. Conf. Learning Representations*. Scottsdale, Arizona, USA, 2013.

[107] Ines Montani et al. *explosion/spaCy: New Span Ruler component, JSON (de)serialization of Doc, span analyzer and more*. Version v3.3.1. 2022. DOI: `10.5281/zenodo.6621076`.

[108] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization." In: *Proc. 2018 Conf. Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 1797–1807. DOI: 10.18653/v1/D18-1206.

[109] Gonzalo Navarro. "A guided tour to approximate string matching." In: *ACM Computing Surveys* 33.1 (2001), pp. 31–88. DOI: 10.1145/375360.375365.

[110] Preksha Nema and Mitesh M. Khapra. "Towards a Better Metric for Evaluating Question Generation Systems." In: *Proc. 2018 Conf. Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3950–3959. DOI: 10.18653/v1/D18-1429.

[111] Ani Nenkova and Lucy Vanderwende. *The impact of frequency on summarization*. Technical Report No. MSR-TR-2005, Microsoft Research, Redmond, Washington. 2005.

[112] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. "Facebook FAIR's WMT19 News Translation Task Submission." In: *Proc. 4th Conf. Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 314–319. DOI: 10.18653/v1/W19-5333.

[113] Dat Q. Nguyen and Anh T. Vu Thanh andNguyen. "BERTweet: A pre-trained language model for English Tweets." In: *Proc. 2020 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*. Virtual Event: Association for Computational Linguistics, 2020, pp. 9–14. DOI: 10.18653/v1/2020.emnlp-demos.2.

[114] Timothy Niven and Hung-Yu Kao. "Probing neural network comprehension of natural language arguments." In: *Proc. 57th Annu. Meeting Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 4658–4664. DOI: 10.18653/v1/P19-1459.

[115] Erol Ozcelik, Ismahan Arslan-Ari, and Kursat Cagiltay. "Why does signaling enhance multimedia learning? Evidence from eye movements." In: *Computers in Human Behavior* 26.1 (2010), pp. 110–117. DOI: 10.1016/j.chb.2009.09.001.

[116] Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. *Recent advances in neural question generation*. arXiv:1905.08949, preprint. May 2019. DOI: 10.48550/arXiv.1905.08949.

[117] Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. "Semantic Graphs for Generating Deep Questions." In: *Proc. 58th Annu. Meeting Association for Computational Linguistics*. Virtual Event: Association for Computational Linguistics, 2020, pp. 1463–1475. DOI: 10.18653/v1/2020.acl-main.135.

[118] Zilong Pan, Chenglu Li, and Min Liu. "Learning Analytics Dashboard for Problem-Based Learning." In: *Proc. 7th ACM Conf. Learning @ Scale*. Virtual Event: Association for Computing Machinery, 2020, pp. 393–396. DOI: 10.1145/3386527.3406751.

[119] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In: *Proc. 40th Annu. meeting Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.

[120] Philip I. Pavlik Jr., Andrew M. Olney, Amanda Banker, Luke Eglington, and Jeffrey Yarbro. "The Mobile Fact and Concept Textbook System (MoFaCTS)." In: *Proc. 2nd Int. Workshop Intelligent Textbooks*. Virtual Event, 2020, pp. 35–59.

[121] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation." In: *Proc. 2014 Conf. Empirical Methods in Natural Language Processing*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[122] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc., 2012.

[123] Weizhen Qi et al. "ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training." In: *Proc. 2020 Findings Association for Computational Linguistics: EMNLP*. Virtual Event: Association for Computational Linguistics, 2020, pp. 2401–2410.

[124] Weizhen Qi et al. "ProphetNet-X: Large-Scale Pre-training Models for English, Chinese, Multi-lingual, Dialog, and Code Generation." In: *Proc. 59th Annu. Meeting Association for Computational Linguistics and the 11th Int. Joint Conf. Natural Language Processing: System Demonstrations*. Virtual Event: Association for Computational Linguistics, 2021, pp. 232–239. DOI: 10.18653/v1/2021.acl-demo.28.

[125] Fanyi Qu, Xin Jia, and Yunfang Wu. "Asking Questions Like Educational Experts: Automatically Generating Question-Answer Pairs on Real-World Examination Data." In: *Proc. 2021 Conf. Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 2583–2593.

[126] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. *Improving language understanding by generative pre-training*. Technical Report, OpenAI. 2018.

[127] Pranav Rajpurkar, Robin Jia, and Percy Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD." In: *Proc. 56th Annu. Meeting Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 784–789. DOI: 10.18653/v1/P18-2124.

[128] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: *Proc. 2016 Conf. Empirical Methods in Natural Language Processing*. Austin, Texas, USA: Association for Computational Linguistics, 2016, pp. 2383–2392. DOI: `10.18653/v1/D16-1264`.

[129] Ines Rehbein and Josef Ruppenhofer. "A new resource for German causal language." In: *Proc. 12th Language Resources and Evaluation Conf.* Marseille, France: European Language Resources Association, 2020, pp. 5968–5977.

[130] John P. Rickards. "Interaction of position and conceptual level of adjunct questions on immediate and delayed retention of text." In: *Journal of Educational Psychology* 68.2 (1976), pp. 210–217. DOI: `10.1037/0022-0663.68.2.210`.

[131] Henry L. Roediger III and Andrew C. Butler. "The critical role of retrieval practice in long-term retention." In: *Trends in Cognitive Sciences* 15.1 (2011), pp. 20–27. DOI: `10.1016/j.tics.2010.09.003`.

[132] Julian Roelle and Kirsten Berthold. "Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters." In: *Learning and Instruction* 49 (2017), pp. 142–156. DOI: `10.1016/j.learninstruc.2017.01.008`.

[133] Jean-François Rouet and Eduardo Vidal-Abarca. "Mining for meaning: Cognitive effects of inserted questions in learning from scientific text." In: *The Psychology of Science Text Comprehension* (2002), pp. 417–436.

[134] Sylvio Rüdian, Alexander Heuts, and Niels Pinkwart. "Educational Text Summarizer: Which sentences are worth asking for?" In: *Proc. 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik - DELFI*. Virtual Event: Gesellschaft für Informatik e.V., 2020, pp. 277–288.

[135] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv:1910.01108, preprint. Mar. 2019. DOI: `10.48550/arXiv.1910.01108`.

[136] Cicero dos Santos, Bing Xiang, and Bowen Zhou. "Classifying Relations by Ranking with Convolutional Neural Networks." In: *Proc. 53rd Annu. Meeting Association for Computational Linguistics and the 7th Int. Joint Conf. Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, 2015, pp. 626–634. DOI: `10.3115/v1/P15-1061`.

[137] Shlomo S. Sawilowsky. "New effect size rules of thumb." In: *Journal of Modern Applied Statistical Methods* 8.2 (2009), pp. 597–599. DOI: `10.22237/jmasm/1257035100`.

[138] Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. "The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task." In: *Proc. 21st Conf. Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 15–25. DOI: `10.18653/v1/K17-1004`.

[139] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units." In: *Proc. 54th Annu. Meeting Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162.

[140] Murray Singer and Nathalie Gagnon. "Detecting causal inconsistencies in scientific text." In: *Narrative Comprehension, Causality, and Coherence: Essays in Honor of Tom Trabasso* (1999), pp. 179–194.

[141] Megan A. Smith and Jeffrey D. Karpicke. "Retrieval practice with short-answer, multiple-choice, and hybrid tests." In: *Memory* 22.7 (2014), pp. 784–802. DOI: 10.1080/09658211.2013.831454.

[142] Sasha Spala, Nicholas A. Miller, Franck Dernoncourt, and Carl Dockhorn. "SemEval-2020 Task 6: Definition Extraction from Free Text with the DEFT Corpus." In: *Proc. 14th Workshop Semantic Evaluation*. Virtual Event: International Committee for Computational Linguistics, 2020, pp. 336–345. DOI: 10.18653/v1/2020.semeval-1.41.

[143] Sasha Spala, Nicholas A. Miller, Yiming Yang, Franck Dernoncourt, and Carl Dockhorn. "DEFT: A corpus for definition extraction in free- and semi-structured text." In: *Proc. 13th Linguistic Annotation Workshop*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 124–131. DOI: 10.18653/v1/W19-4015.

[144] Karen Sparck Jones. "A Statistical Interpretation of Term Specificity and its Application in Retrieval." In: *Journal of Documentation* 28.1 (1972), pp. 11–21. DOI: 10.1108/eb026526.

[145] Nancy J. Spencer. "Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability." In: *Journal of Psycholinguistic Research* 2.2 (1973), pp. 83–98. DOI: 10.1007/BF01067203.

[146] Daniel Stafford and Robert Flatley. "Openstax." In: *The Charleston Advisor* 20.1 (2018), pp. 48–51. DOI: doi:10.5260/chara.20.1.48.

[147] Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A. Hearst. "Automatically Generating Cause-and-Effect Questions from Passages." In: *Proc. 16th Workshop Innovative Use of NLP for Building Educational Applications*. Virtual Event: Association for Computational Linguistics, 2021, pp. 158–170.

[148] Ralf Steinmetz and Klara Nahrstedt. *Multimedia Applications*. 1st ed. Heidelberg, Germany: Springer-Verlag, 2004. DOI: 10.1007/978-3-662-08876-0.

[149] Tim Steuer, Leonard Bongard, Jan Uhlig, and Gianluca Zimmer. "On the Linguistic and Pedagogical Quality of Automatic Question Generation via Neural Machine Translation." In: *Proc. 2021 European Conf. Technology Enhanced Learning*. Virtual Event: Springer International Publishing, 2021, pp. 289–294. DOI: 10.1007/978-3-030-86436-1_22.

[150] Tim Steuer, Anna Filighera, Tobias Meuser, and Christoph Rensing. *I Do Not Understand What I Cannot Define: Automatic Question Generation With Pedagogically-Driven Content Selection*. arXiv:2110.04123v1, preprint. Under review from IEEE Transactions on Learning Technologies since 02/2021. Oct. 2021. DOI: `10.48550/arXiv.2110.04123`.

[151] Tim Steuer, Anna Filighera, Nina Mouhammad, Gianluca Zimmer, and Thomas Tregel. "Learning-Relevant Concept Extraction By Utilizing Automatically Generated Textbook Corpora." In: *Proc. 2022 Int. Conf. Advanced Learning Technologies*. Bucharest, Romania: IEEE Computer Society, 2022, pp. 379–383. DOI: `10.1109/ICALT55010.2022.00117`.

[152] Tim Steuer, Anna Filighera, and Christoph Rensing. "Exploring Artificial Jabbering for Automatic Text Comprehension Question Generation." In: *Proc. 2020 European Conf. Technology Enhanced Learning*. Virtual Event: Springer International Publishing, 2020, pp. 1–14. DOI: `10.1007/978-3-030-57717-9_1`.

[153] Tim Steuer, Anna Filighera, and Christoph Rensing. "Remember the Facts? Investigating Answer-Aware Neural Question Generation for Text Comprehension." In: *Proc. 2020 Int. Conf. Artificial Intelligence in Education*. Virtual Event: Springer International Publishing, 2020, pp. 512–523. DOI: `10.1007/978-3-030-52237-7_41`.

[154] Tim Steuer, Anna Filighera, and Thomas Tregel. "Investigating Educational and Noneducational Answer Selection for Educational Question Generation." In: *IEEE Access* 10 (2022), pp. 63522–63531. DOI: `10.1109/ACCESS.2022.3180838`.

[155] Tim Steuer, Anna Filighera, Thomas Tregel, and André Miede. "Educational Automatic Question Generation Improves Reading Comprehension in Non-Native Speakers: A Learner-Centric Case Study." In: *Frontiers in Artificial Intelligence* 5 (2022), pp. 1–14. DOI: `10.3389/frai.2022.900304`.

[156] Tim Steuer, Anna Filighera, Gianluca Zimmer, and Thomas Tregel. "What Is Relevant for Learning? Approximating Readers' Intuition Using Neural Content Selection." In: *Proc. 2022 Int. Conf. Artificial Intelligence in Education*. Durham, United Kingdom: Springer International Publishing, 2022, pp. 505–511. DOI: `10.1007/978-3-031-11644-5_41`.

[157] Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. "Multihop Question Generation with Graph Convolutional Network." In: *Proc. 2020 Findings Association for Computational Linguistics: EMNLP*. Virtual Event: Association for Computational Linguistics, 2020, pp. 4636–4647. DOI: `10.18653/v1/2020.findings-emnlp.416`.

[158] Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. "Neural Models for Key Phrase Extraction and Question Generation." In: *Proc.2018 Workshop Machine Reading for Question Answering*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 78–88. DOI: `10.18653/v1/W18-2609`.

[159] Yibo Sun et al. “Joint learning of question answering and question generation.” In: *IEEE Transactions on Knowledge and Data Engineering* 32.5 (2019), pp. 971–982. DOI: 10.1109/TKDE.2019.2897773.

[160] Yuni Susanti, Takenobu Tokunaga, and Hitoshi Nishikawa. “Integrating automatic question generation with computerised adaptive test.” In: *Research and Practice in Technology Enhanced Learning* 15 (2020), pp. 1–22. DOI: 10.1186/s41039-020-00132-w.

[161] Rohail Syed et al. “Improving learning outcomes with gaze tracking and automatic question generation.” In: *Proc. 2020 Web Conf.* Taipei, Taiwan: Association for Computing Machinery, 2020, pp. 1693–1703. DOI: 10.1145/3366423.3380240.

[162] Lasang J. Tamang, Rabin Banjade, Jeevan Chapagain, and Vasile Rus. “Automatic Question Generation for Scaffolding Self-explanations for Code Comprehension.” In: *Proc. 2022 Int. Conf. Artificial Intelligence in Education*. Durham, United Kingdom: Springer International Publishing, 2022, pp. 743–748. DOI: 10.1007/978-3-031-11644-5_77.

[163] Rachel Van Campenhout, Nick Brown, Bill Jerome, Jeffrey S. Dittel, and Benny G. Johnson. “Toward Effective Courseware at Scale: Investigating Automatically Generated Questions as Formative Practice.” In: *Proc. 8th ACM Conf. Learning @ Scale*. Virtual Event: Association for Computing Machinery, 2021, pp. 295–298. DOI: 10.1145/3430895.3460162.

[164] Paul Van den Broek, Sandra Virtue, Michelle G. Everson, Yuhtsuen Tzeng, and Yung-chi Sung. “Comprehension and memory of science texts: Inferential processes and the construction of a mental representation.” In: *The Psychology of Science Text Comprehension* (2002), pp. 131–154.

[165] Chris Van der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. “Best practices for the human evaluation of automatically generated text.” In: *Proc. 12th Int. Conf. Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, 2019, pp. 355–368. DOI: 10.18653/v1/W19-8643.

[166] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[167] Stalin Varanasi, Saadullah Amin, and Guenter Neumann. “CopyBERT: A Unified Approach to Question Generation with Self-Attention.” In: *Proc. 2nd Workshop Natural Language Processing for Conversational AI*. Virtual Event: Association for Computational Linguistics, 2020, pp. 25–31. DOI: 10.18653/v1/2020.nlp4convai-1.3.

[168] Ashish Vaswani et al. “Attention is All you Need.” In: *Proc. 30st Int. Conf. Neural Information Processing*. Long Beach, California, USA: Curran Associates, Inc., 2017, pp. 6000–6010. DOI: 10.5555/3295222.329534.

[169] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. "Pointer Networks." In: *Proc. 28th Int. Conf. Neural Information Processing Systems*. Montréal, Canada: Curran Associates, Inc., 2015, pp. 2692–2700. DOI: `10.5555/2969442.2969540`.

[170] Basil S. Walker. "Effects of inserted questions on retroactive inhibition in meaningful verbal learning." In: *Journal of Educational Psychology* 66.4 (1974), pp. 486–490. DOI: `10.1037/h0036749`.

[171] Bingning Wang, Xiaochuan Wang, Ting Tao, Qi Zhang, and Jingfang Xu. "Neural question generation with answer pivot." In: *Proc. 2020 AAAI Conf. Artificial Intelligence*. New York, New York, USA: AAAI Press, 2020, pp. 9138–9145. DOI: `10.1609/aaai.v34i05.6449`.

[172] Xinyu Wang et al. "Automated Concatenation of Embeddings for Structured Prediction." In: *Proc. 59th Annu. Meeting Association for Computational Linguistics and the 11th Int. Joint Conf. Natural Language Processing (Volume 1: Long Papers)*. Virtual Event: Association for Computational Linguistics, 2021, pp. 2643–2660. DOI: `10.18653/v1/2021.acl-long.206`.

[173] Zhen Wang, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang. "Diversify question generation with continuous content selectors and question type modeling." In: *Proc. 2020 Findings Association for Computational Linguistics: EMNLP*. Virtual Event: Association for Computational Linguistics, 2020, pp. 2134–2143. DOI: `10.18653/v1/2020.findings-emnlp.194`.

[174] Zichao Wang, Andrew S. Lan, Weili Nie, Andrew E. Waters, Phillip J. Grimaldi, and Richard G. Baraniuk. "QG-Net: A Data-Driven Question Generation Model for Educational Content." In: *Proc. 4th Annu. ACM Conf. Learning at Scale*. London, United Kingdom: Association for Computing Machinery, 2018, pp. 1–10. DOI: `10.1145/3231644.3231654`.

[175] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. "How to Use t-SNE Effectively." In: *Distill* (2016). DOI: `10.23915/distill.00002`.

[176] Stefan Wellek. *Testing statistical hypotheses of equivalence and noninferiority*. 2nd ed. CRC Press, Taylor and Francis Group, 2010. DOI: `10.1201/EBK1439808184`.

[177] Angelica Willis, Glenn Davis, Sherry Ruan, Lakshmi Manoharan, James Landay, and Emma Brunskill. "Key phrase extraction for generating educational question-answer pairs." In: *Proc. 6th ACM Conf. Learning @ Scale*. Chicago, Illinois, USA: Association for Computing Machinery, 2019, pp. 1–10. DOI: `10.1145/3330430.3333636`.

[178] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing." In: *Proc. 2020 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*. Virtual Event: Association for Computational Linguistics, 2020, pp. 38–45. DOI: `10.18653/v1/2020.emnlp-demos.6`.

[179] Yonghui Wu et al. *Google's neural machine translation system: Bridging the gap between human and machine translation*. arXiv:1609.08144, preprint. Oct. 2016. DOI: `10.48550/arXiv.1609.08144`.

[180] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. "Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths." In: *Proc. 2015 Conf. Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1785–1794. DOI: 10.18653/v1/D15-1206.

[181] Bei Yu, Yingya Li, and Jun Wang. "Detecting Causal Language Use in Science Findings." In: *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 4664–4674. DOI: 10.18653/v1/D19-1473.

[182] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. "Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate?" In: *BMC Medical Research Methodology* 16 (2016), pp. 1–10. DOI: 10.1186/s12874-016-0200-9.

[183] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. "PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization." In: *Proc. 37th Int. Conf. Machine Learning*. Virtual Event: PMLR, 2020, pp. 11328–11339.

[184] Lishan Zhang and Kurt VanLehn. "How do machine-generated questions compare to human-generated questions?" In: *Research and Practice in Technology Enhanced Learning* 11.1 (2016), pp. 1–28. DOI: 10.1186/s41039-016-0031-7.

[185] Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. "Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks." In: *Proc. 2018 Conf. Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3901–3910. DOI: 10.18653/v1/D18-1424.

[186] Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. "Neural Question Generation from Text: A Preliminary Study." In: *Proc. 2018 Natural Language Processing and Chinese Computing*. Dalian, China: Springer International Publishing, 2018, pp. 662–671. DOI: 10.1007/978-3-319-73618-1_56.

[187] Wenjie Zhou, Minghua Zhang, and Yunfang Wu. "Question-type Driven Question Generation." In: *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 6032–6037. DOI: 10.18653/v1/D19-1622.

[188] Yukun Zhu et al. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." In: *Proc. 2015 IEEE Int. Conf. Computer Vision*. Boston, Massachusetts, USA, 2015, pp. 19–27. DOI: 10.1109/ICCV.2015.11.

All web pages cited in this work have been checked in August 2022. However, due to the dynamic nature of the World Wide Web, their long-term availability cannot be guaranteed.

A

APPENDIX

A.1 RACE PLOTS OF THE PILOT STUDY

The following plots show the Wh-word and BLEU-4 distribution on the RACE corpus for the study conducted in Chapter 4. The plots show very similar results to the corresponding LearningQ plots.

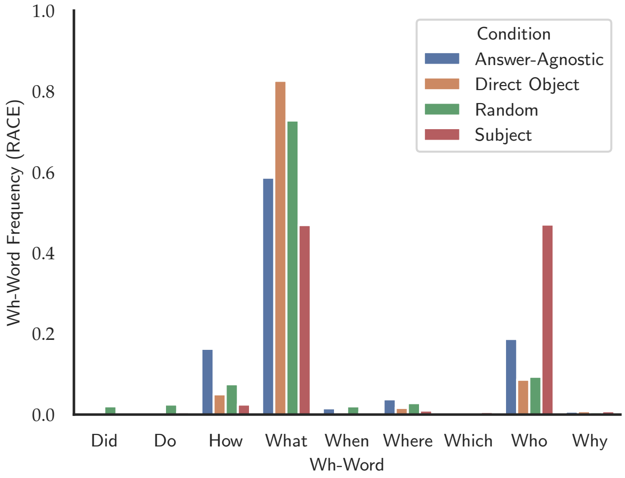


Figure 28: Distribution of Wh-Words on RACE.

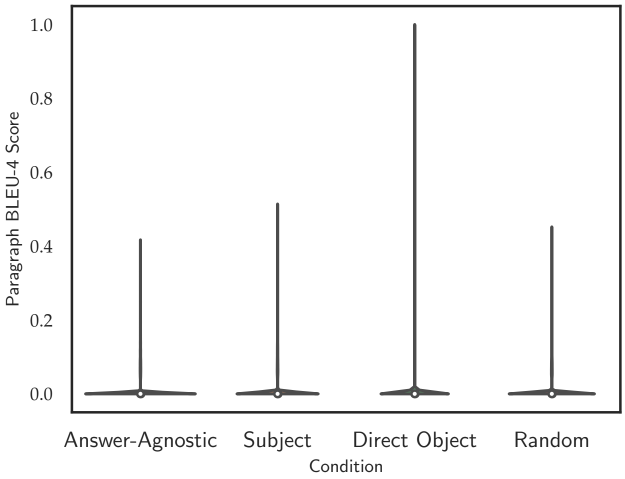


Figure 29: Violin plot of the distribution of the mean paragraph BLEU-4 scores on RACE.

A.2 CAUSE AND DEFINITION ANNOTATION SCHEMA

The following sections are direct excerpts of the annotation scheme used to annotate the definitions and causal sentences, primarily authored in collaboration with Friederike Lenke. They are provided to give readers a better understanding of the detail level and typical usage of the corresponding annotation scheme. The excerpts are given in German because the study was conducted with German texts and annotators. Note that this does not describe the complete annotation scheme but only representative sections.

A.2.1 *Was ist ein Tag / Was ist eine Definition*

Ein Tag ist eine benannte Markierung im Text. Durch den Namen kann man verschiedene Markierungen auseinanderhalten.

> Beispiel: "**Ein Hund** ist *ein vierbeiniges Säugetier*. "

Wir haben hier zwei Tags, die den zu definierenden Term (fettgedruckt) und die Definition (kursiv) auseinanderhalten. Eine Relation verbindet zwei Markierungen im Text. Dadurch können zwei Markierungen in Bezug zueinander gestellt werden.

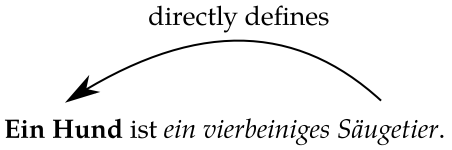


Abbildung 30: Ein Beispiel für eine Relation.

Im Beispiel wird über die Relation (directly defines) klargestellt, dass die Definition den Term definiert.

A.2.2 *Mindset beim Annotieren*

- Welchen Zweck verfolgen wir? Automatisierte Generierung von Fragen, deren Beantwortung textuell vermittelte Themenkomplexe didaktisch festigen sollen.
- Annotiere ich auch Sätze, die zwar laut Annotationsschema markiert gehören, die aber fürs Lernen nicht sinnvoll sind? Ja, alle Sätze, bei denen ein Annotationsschema trifft, werden annotiert, losgelöst, ob diese wichtig fürs Lernen sind. Für die spätere automatische Erkennung ist es von höchster Wichtigkeit, dass wir alle Sätze nach unserem Regelsatz auswählen und nicht weitere implizite Annahmen einfließen lassen.
- Wie und wie weitreichend sollen die Annotation vorgenommen werden? So viel Kontext, wie zum Verständnis der einzelnen Annotation nötig ist, nicht mehr! Eine Annotation sollte in dem minimalen Umfang getätigt werden, der für ihre

präzise Nachvollziehbarkeit und eindeutige Zuweisung durch Andere verständlich scheint, jedoch nicht so weit gefasst werden, dass Andere womöglich andere Annotationen vornehmen würden.

Dem Annotieren liegt immer ein subjektiver Entscheidungsprozess zugrunde, der oft keine komplette Übereinstimmung mit Anderen zulässt! Wir versuchen dennoch, unsere subjektiven Einschätzungen durch klare Annotationsschemata so weit wie möglich aneinander anzupassen. Hierbei möchten wir uns nicht (nur) auf "linguistische Marker"festlegen, da diese für unseren Anspruch viel zu eng gefasst sind und mögliche, komplexere Anwendungsfälle von vornherein ausschließen könnten; wir arbeiten mit abstrakteren Konzeptbegriffen, die bei aller Komplexität dennoch praktisch umsetzbar und für andere nachvollziehbar bleiben müssen.

A.2.3 *Annotationsschema Definitionen*

Eine explizite Definition in unserem Sinn erläutert einen Begriff (ein oder mehrere Wörter), indem sie eindeutig auf diesen referiert. Sie besteht zumindest aus einem Term und einer Definition sowie der Beziehung zwischen diesen (Direct-defines). Besitzt ein Satz nur einen Term ohne Definition oder eine Definition ohne Definitionsmarker, so annotieren wir ihn nicht.

Ein Term ist ein griffiger, später wiederverwendbarer Term. So kann er mit einem Glossarbegriff verglichen werden. Der Term wird immer mit seinem Artikel annotiert. Eine Definition ist eindeutig und definiert den Gegenstand im Sinne des Autors vollständig.

Tags:

- Term (der zu definierende primäre Begriff)
- Definition (definiert eindeutig einen Term. Die Definition ist integral für Term, aber nicht notwendigerweise eine 100% Erklärung)
- Alias Term [Optional] (ein Ersatzbegriff für den primären Begriff, Synonym)
- Definition Marker (Token, das den Satzteil als eine Definition ausweist (z.B. ist, nennt)
- Definition_Continued (Weiterführung einer aufgeteilten Definition)
- Term_Continued (Weiterführung eines aufgeteilten Terms)

Relationen:

- Direct-defines (Verbindung Term -> Definition)
- AKA (Verbindung Term -> Alias Term)
- Specifies (Verbindung Definition Marker ->Term)

- Continues Definition (Verbindung Definition -> Continued_Definition)
- Continues Term (Term -> Continued_Term)

Beispiel (keine Def.):

> *"Eine Universität beinhaltet Studierende und Professoren."*

Dies beschreibt zwar Eigenschaften einer Uni, jedoch fehlen wichtige Aspekte (z.B. Forschung). Da wir im Zweifel nicht alle wichtigen Aspekte kennen, schauen wir, ob im umliegenden Text noch wichtige Aspekte genannt werden, die aber nicht im Definitionskandidatensatz sind. Wir sind uns bewusst, dass dies zwangsläufig zu teilweise subjektiven Entscheidungen führt, versuchen aber bei jedem Satz bestmöglich die Intention des Autors zu treffen. Koreferenzen (Rückverweise) können auch als Term oder Definition dienen. Sofern mehr als ein Satz zwischen dem ursprünglichen Term und der Koreferenz liegen, wird das referenzierende Wort annotiert.

A.2.4 *Annotationsschema Kausalsätze*

Eine explizite Kausalität in unserem Sinn beschreibt ein Ursachen-Wirkungs-Verhältnis und wird über ein Triggerwort eingeleitet. Wir annotieren Sätze, die mindestens einen Cause und Effect beinhalten und durch ein Triggerwort gekennzeichnet sind. Ausnahme: Sätze, die ein Triggerwort enthalten, aber keine Kausalrelation beschreiben: diese bekommen das Triggerwort explizit als CAUSE_NO_TRIGGER markiert.

Tags:

- Cause (ein[e] Wirkkraft, Prozess, Ereignis, Handlung, etc., die einen Effekt herbeiführt)
- Effect (das Resultat, Ergebnis, die Folge, etc., die auf den Cause zurückzuführen sind)
- CAUSE_TRIGGER (der Ausdruck, der den Ausschlag für den Cause-Effect Zusammenhang liefert)
- CAUSE_NO_TRIGGER [Ausnahmefall] (der Ausdruck, der normalerweise einen Cause-Effect Zusammenhang liefert, macht das in diesem Fall nicht)
- Cause_Continued (Weiterführung eines aufgeteilten Cause)
- Effect_Continued (Weiterführung eines aufgeteilten Effect)

Relationen:

- Consequence (Verbindung zwischen Cause und Effect; kann naturgesetzlich oder intentional durch einen Actor hervorgerufen sein)
- Causes (Verbindung Causal Trigger -> Cause)

- Continues Cause (Verbindung Cause -> Cause_Continued)
- Continues Effect (Verbindung Effect -> Continued_Effect)

Die Kausalität muss sich ausschließlich aus der vorhandenen Textebene ableiten lassen. Sie steckt immer genau in einem Satz und geht nicht über Satzgrenzen hinweg. Eine Kausalitätsstruktur muss einen CAUSE_TRIGGER enthalten. Falls eine Relation mehrere Trigger enthält, werden alle als solche annotiert. Trigger, die angezeigt werden, aber nicht als solche angewendet werden, werden als CAUSE_NO_TRIGGER annotiert.

A.3 OVERVIEW OF ALL MODEL HYPERPARAMETERS

The following table provides an overview of the training procedure of the transformer models used in the different chapters of the thesis. All models have been trained with the HuggingFace library in Python on NVIDIA RTX cards and the Adam optimizer with a learning rate of 0.00002.

	Snapshot	Parameters	BS	Epochs	Remarks
Chapter 5					
1	bert-base-german-cased	108 million	16	5	2 models
Chapter 6					
1	albert-base-v2	11 million	4,8,16,32	6	
2	distilbert-base-cased	65 million	4,8,16,32	6	
3	bert-base-cased	107 million	4,8,16,32	6	
4	deberta-base	138 million	4,8,16,32	6	
5	roberta-base	124 million	4,8,16,32	6	
6	spanbert-base-cased	107 million	4,8,16,32	6	
Chapter 7					
1	bert-base-german-cased	108 million	4,8,16,32,64	6	2 models
2	bert-base-cased	107 million	4,8,16,32,64	6	2 models
Chapter 8					
1	distilbert-base-cased	65 million	16	6	

Table 21: An overview of the different models trained in this thesis. The column *BS* indicates the respective batch sizes while training.

A.4 LIST OF ACRONYMS

AQG	Automatic Question Generation
BLEU	Bilingual Evaluation Understudy
CEFR	Common European Framework Of Reference For Languages
DEFT	Definition Extraction From Texts
GRU	Gated Recurrent Unit
IAA	Inter-Annotator Agreement
IOB	Inside–Outside–Beginning
LSTM	Long Short-Term Memory
NER	Named Entity Recognition
NLG	Natural Language Generation
NLU	Natural Language Understanding
NQ	Natural Questions
PDF	Portable Document Format
RACE	Reading Comprehension Dataset From Examinations
ROC	Receiver Operating Characteristic
SQuAD	Stanford Question Answering Dataset
TQA	Textbook Question Answering
TQA-A	Textbook Question Answering With Answer Spans

A.5 SUPERVISED STUDENT THESES

[1] Viraj Kulkarni. “Automatically Generating Questions for Self Assessment Of Reading Comprehension.” Master Thesis. TU Darmstadt, 2019.

[2] Nina Kolmar. “Text Simplification for the German Language via Neural Networks.” Bachelor Thesis. TU Darmstadt, 2020.

[3] Tim Unverzagt. “Application of the Lottery Ticket Hypothesis in NLP and Early Pruning.” Bachelor Thesis. TU Darmstadt, 2020.

[4] Tianyang Zhou. “Detecting Question-Worthy Sentences with Extractive Summarization Methods.” Master Thesis. TU Darmstadt, 2020.

[5] Frederik Röper. “Design und Implementierung einer Evaluationsplattform für Fragegeneratoren im Bildungsbereich.” Bachelor Thesis. TU Darmstadt, 2020.

[6] Salma Loussaief. “Generation of Multi-Language Domain-Specific Vocabulary from Unstructured Data.” Master Thesis. TU Darmstadt, 2020.

[7] Lin Li. “Investigating Automatic Answer Extraction Methods for Automatic Question Generation on the SQuAD Dataset and in Education.” Master Thesis. TU Darmstadt, 2021.

[8] Aron Kaufmann. “An Empirical Comparison of German Neural Language Models through NLU Tasks.” Master Thesis. TU Darmstadt, 2021.

[9] Gianluca Zimmer. “Investigating Cause and Effect Extraction as Content Selection Method for Automatic Question Generation.” Bachelor Thesis. TU Darmstadt, 2021.

[10] Yassine Aziani. “Automatic Detection of Information Inconsistencies in Automatic Summaries.” Master Thesis. TU Darmstadt, 2021.

[11] Jan-Christoph Cramer. “Morphemepiece in Python: A Case Study of the Effects of a Morpheme Tokenizer on Language Model Performance.” Bachelor Thesis. TU Darmstadt, 2022.

B

AUTHOR'S PUBLICATIONS

JOURNAL PUBLICATIONS

[1] Tim Steuer, Anna Filighera, and Thomas Tregel. "Investigating Educational and Noneducational Answer Selection for Educational Question Generation." In: *IEEE Access* 10 (2022), pp. 63522–63531. DOI: `10.1109/ACCESS.2022.3180838`.

[2] Tim Steuer, Anna Filighera, Thomas Tregel, and André Miede. "Educational Automatic Question Generation Improves Reading Comprehension in Non-Native Speakers: A Learner-Centric Case Study." In: *Frontiers in Artificial Intelligence* 5 (2022), pp. 1–14. DOI: `10.3389/frai.2022.900304`.

CONFERENCE PUBLICATIONS

[3] Tim Steuer, Anna Filighera, Gianluca Zimmer, and Thomas Tregel. "What Is Relevant for Learning? Approximating Readers' Intuition Using Neural Content Selection." In: *Proc. 2022 Int. Conf. Artificial Intelligence in Education*. Durham, United Kingdom: Springer International Publishing, 2022, pp. 505–511. DOI: `10.1007/978-3-031-11644-5_41`.

[4] Tim Steuer, Anna Filighera, Nina Mouhammad, Gianluca Zimmer, and Thomas Tregel. "Learning-Relevant Concept Extraction By Utilizing Automatically Generated Textbook Corpora." In: *Proc. 2022 Int. Conf. Advanced Learning Technologies*. Bucharest, Romania: IEEE Computer Society, 2022, pp. 379–383. DOI: `10.1109/ICALT55010.2022.00117`.

[5] Tim Steuer, Leonard Bongard, Jan Uhlig, and Gianluca Zimmer. "On the Linguistic and Pedagogical Quality of Automatic Question Generation via Neural Machine Translation." In: *Proc. 2021 European Conf. Technology Enhanced Learning*. Virtual Event: Springer International Publishing, 2021, pp. 289–294. DOI: `10.1007/978-3-030-86436-1_22`.

[6] Tim Steuer, Anna Filighera, and Christoph Rensing. "Exploring Artificial Jabbering for Automatic Text Comprehension Question Generation." In: *Proc. 2020 European Conf. Technology Enhanced Learning*. Virtual Event: Springer International Publishing, 2020, pp. 1–14. DOI: `10.1007/978-3-030-57717-9_1`.

[7] Tim Steuer, Anna Filighera, and Christoph Rensing. "Remember the Facts? Investigating Answer-Aware Neural Question Generation for Text Comprehension." In: *Proc. 2020 Int. Conf. Artificial Intelligence in Education*. Virtual Event: Springer International Publishing, 2020, pp. 512–523. DOI: `10.1007/978-3-030-52237-7_41`.

[8] Tim Steuer and Christoph Rensing. "Themenübergreifende Diskursklassifikation auf Basis von Word Embeddings und Sequenzfeatures." In: *Proc. 17. Fachtagung Bildungstechnologien der Gesellschaft für Informatik - DELFI*. Berlin, Germany: Gesellschaft für Informatik e.V., 2019, pp. 45–56. DOI: 10.18420/delfi2019_234.

TECHNICAL REPORTS

[9] Tim Steuer, Anna Filighera, Tobias Meuser, and Christoph Rensing. *I Do Not Understand What I Cannot Define: Automatic Question Generation With Pedagogically-Driven Content Selection*. arXiv:2110.04123v1, preprint. Under review from IEEE Transactions on Learning Technologies since 02/2021. Oct. 2021. DOI: 10.48550/arXiv.2110.04123.

CO-AUTHORED PUBLICATIONS

[1] Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. "Your Answer is Incorrect... Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset." In: *Proc. 60th Annu. Meeting Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 8577–8591. DOI: 10.18653/v1/2022.acl-long.587.

[2] Pegah Golchin, Ralf Kundel, Tim Steuer, Rhaban Hark, and Ralf Steinmetz. "Improving DDoS Attack Detection Leveraging a Multi-aspect Ensemble Feature Selection." In: *Proc. 2022 IEEE/IFIP Network Operations and Management Symposium*. Budapest, Hungary: IEEE, 2022, pp. 1–5. DOI: 10.1109/NOMS54207.2022.9789763.

[3] Anna Filighera, Joel Tschesche, Tim Steuer, Thomas Tregel, and Lisa Wernet. "Towards Generating Counterfactual Examples as Automatic Short Answer Feedback." In: *Proc. 2022 Int. Conf. Artificial Intelligence in Education*. Durham, United Kingdom: Springer International Publishing, 2022, pp. 206–217. DOI: 10.1007/978-3-031-11644-5_17.

[4] Anna Filighera, Leonard Bongard, Tim Steuer, and Thomas Tregel. "Towards A Vocalization Feedback Pipeline for Language Learners." In: *Proc. 2022 Int. Conf. Advanced Learning Technologies*. Bucharest, Romania: IEEE Computer Society, 2022, pp. 248–252. DOI: 10.1109/ICALT55010.2022.00081.

[5] Anna Filighera, Tim Steuer, and Christoph Rensing. "Fooling It - Student Attacks on Automatic Short Answer Grading." In: *Proc. 2020 European Conf. Technology Enhanced Learning*. Virtual Event: Springer International Publishing, 2020, pp. 347–352. DOI: 10.1007/978-3-030-57717-9_25.

[6] Anna Filighera, Tim Steuer, and Christoph Rensing. "Fooling Automatic Short Answer Grading Systems." In: *Proc. 2020 Int. Conf. Artificial Intelligence in Education*. Virtual Event: Springer International Publishing, 2020, pp. 177–190. DOI: 10.1007/978-3-030-52237-7_15.

[7] Anna Filighera, Tim Steuer, and Christoph Rensing. "Automatic Text Difficulty Estimation Using Embeddings and Neural Networks." In: *Proc. 2019 European Conf. on Technology Enhanced Learning*. Delft, Netherlands: Springer International Publishing, 2019, pp. 335–348. DOI: 10.1007/978-3-030-29736-7_25.

[8] Oliver Schneider, Hendrik Drachsler, Christoph Rensing, Steuer Tim, and Jan Hansen. "Trusted Learning Analytics." In: *Proc. 17. Fachtagung Bildungstechnologien der Gesellschaft für Informatik - DELFI Workshops*. Berlin, Germany: Gesellschaft für Informatik e.V., 2019, p. 51. DOI: 10.18420/delfi2019-ws-106.

[9] Carsten Fuß, Tim Steuer, Kevin Noll, and André Miede. "Teaching the Achiever, Explorer, Socializer, and Killer – Gamification in University Education." In: *Games for Training, Education, Health and Sports*. Darmstadt, Germany: Springer International Publishing, 2014, pp. 92–99. DOI: 10.1007/978-3-319-05972-3_11.

C

ERKLÄRUNGEN LAUT PROMOTIONSORDNUNG

§8 Abs. 1 lit. c PromO
Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§8 Abs. 1 lit. d PromO
Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§9 Abs. 1 PromO
Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§9 Abs. 2 PromO
Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient

Darmstadt, 19. September 2022

Tim Steuer

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

`https://bitbucket.org/amiede/classicthesis/`

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

`http://postcards.miede.de/`

Final Version as of February 1, 2023 (`classicthesis`).