Towards Adaptive Language Learning Acknowledging Linguistic Variability and Curricular Constraints

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M.A. Tanja Schmidt, geb. Heck

aus Balingen

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	07.05.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Detmar Meurers
2. Berichterstatter:	Prof. Dr. Torsten Zesch

Acknowledgments

While working on my doctoral project, I received varied support that modulated the difficulty of the task enough to allow me to successfully complete it. I would like to thank everyone who contributed to this network of support, both at work and in my personal life.

To my supervisors Detmar Meurers and Thorsten Zesch, I am grateful for the freedom to develop my research in the direction of my own interests, but also for the feedback that pushed me towards ever improved versions of my work and provided new directions to look into.

To my colleagues, most of all Leona Colling and Stephen Bodnar, who were there throughout the entire duration of the PhD project, but also my more recent colleagues Walid El Hefney, Mariia Soliar, and Kristian Lange, for the great collaboration in the RCT projects as well as for the personal exchanges and emotional support. My thanks also include my colleagues from other institutions who provided the data from previous studies with the FeedBook or helped make the current RCTs a success and were always willing to provide input for data analyses. In extension, I also thank my colleagues who were not part of the RCT projects, but created an agreeable work environment and were willing to help out whenever asked.

To my proof-reader Melanie Heck who did not spare with comments, corrections and suggestions for spelling, grammar and logical flow issues she came across while reading the entire dissertation.

Finally, thank you to all those loved ones who had my back and treated me with so much patience and understanding during these challenging years.

Abstract

Learning foreign languages constitutes a central aspect of modern education, yet teachers lack the resources for catering to each student's individual needs. Intelligent tutoring systems can, in principal, fill this gap. They provide individualized practice for a subject matter in a digital environment by adapting learning material organized in a domain model to learner characteristics tracked in the system's learner model according to an instructional model. Their adaptation strategies encompass macro-adaptivity to select the best suitable practice material as well as micro-adaptivity to provide support on the selected material. Yet, most intelligent language tutoring systems neglect the needs of instructional settings. What is more, their strong focus on matching exercise difficulty to learner abilities allows little room for additional adaptation. Yet linguistic variability constitutes an important aspect in language learning as it is omnipresent in languages. This thesis therefore investigates how and to what extent linguistic variability should and can be considered in adaptive language tutoring systems accompanying classroom teaching that focus on grammar practice in contextualized exercises.

We first investigated the effect of linguistically varied exercises on learning. We found that varied practice fosters variability in learner productions, but that it also affects practice difficulty. In addition, it seems to strengthen declarative knowledge but slow down proceduralization. Learners would thus benefit from macro-adaptive exercise selection that controls linguistic variability. Since different linguistic variants have different effects on exercise difficulty, macro-adaptivity must split its focus between exercise difficulty and linguistic variability.

We therefore explored meso-adaptivity as an alternative means to mitigate exercise difficulty. We found that integrating this hybrid between macro- and micro-adaptivity to step by step reduce an exercise's difficulty when a learner struggles decreases the number of incorrect answers. Yet positive effects on learning gains manifested only when learners received meso-adaptivity a) in closed exercises if practice difficulty was elevated and b) in half-open exercises if the learner was already familiar with the supportive adjustments. In addition, learners with very low general language proficiency or high strategic competence may also benefit from the treatment. We also saw that mitigating exercise difficulty is not necessary if the prac-

tice material's difficulty is tailored towards the target group. Maintaining different pools of practice material for different target groups could thus sufficiently control exercise difficulty and free up macro-adaptivity to focus on linguistic variability.

Addressing the practicability of this approach, we present means to automatically generate large pools of systematically varied exercises. With a focus on ensuring grammatical and pedagogical validity, we interleave exercise generation with revision steps for more efficient quality assurance. We further refine the pedagogically informed initial pool of exercises through learning analytics based on learner errors.

The contribution of this thesis is thus twofold: (1) It addresses conceptual questions concerning the attention to variability in macro-adaptive language tutoring systems while considering curricular constraints. (2) It presents a fully implemented approach to adaptivity, which has been evaluated in several studies.

Zusammenfassung

Erlernen von Fremdsprachen ist ein zentraler Bestandteil heutiger Schulbildung, doch Lehrkräften fehlen oft die Ressourcen, um auf die individuellen Bedürfnisse der Schüler einzugehen. Intelligente Tutorensysteme können diese Schwäche prinzipiell überwinden. Sie ermöglichen individualisiertes Üben einer Materie in einer digitalen Lernumgebung, indem sie Lernmaterial eines Domänenmodells an Lernereigenschaften eines von ihnen gepflegten Lernermodells gemäß einem pädagogischen Modell anpassen. Ihre Adaptivitätsstrategien umfassen dabei Makroadaptivität zur Auswahl geeigneten Lernmaterials, wie auch Mikroadaptivität für Unterstützung bei der Bearbeitung der ausgewählten Materialien. Allerdings berücksichtigen die meisten intelligenten Sprachlernsysteme die Anforderungen, die durch die Einbindung in den Schulkontext entstehen, nur unzureichend. Zudem lässt ihr starker Fokus auf die Abstimmung von Aufgabenschwierigkeit auf die Fähigkeiten eines Schülers wenig Spielraum für weitere Anpassungen an den Lernenden. Insbesondere sprachliche Variabilität, die in Sprachen allgegenwärtig ist, spielt im Spracherwerb jedoch eine wichtige Rolle. Diese Dissertation untersucht daher, wie und in welchem Umfang sprachliche Variabilität in adaptiven Sprachlernsystemen, die mit kontextualisierten Grammatikaufgaben den Unterricht begleiten, berücksichtigt werden sollte und kann.

Zunächst wurden die Auswirkungen sprachlich variabler Aufgaben auf das Lernen untersucht. Dabei stellte sich heraus, dass variable Aufgaben zum einen Variabilität in von Lernern produzierten Texten, zum anderen jedoch auch die Aufgabenschwierigkeit beeinflussen. Zudem neigen variable Aufgaben dazu, deklaratives Wissen zu festigen, Prozeduralisierung jedoch zu verlangsamen. Eine lernförderliche Aufgabenauswahl zeichnet sich daher durch adaptive Anpassung sprachlicher Variabilität aus. Da sich verschiedene Varianten eines sprachlichen Mittels unterschiedlich auf die Aufgabenschwierigkeit auswirken, muss Makroadaptivität eine Balance zwischen einem Fokus auf Aufgabenschwierigkeit und auf sprachlicher Variabilität finden.

Daher wurde in der vorliegenden Arbeit Mesoadaptivität als alternativer Ansatz, die Schwierigkeit einer Aufgabe zu verringern, untersucht. Die Integration dieser Zwischenform von Mikro- und Makroadaptivität, die bei Problemen eines Schülers,

eine Aufgabe zu lösen, die Schwierigkeit der Aufgabe schrittweise reduziert, verringerte die Anzahl an falschen Schülerabgaben. Eine positive Auswirkung auf den Lernerfolg zeigte sich jedoch nur a) bei Anpassungen von geschlossenen Aufgaben unter erschwerten Übungsbedingungen und b) von offeneren Aufgaben, wenn die Lernenden mit den Hilfestellungen vertraut waren. Zudem scheinen Lernende mit geringer allgemeiner sprachlicher Kompetenz oder hoher strategischer Kompetenz ebenfalls von Mesoadaptivität zu profitieren. Es zeigte sich allerdings, dass eine Verringerung der Aufgabenschwierigkeit nicht notwendig ist, wenn die Schwierigkeit des Übungsmaterials auf die Zielgruppe abgestimmt ist. Unterschiedliche Aufgabenportfolios für unterschiedliche Zielgruppen könnten somit die Aufgabenschwierigkeit hinreichend regulieren, sodass der Fokus von Makroadaptivität auf die Anpassung sprachlicher Variabilität gerichtet werden kann.

Um die praktische Anwendung dieses Ansatzes zu erleichtern, stellt die Arbeit Mittel zur automatischen Generierung großer, systematisch variabler Aufgabenportfolios vor. Um einerseits grammatikalische und pädagogische Korrektheit der Aufgaben und andererseits eine effiziente Qualitätssicherung zu gewährleisten, werden mehrere Prüf- und Überarbeitungsschritte in die Aufgabengenerierung integriert. Zudem wird das ursprüngliche, pädagogisch motivierte Aufgabenportfolio mithilfe von Learning Analytics basierend auf von Schülern gemachten Fehlern verbessert.

Somit leistet diese Thesis sowohl einen konzeptionellen als auch einen technischen Beitrag: Zum einen erforscht sie die Berücksichtigung von Variabilität in makroadaptiven Sprachtutorensystemen unter Beachtung des Schulkontextes. Zum anderen stellt sie einen vollständig implementierten Adaptivitätsansatz vor, der in mehreren Studien evaluiert wurde.

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms and Abbreviations

AI	Artificial Intelligence
AJAX	Asynchronous JavaScript and XML
BKT	Bayesian Knowledge Tracing
CAF	Complexity, Accuracy, and Fluency
CEFR	Common European Framework of Reference
CI	Continuous Improvement
CSS	Cascading Style Sheets
DDF	Desirable Difficulty Framework
Didi	DigBindiff
EGP	English Grammar Profile
GUI	Graphical User Interface
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
I4S	Interact4School
IAA	Inter-Annotator Agreement
ICALL	Intelligent Computer-Assisted Language Learning
ILTS	Intelligent Language Tutoring System
IRT	Item Response Theory
ITS	Intelligent Tutoring System
JSON	JavaScript Object Notation
KC	Knowledge Component
KT	Knowledge Tracing

LSA	Latent Semantic Analysis
NLG	Natural Language Generation
NLP	Natural Language Processing
POLKE	Pedagogically oriented language knowledge extraction and readability-controllable natural language generation
POS	part-of-speech
PPP	Presentation-Practice-Production
RCT	Randomized Controlled Trial
RQ	research question
SLA	Second Language Acquisition
SRL	Self-Regulated Learning
TAP	Transfer-Appropriate Processing
TBLT	Task-Based Language Teaching
TSLT	Task-Supported Language Teaching
ZPD	Zone of Proximal Development

Chapter 1

Introduction and Motivation

In a more and more globalized world, knowing one or more foreign languages has become an indispensable skill. However, learning another language is not a trivial task. In order for practice to be effective, it should be deliberate, systematic, transfer-appropriate, with feedback, and of desirable difficulty (Suzuki, 2023). Human instructors struggle to consider all five dimensions in their teaching practices as they generally lack time, appropriate material, and suitable diagnostics (Aftab, 2016). The field of second language instruction is no exception. Counteracting this, modern technology has rapidly made its way into classrooms, first in reaction to the COVID-19 pandemic (Pozo et al., 2021), and driven further through the rise and potentials of generative AI tools such as ChatGPT[1] (Saville, 2023; Kohnke et al., 2023). Intelligent Language Tutoring Systems (ILTSs) can, in principle, account for all characteristics of effective language practice. Examples of such learning tools that are successfully used in real life include Heift's (2010) E-Tutor, Nagata's (2009) Robo-Sensei, and Amaral and Meurers's (2011) Tagarela. ILTSs individualize practice through two types of user-adaptivity (Rus et al., 2015). Referred to as *outer loop*, *curriculum sequencing* or *macro-adaptivity*, the first type focuses on determining the next practice item. The second type, often interchangeably called *inner loop*, *problem solving and solution analysis tutor*, or *micro-adaptivity*, provides support for a selected practice item while the learner works on it (Pelánek, 2017). We will subsequently refer to these two types as macro-adaptivity and micro-adaptivity. Macro-adaptive exercise sequencing can account for systematic presen-

[1] ChatGPT is available at https://chatgpt.com/

tation of transfer-appropriate practice material at desirable levels of difficulty. On this practice material, micro-adaptive scaffolding can provide feedback for learner input in order to successively guide a learner towards the correct answer. For deliberate practice of goal-oriented, individualized exercises focusing on specific linguistic constructions, Mastery Learning has become a widely adopted teaching approach promoting that with appropriate instruction and effort, most learners can eventually master a topic (Pelánek and Řihák, 2017). Mastery of a domain is a multifaceted construct representing a learner's knowledge state of many and diverse Knowledge Components (KCs). Intelligent Tutoring Systems (ITSs) can, thanks to their ability to maintain detailed learner models, track and monitor a learner's progress for each of these KCs. ITSs then typically either use dashboards to expose learners to their progressing knowledge states in order to enable them to make informed decisions about how to navigate through the learning material (Bull et al., 2016; Twigg, 2003), or use the learner models to adaptively select learning material tailored to a knowledge state (Chrysafiadi and Virvou, 2013). They thus constitute digital tools that aim to teach knowledge of a domain in individualized practice (Graesser et al., 2012). To this purpose, they pursue macro-adaptive strategies that identify the most suitable learning material for a learner at a given point in time, and micro-adaptive strategies that provide support tailored to the individual student's needs. The adaptivity strategies rely on a domain model that defines the knowledge to acquire, on a learner model that tracks a learner's evolving traits, and on an instructional model that defines the strategies to adapt learning material of the domain model to learner characteristics of the learner model (Burns and Capps, 1988).

Adaptive exercise selection focuses on matching exercise difficulty to a learner's abilities (Park and Lee, 2004). This is straightforward for so-called well-defined domains (Fournier-Viger et al., 2010). In ill-defined domains, on the other hand, tutoring systems usually require domain-specific approaches. Language learning is considered such an ill-defined domain. This is usually attributed to ambiguity inherent to language. Out of the five characteristics of ill-defined domains listed by Fournier-Viger et al. (2010), overlapping sub-problems constitutes the one that most notably applies to language learning, as any sentence involves multiple and varied sub-problems in the shape of linguistic forms. Complicating this further, ITSs are best combined with teacher-orchestrated instruction (Corbett et al., 2001; Slavuj et al., 2015). Such blended learning settings are often subject to curricular constraints imposed by the

instructional setting (Phillips et al., 2020). These constraints may restrict the scope of linguistic forms included in the practice material. Yet due to exercise-intrinsic linguistic variability it is not always possible to exclude all undesired linguistic forms (Katinskaia et al., 2023). Existing approaches to domain-specific, adaptive sequencing of form-focused grammar exercises fail to take into account both the inherent linguistic variability of practice items and external curricular constraints on exercise selection. This thesis aims to pinpoint the nature of these two factors, outline their impact on adaptive exercise selection, and present a novel approach to exercise selection that remedies this fundamental shortcoming of existing approaches.

Figure 1.1 provides a simplified overview of the central research questions around which this thesis is centered. The basic idea assumes that a macro-adaptive ILTS aims to select exercises matching the learner's abilities from a limited pool of exercises. The curriculum that a learner follows restricts the exercises within this pool. Exercise difficulty further restricts the exercises that are suitable for the learner. The approach of this thesis in addition considers linguistic variability in learner-adaptive exercise selection as a third constraint. In order to increase the number of suitable exercise candidates, we hypothesize that an intermediate form of adaptivity operating between macro- and micro-adaptivity, which we call *meso-adaptivity*, mitigates exercise difficulty so that otherwise too difficult exercises may be considered for selection.

In order to verify this hypothesis, we focus on contextualized form-focused grammar exercises in adaptive ILTSs. In terms of practice material, we consider grammar topics relevant to 7^{th} grade students of English in German secondary schools.[2] We cover both closed exercises, where learners need to identify to correct answer from a limited set of candidates, such as Mapping, Jumbled Sentence, Mark-the-Words, and Categorization exercises, and half-open exercises, where learners need to produce the target answer, including Gap-Filling and Short Answer exercises. Since all exercises are contextualized, they comprise a target encompassing the linguistic form to practice, which is embedded into some co-text. In order to develop the concept for our envisioned system, we first assess the implications of integrating an ILTS

[2]In order to make explicit that the setting constitutes one of formal instruction, learners may be referred to as *students* (Applefield et al., 2000). Since all work presented in this thesis focuses on school settings, we here use the term *student* and the more general term *learner* (Pelánek, 2022) interchangeably.

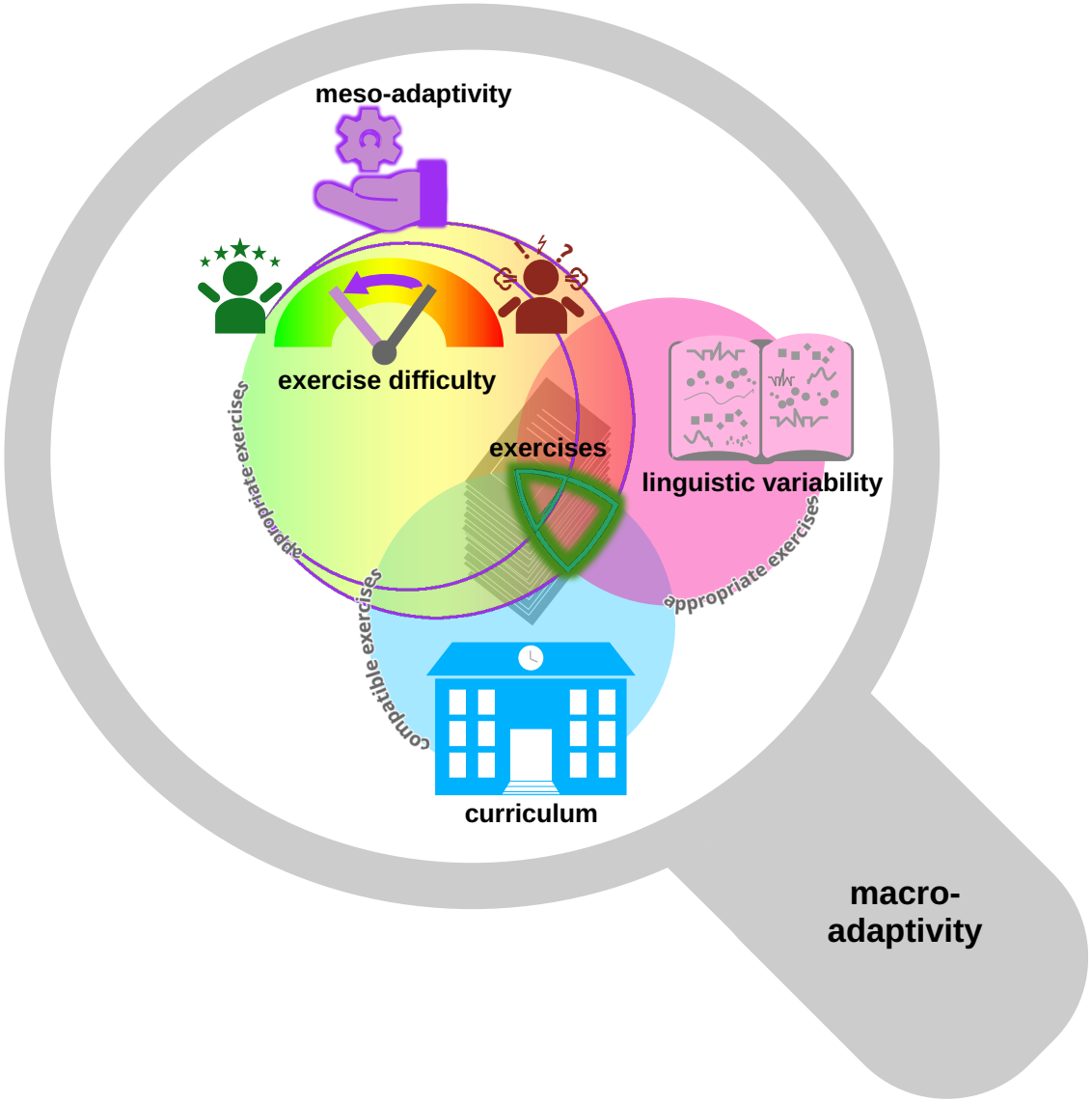


Figure 1.1: Overview of central research questions.
The central hypotheses assume that meso-adaptivity can increase the number of exercises that are suitable for a learner. Suitable exercises are restricted by the learner's curriculum, exercise difficulty, and linguistic variability.

into an instructional setting in Chapter 2, of considering linguistic variability in Chapter 3, and of reducing the focus on exercise difficulty when selecting exercises in Chapter 4. Based on the findings from these assessments, Chapter 5 presents the implementation of an adaptive ILTS that balances all three aspects. While the mechanism of adaptive language learning constitutes the functional goal of our approach, Chapter 6 illustrates the means to make it feasible in the form of automatic

exercise generation. Chapter 7 concludes with directions for future work.

This thesis is structured along ten research questions (RQs). Figure 1.2 summarizes its scope and visualizes links between the RQs. Since the thesis is based on data-driven analyses, connections between some nodes of the graph are verified or rejected as we answer the corresponding research questions.

With respect to the curricular constraints assessed in Chapter 2, typical macro-adaptive sequencing differs from instructional practices in school settings in the way practice material is organized and sequenced. The representation of a learner's knowledge state can be used in adaptive ITSs to determine when to move on to the next KC according to Mastery Learning (Pelánek, 2018). In contrast, most instructional settings structure the introduction and practice of grammar topics in a fixed sequence that the institution or the teacher deems pedagogically valuable (Irfani, 2017). When combining both approaches, the order in which KCs are practiced is therefore not entirely left to adaptive sequencing, but constrained by this curricular order. We investigate whether a prototypical structure defined by an instructional setting interferes with adaptive sequencing that aims to select practice items matching a learner's abilities.

Research Question 1

Is a prototypical curricular organization of learning targets compatible with adaptive sequencing implementing Mastery Learning?

A second impact of curricular structures concerns the notion of forgetting. Adaptive algorithms often integrate this concept and interleave revision exercises and exercises of not yet mastered KCs (Song and Luo, 2023). Institutional curricula, on the other hand, tend to include separate revision units for a specific linguistic learning target (Varanoglulari et al., 2008). We therefore discuss possible compromises to combine the two approaches with the aim of integrating the benefits of both.

Research Question 2

How can revision practice combine designated learning units introduced by a curriculum and adaptive, interleaved practice?

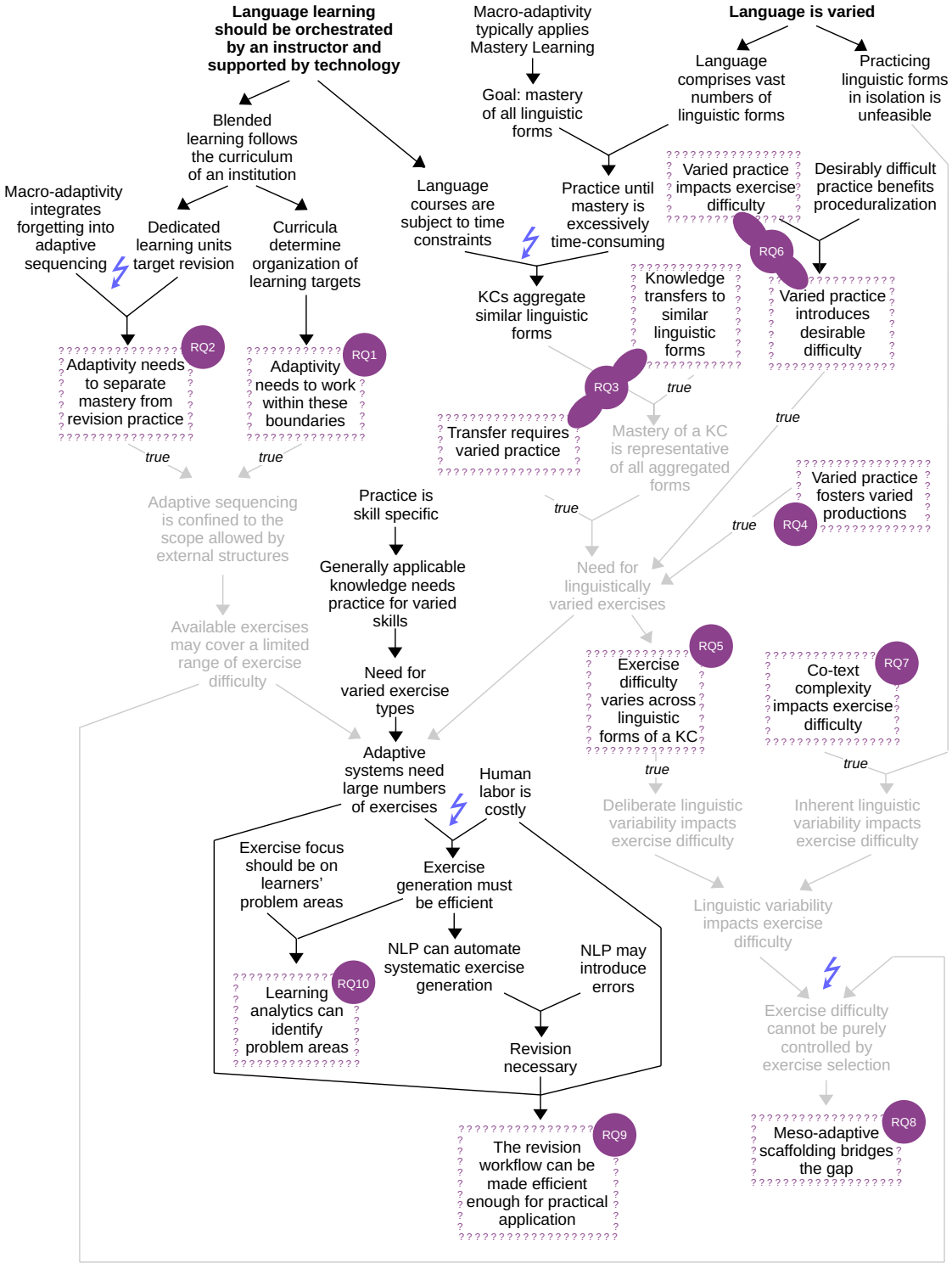


Figure 1.2: Scope of the thesis.
The arrows represent conditional dependencies between the nodes of the graph. Nodes that apply only if the preceding RQs are answered to the affirmative are shaded in gray.

While curricular structures restrict exercise selection by narrowing its scope, variability – assessed in Chapter 3 – affects it through the inherent difficulty of each variation. Since, as Raviv et al. (2022) point out, 'all experiences are unique', the acquired knowledge needs to be generalizable, which may be achieved through variability in the input. Diverse linguistic variants can either be inherent to the linguistic co-text material of the practice item, or deliberately introduced. Variability is therefore not only omnipresent in exercises, but also often intentionally introduced in order to foster transfer of knowledge (e.g., Lightbrown, 2007; DeKeyser, 2018). In language learning, research evolving around transfer has so far primarily focused on skill specificity of knowledge (DeKeyser, 2018), as well as on transfer to new contexts (Suzuki, 2022). Linguistic variability, although extensively researched, has so far not been considered in instructional contexts (Raviv et al., 2022).

The motivation for deliberately introducing linguistic variability originates from the definition of KCs on which Mastery Learning builds. They are defined as well-specified, fine-grained items (Pelánek and Řihák, 2017). The space of relevant KCs is usually represented in a domain model (Pelánek, 2019). For practical applications, Pelánek and Řihák (2018) point out the challenge of considering the hundreds of KCs comprising the domain covered by an educational system. In language learning, where linguistic forms constitute KCs, their sheer numbers make it infeasible in school settings to expect learners to practice them all until mastery. When considering Skill Acquisition Theory this issue is further aggravated by the fact that mastery of linguistic forms does not transfer to other contexts, tasks, or skills (Lyster and Sato, 2013). A practical approach therefore needs to define KCs on a higher level that aggregates multiple linguistic forms with similar properties (Pelánek, 2018). In such an approach, mastery can only be representative of the entire KC if mastery of one linguistic form transfers to the other forms of the aggregate KC. However, implementations of Mastery Learning have so far not established whether this is the case for the KCs they define. We therefore in a first step analyze whether transfer of knowledge occurs between linguistic forms with similar properties and whether such transfer is subject to conditions such as varied practice of different linguistic forms.

Research Question 3

Does knowledge of a linguistic form transfer to other linguistic forms with similar properties? Is this transfer conditioned by varied practice?

Even if KCs cannot validly be defined as aggregates of linguistic forms with similar properties, this approach might still provide practicable approximations for mastery of KCs in instructional settings with constraints on the time that is available for practice (Bloom, 1968). Varied practice raises a learner's awareness of varied linguistic forms. Through practice, the process of proceduralization incorporates declarative knowledge into production rules, which reduces processing time and the likelihood of errors (Anderson, 1983). Since noticing is a prerequisite for proceduralization (Schmidt, 1990), varied practice potentially prevents under-use and overgeneralization of knowledge (Hirschmann et al., 2013; James, 2018). This raises an additional question in whether varied practice is beneficial for proceduralization and for varied learner productions in open tasks. In this case, linguistic forms within each KC should be varied in order to cover as many of them as possible. This broader coverage of linguistic variants in addition increases the likelihood of a learner having seen the exact form that they encounter after practice, which they can then retrieve from memory (Raviv et al., 2022).

Research Question 4

Is varied practice beneficial for proceduralization? Does it result in more varied learner productions?

Each form of an aggregate KC also introduces properties that are not shared with the other forms of the KC. Different properties might have different effects on difficulty (Pandarova et al., 2019) and thus introduce variation in complexity within each KC. We therefore analyze whether different linguistic forms of the same aggregate KC affect exercise difficulty differently.

Research Question 5

Does exercise difficulty vary across linguistic forms of similar properties?

Difficulty might not only vary across linguistic forms, but also be affected by the

linguistic diversity in the practice material. Linguistic forms may appear in varied combinations in practice material. Allowing more diverse forms in an exercise broadens the focus of the learning target. While practice is most effective for proceduralization and long-term learning gains if it is desirably difficult, research yet needs to establish whether difficulty introduced by linguistically varied practice is indeed desirable.

Research Question 6

Does linguistically varied practice increase practice difficulty? If so, is this difficulty desirable?

In order to assign exercises to learners that are adequately difficult for them, adaptive ITSs attempt to sequence and provide scaffolding support for exercises in such a way that they always present a slight challenge to the learner (Murray and Arroyo, 2002). In Flow Theory, this corresponds to the flow channel in which the learner is neither bored nor too anxious (Csíkszentmihályi, 1991). In language learning, it is referred to as the learner's Zone of Proximal Development (ZPD) (Vygotsky, 1978). In most domains, including content domains such as History or Geography, as well as Maths, the next KC to practice is therefore determined based on its difficulty in relation to the learner's knowledge state. When providing contextualized exercises, language learning again poses challenges to this task due to linguistic variability inherent to the exercises. Sentences constitute complex linguistic constructs comprised of a variety of linguistic forms whose co-occurrences vary across as well as within speakers (Yildirim et al., 2013). It is therefore almost impossible to practice a single KC in isolation. Language practice thus inevitably makes use of multiple and diverse linguistic forms in the co-text of each practice item which may influence the difficulty of that item in different ways (Pandarova et al., 2019). This linguistic variability in the co-text is even more pronounced when exercises are based on authentic language material. The advent of content and meaning focused instruction doctrines such as Task-Based Language Teaching (TBLT) invite authentic material in order to promote practice of linguistic forms in context and with a focus on a functional target task (Thomas and Reinders, 2013). When integrating authentic language material, it is not even feasible to always practice a KC in similar linguistic contexts. However, linguistic forms of the co-text are not the focus of the practice

item so that passive knowledge of them is usually sufficient. Insights from reading comprehension analyses therefore provide preliminary indicators of whether varied linguistic forms in the co-text impact exercise difficulty. Research in this domain has indeed found effects of text complexity on exercise difficulty (Hartig et al., 2012). Whether this only applies to meaning-focused tasks or whether co-text complexity also affects the difficulty of form-focused grammar exercises, on the other hand, remains an open research question that we aim to answer.

Research Question 7

Does co-text complexity impact the difficulty of form-focused grammar exercises?

Considering how linguistic variability affects exercise difficulty, the pool of exercises restricted by curricular constraints will not always cover the difficulty level required by the learner. Aggravating this further, in order for the acquired knowledge to be more widely applicable, Transfer-Appropriate Processing (TAP) requires it to be practiced in a variety of exercises targeting different skills (Lightbrown, 2007; DeKeyser, 2018). These are typically implemented as different exercise types, which also have inherent difficulties (Grellet, 1981, p. 5). Learners should on the one hand be exposed to all exercise types within the time available for practice, and on the other hand always be able to solve the exercises in order to avoid de-motivation and frustration (Murray and Arroyo, 2002). Chapter 4 therefore assesses means to balance time constraints against learning goals without overloading learners. The ZPD advocates material that learners can successfully process with scaffolding help (Vygotsky, 1978). If adaptive exercise sequencing pre-selects exercise candidates that comply with developmental readiness to some degree, scaffolding might be able to bridge the remaining gap between learner abilities and exercise difficulty. We therefore investigate whether providing extensive scaffolding in the form of meso-adaptivity that dynamically adjusts an exercise can mitigate exercise difficulty sufficiently to enable learners to successfully complete most, if not all, exercises. If this proves to be the case, then adaptivity in language learning can be reformed to consider linguistic variability and curricular constraints through macro-adaptive exercise sequencing. Exercise difficulty can then be fine-tuned to the learner's abilities through dynamic meso-adaptive adjustments of exercise elements.

Research Question 8

Can meso-adaptivity mitigate exercise difficulty sufficiently to enable learners to solve (almost) all exercises?

Once these RQs have been answered, we know the requirements for our envisioned macro-adaptive ILTS presented in Chapter 5.

An approach to adaptive practice that limits the pool of exercise candidates by considering linguistic variability and curricular constraints requires large numbers of exercises with systematically varied linguistic forms for different semantic topics. This process, discussed in Chapter 6, should therefore be automated as much as possible. In order to still ensure that learners are not exposed to incorrect practice material, all exercises should be reviewed and corrected if necessary (Keller, 2021). If revision is not integrated into exercise generation in an efficient way, thorough revision becomes infeasible in real-world contexts given the time required to review all generated exercises. In order to remove this hurdle on the way to bringing ILTSs that adaptively consider variability into real-world instructional settings, we outline an approach that supports content creation for such systems.

Research Question 9

How can exercise generation efficiently support the creation of valid and linguistically varied exercises?

In instructional settings, pedagogical validity of the practice material is particularly important. ILTSs typically rely on domain experts to determine what constitutes relevant and pedagogically valid content (Katinskaia et al., 2023). These initial conceptualizations basically represent human intuition. Once learner data has been collected, however, learning analytics based on learner errors can identify learners' problem areas that the practice material should cover. To this purpose, we propose and evaluate an application of Continuous Improvement (CI) based on learning analytics. By applying the suggested analyses and adjustments on a regular basis, exercise generation can be continuously improved as more learner data becomes available.

Research Question 10

How can learning analytics support continuous improvement of exercise generation?

By outlining, in the following chapters, the approach to considering linguistic variability in macro-adaptive systems used in instructional settings in such a comprehensive manner, we aim to facilitate its adaptation into real-world contexts as much as possible.

Chapter 2

Curricular Constraints Restricting Exercise Selection

Parts of this chapter are based on Publications 1, 3, 4, 5 and 6.

Learning one or more foreign languages is compulsory for students in all European countries, as well as in other parts of the world including Canada, Morocco, Thailand, and Kazakhstan (Pufahl et al., 2001). However, classroom teachers lack the time and means to provide individualized practice for their students (Aftab, 2016). ILTSs, on the other hand, integrate the benefits of modern technology that allow to cater to each student's individual needs (Slavuj et al., 2015). Intelligent Computer-Assisted Language Learning (ICALL) therefore aims to integrate the use of ILTSs in instructional settings, thus combining the benefits of both (Graham, 2005).

Integrating an ILTS into classroom teaching only makes sense if using the system helps learners to better or faster reach the course goals. The ILTS therefore needs to always align with the content taught by the teacher (Gündüz, 2005). This requires the material to on the one hand cover the lexical and grammatical content of the lessons and to on the other hand be presented to the learners at about the same time as it is covered in class (Heift, 2016). Since the range of language classes following different lesson plans is immense (Bliesener, 1998), a widely used ILTS needs to be able to flexibly adjust to different lesson plans. Such flexibility can only be achieved through the teacher's input, who determines the schedule and specific content of each lesson (Kertz-Welzel, 2004). The teacher retains the role of an orchestrator

and uses the ILTS to supplement classroom instruction.

This implies that the decision about what learning material is presented to learners at what point in time does not entirely reside with the system. Instead, the ILTS must provide a certain degree of navigational freedom.

In this chapter, we focus on research questions 1 and 2 with the aim of identifying curricular constraints that might restrict exercise selection. While RQ 1 assesses practice exercises in the context of Mastery Learning, RQ 2 examines the integration of revision exercises in contrast to introductory exercises on the one hand and differentiating between remedial and refresher exercises on the other hand.

Research Questions

RQ1 Is a prototypical curricular organization of learning targets compatible with adaptive sequencing implementing Mastery Learning?

RQ2 How can revision practice combine designated learning units introduced by a curriculum and adaptive, interleaved practice?

2.1 Adaptive Exercise Selection within Curricular Structures

Since the use of technology in school classrooms is being expected to an increasing extent (Lim et al., 2013), ITSs are emerging more and more in teacher-orchestrated instruction settings. The benefits are obvious as the systems allow learners to engage with practice material outside of class without the time limitations set by classroom teaching. However, the integration is not always successful, mostly because common ITSs do not easily support classroom settings (Phillips et al., 2020). In order to facilitate a successful integration, in RQ 1 we look at two particularly challenging aspects, the structure of learning material and revision of previously introduced learning targets, and analyze how the approach of adaptive ILTSs can be reconciled with that of teacher-orchestrated instruction in school settings.

2.1.1 Aims and Scope of Adaptive Exercise Selection

Intelligent Tutoring Systems provide ample opportunities for adaptation. While Adaptive Hypermedia Systems focus on adapting the presentation of the content, ITSs adapt on the two levels of macro- and micro-adaptivity (Vanlehn, 2006; Brusilovsky, 2000). These two levels of adaptivity in ITSs have received considerable attention in various domains, often borrowing from practices in non-educational contexts. Their purpose is to improve learning outcomes by adapting to the students' individual needs (Phobun and Vicheanpanya, 2010).

If the goal of macro-adaptivity is to find an optimal order in which to practice a given set of exercises, Tabari and Cho (2022) confirmed for writing tasks with varying amounts of planning time that simple-to-complex sequences were most effective for learning outcomes. If, on the other hand, the goal is not to complete all exercises but to instead reach mastery of a learning target, each learner may require different exercises in varying numbers. Since the introduction of the concept of Mastery Learning by Bloom (1968) some 65 years ago, the focus of most adaptive ITSs has been on individualizing practice for learners in a way that allows most learners to reach the learning goals at some point. The task of macro-adaptive exercise selection then consists in identifying those exercises from a large pool of candidates that are suitable for the given learner at the current point in time.

In order to select exercises adapted to the learner, the system needs to keep a record of the learner's characteristics on which it bases the selection criteria. This is generally done with a learner model that is populated and adjusted as the learner interacts with the system. The type of data and the structure in which it is stored depend on the specific learner model. In their literature review, Chrysafiadi and Virvou (2013) pinpoint which learner model types are best suited to represent what kind of learner data, as outlined in Table 2.2.

Stereotype learner models focus on demographics and cognitive characteristics (Muhammad et al., 2016; Wang et al., 2009; Hafidi and Bensebaa, 2014). They differentiate between a defined number of learning styles for which they provide customized learning experiences (Rich, 1979). Which learning styles the model differentiates is usually based on an established taxonomy (Troussas et al., 2021; Grubišić

	Learner characteristics
Stereotype model (Muhammad et al., 2016; Wang et al., 2009; Hafidi and Bensebaa, 2014)	- demographics: age, gender, language, educational background, background knowledge, initial proficiency - cognitive characteristics: learning styles, preferences, goals, personality, cognitive style, language affinity, memory, attention, concentration, learning affinity, perception, problem solving, decision-making/ analytical abilities, critical thinking/ communication/ collaboration skills
Scalar model (Brusilovsky and Millán, 2007)	- proficiency
Overlay model (Brusilovsky, 1992)	- domain knowledge
Bayesian model (Peña et al., 2011; Millán et al., 2001)	- domain knowledge - affective factors: motivation, emotions, feelings - meta-cognitive factors: reflection, self-awareness/ -monitoring/-regulation/-assessment/-management/ -explanation - learning styles - misconceptions
Bug model (Brusilovsky, 1994)	- errors - mistakes

Table 2.2: Learner characteristics represented in different learner models.

Different learner models maintain records of different learner characteristics.

et al., 2013). New users are then associated with the most probable stereotype according to their characteristics. While stereotype models consider a range of learner characteristics, **scalar models** take only a learner's domain knowledge into account (Brusilovsky and Millán, 2007). They represent this knowledge, typically assessed through self-estimation, as a single value on a scale from low to high proficiency. A richer representation is available in **overlay models**. Focusing on the learner's mastery of the domain knowledge, they constitute subsets of a domain model enriched with the learner's mastery of the domain model components as boolean or scalar values (Brusilovsky, 1992). A special sub-variant, the differential model, additionally differentiates between compulsory and optional domain model knowledge (Millán et al., 2010). **Bayesian learner models** can in addition take into consideration dependencies between domain model components and other learner characteristics

such as affective factors, meta-cognitive factors, learning styles, and also misconceptions (Peña et al., 2011; Millán et al., 2001). To explicitly model systematic learner errors and accidental mistakes, **bug models** are often used (Brusilovsky, 1994). Also referred to as *buggy models* or *error models*, they come in different shapes (Brusilovsky and Millán, 2007). Perturbation models extend overlay models with misconceptions that they identify and track through learner errors. Genetic models, while allowing to trace a learner's development to more and more specialized and generalized knowledge and incorporating incorrect knowledge representations, have never been used in real life due to their high complexity. Constraint-based models rely on a domain specification representing constraints (Ohlsson, 1994). A constraint thereby consists of a relevance clause determining whether the constraint is applicable, and a satisfaction clause that determines whether the constraint has been violated. The bug model then contains constraints for which the relevance clause was fulfilled but the satisfaction clause violated.

A practical differentiation of learner model components is given by Esichaikul et al. (2011). They maintain a knowledge model that tracks the learner's progress in the domain model separately from a collection of domain independent information including cognitive abilities, affective state, experience, and learning style. Bayesian models have been developed in a range of variants. The existence of an efficiency-centric version with reduced dependency modeling (Mayo and Mitrovic, 2001) suggests that there is a trade-off between efficiency in using the model and explicitness in the representations. While there may be advantages in modeling all aspects of a learner model jointly, and possibly with dependencies between them, Esichaikul et al.'s (2011) approach constitutes a more practical approach. Domain independent characteristics can be treated similarly across domains whereas the knowledge model requires special attention depending on the content the ITS aims to transmit. In language learning, this is where computational linguistics comes into play.

Another differentiation highlighted by Chrysafiadi and Virvou (2013) separates static learner characteristics such as age or mother tongue from dynamic characteristics including prior and current 'knowledge and skills, errors and misconceptions, learning styles and preferences, [and] affective, ... cognitive ... and meta-cognitive factors'. While static characteristics are collected once upon registration with the system through explicit questionnaires, dynamic characteristics need to be

constantly inferred and updated from a learner's answers to practice material and interactions with the tutoring system (Devedžic, 2006).

Based on the learner model, the system can adapt the content to the learner. Some approaches to adaptive sequencing in addition consider environment parameters such as the device, location, noise, connectivity, and time constraints (Ennouamani and Mahani, 2017; Muhammad et al., 2016).

Macro-adaptive sequencing of learning content in principle constitutes a robot path planning problem (Mac et al., 2016). Macro-adaptive systems therefore pursue either global or local path planning strategies. Global approaches determine the entire learning path for each individual learner prior to starting a distinct learning unit. Local approaches determine the next practice material within each learning unit dynamically after completing the previous one.

Global approaches usually estimate learner abilities in a pre-test or questionnaire the learner has to complete before starting the learning unit. A rather simple approach to macro-adaptivity constitutes **fuzzy logic**, which creates a learning path by computing the similarity of learning objects to the learner's misconceptions identified in a pre-test and ordering them by relevance for remedial practice (Hsieh et al., 2013). Relying on expert knowledge, **decision trees** classify learners into profiles that are associated with learning paths. They are easily interpretable, yet quickly become hard to manage the more parameters are considered (Wang et al., 2009). **Evolutionary algorithms** try to iteratively improve a naively initialized learning path through a set of well-defined transformation rules (Rasheed and Wahid, 2019). They select the learning path based on learner characteristics and have been explored in various forms to this purpose, including genetic algorithms, Ant Colony Optimization, Particle Swarm Optimization, and Harmony Search. In genetic algorithm approaches, exercises are represented as genes and assembled into chromosomes representing the individualized learning paths based on knowledge gaps identified in the pre-test (Chen, 2008). Ant Colony Optimization can be used to identify the shortest learning path that leads towards the learning goal by determining the path most often followed by learners who have already completed the learning unit (Elshani, 2021). Also relying on available learner data, Particle Swarm Optimization considers local and global maxima in order to find the best path towards completing the learning goal given the learner profile and the domain model (Rasheed and Wahid,

2019). The Harmony Search Algorithm also considers the learner profile and the domain model, but is simpler than the other approaches as its memory encompasses only a single previous iteration (Hnida et al., 2018). **Artificial Neural Networks** are mainly used to adapt to individual learning styles by clustering available data of learners with similar learning paths and assigning new learners to these clusters (Chrysafiadi et al., 2019). Since they need large amounts of learner data for training, they are often combined with expert rules that are overwritten as learner data becomes available (Rasheed and Wahid, 2019). Alternative approaches aiming to adapt to the learning style include **Bayesian Networks** that maintain probabilities conditioned by the learning style, as well as **ontology-based** models that match concept classes representing knowledge components to learning styles (Chrysafiadi et al., 2019; Elshani, 2021). **Learning graph** based approaches determine the shortest possible learning path based on weights assigned to the edges in the graph. The parameters of the suitability function used to calculate the weights are determined based on existing learner data (Karampiperis and Sampson, 2005).

While these approaches provide personalized learning paths, they cannot account for changing learner characteristics or dynamically adjust to identified misconceptions or learning speed as the learner progresses towards the learning goals. In principle, most global approaches could be applied again after each learner interaction with the system to re-calculate the entire learning path with the already covered path as an additional restriction. For more efficient solutions, however, local approaches have been integrated into a number of systems. **Elo Rating Systems**, originally developed to assess chess players, update learner ability and exercise difficulty estimates after each interaction and identify exercises whose difficulty scores match the learner's ability level (Antal, 2013). In static assessments, **Item Response Theory (IRT)** can estimate learner abilities and exercise difficulties simultaneously. Longitudinal IRT emerged from the need to model changes in latent learner traits for adaptive sequencing (Wang and Nydick, 2020). Also aiming to assess a learner's proficiency through a minimal number of exercises, **Space Theory** maintains the domain knowledge in AND/OR graphs where exercises representing the same knowledge state are connected through OR relations and prerequisite knowledge states through AND relations (Doignon and Falmagne, 1985). Approaches using **reinforcement learning** typically maintain a so-called agent for each learner. This agent tries to optimize exercise selection based on feedback in the form of the

evaluation of the learner's answer (Bennane, 2013).

All these approaches, however, assign a global difficulty score to an exercise. By doing so, they fail to consider the many facets of factors contributing to exercise difficulty and the varied instantiations of these factors and facets in learner profiles (Beinborn, 2016). **Multidimensional IRT** (Park et al., 2019) addresses this issue by tracking a multitude of different skills instead of a single, aggregated skill. **Knowledge Tracing (KT)** approaches explicitly focus on and track performance for KCs that may be practiced by multiple exercises instead of considering individual exercise instances (Shen et al., 2024). Each KC is practiced until the learner masters it. KT determines when to move on to the next KC, but does not prescribe which KC to practice next. It encompasses a range of probabilistic, logistic, and deep learning-based approaches that model a learner's progressing knowledge state. Learning Factor Analysis, for instance, identifies over- and under-practiced KCs based on a learner's practice frequency of that KC in relation to the component's difficulty and learning rate (Cen et al., 2006). An extension to that approach, Performance Factor Analysis, in addition takes the learner's performance into account instead of assuming identical learning rates for all learners (Pavlik et al., 2009). Bayesian Knowledge Tracing (BKT) in particular can also account for the concept of forgetting (Pelánek and Řihák, 2017).

Adaptive exercise selection thus aims to select exercises matching various learner characteristics, most notably a learner's proficiency, in order to efficiently master all KCs of the domain. Depending on the approach, the focus of exercise selection is on determining when to practice the next KC, which KC to practice next, or which exercise to practice for a specific KC.

2.1.2 Aims and Scope of Language Curricula

In a classroom instruction setting, the general lesson plan is determined by the teacher or a higher institution (Erss, 2018). The general structure of the learning content is dictated by the syllabus. Within the bounds set by the curriculum, language teachers need to have certain control over the practiced material in order to be able to discuss specific linguistic forms and exercises in class (Burstein et al., 2012; Feng et al., 2014; Singh et al., 2011). Sabbah (2018) summarizes different syllabus approaches where the categories according to which the content is orga-

nized differ considerably. In a **structural syllabus**, the content is structured by grammar topics, usually sequenced by their complexity. In reaction to criticism concerning their lack of contextualization and their mechanical drill notion, alternatives to these traditional syllabi have emerged. **Semantic syllabi** focus on vocabulary, topic, or context. The grammatical constructs taught by a learning unit then depend on the text material that is presented. In an attempt to combine the strengths of these two approaches, **notional syllabi** take notions such as *time* or *cause* as KCs, which aims at more communicative language use by focusing on categories of communicative function. Focusing less on specific language components but rather on pedagogical processes, **process-oriented syllabi** have recently received increasing attention. **Communicative syllabi** aim to teach all skills required for communication including important topics, functions, notions, and vocabulary (Astete, 2015). In **task-based syllabi**, linguistic aspects become even less central.

Task-Based Language Teaching (TBLT) has become the predominant approach to language learning (Müller-Hartmann and Schocker-von Ditfurth, 2013; Ní Chiaráin and Ní Chasaide, 2020) as it can offer a less monotonous learning experience than traditional grammar-focused instruction with de-contextualized exercises (Doughty and Long, 2003). Form-focused grammar exercises have indeed often been criticized for their lack of a communicative context which, according to the critics, prevents learners from transferring the knowledge of practiced grammar constructs to new contexts (Lightbrown, 2019). TBLT, on the other hand, introduces language means in the developmentally natural order in which learners produce them. Fully implementing this communication-based method with a focus on meaning and a functional target task (Thomas and Reinders, 2013) within an ILTS is challenging. Creating complex learning cycles with function-focused final tasks preceded by step-wise pre-task activities supporting practice of the task-essential language aspects requires considerable human effort. Also, research suggests that in communication-based learning, learners often fail to notice, and therefore master, low-frequency linguistic constructions (Lightbrown, 2007). According to Shea's (2002) definition, the functional target task constitutes a complex task. For such tasks, research in motor skill practice has found that instructions and feedback should focus on the outcome of the task, not on technical aspects. Considering that insights from motor practice tend to also apply to language learning (Suzuki et al., 2022; Apfelbaum et al., 2019) this implies that the functional target task should not focus on linguistic aspects (i.e.,

grammar and vocabulary) but on the functional goal. However, with motor skills, poor technique tends to also result in poor outcomes. In language learning, on the other hand, the functional goal can usually be attained even if the accuracy of the speaker's language is poor so that the speaker remains unaware of their errors. It is therefore important to focus on vocabulary and grammar in precursory, simple exercises that prepare for the target task. Form-focused grammar exercises can fulfill all criteria of simple tasks: They can be designed to have a single degree of freedom, allow mastery of the task within one practice session, and seem artificial (Shea, 2002). In order to facilitate the transition from simple to complex tasks, the degree of freedom may be increased gradually by integrating only a single up to multiple linguistic forms into one exercise. However, current implementations of TBLT in ILTSs lack deliberateness of practice due to a lack of focus on form (Peterson, 2013) so that the learner's attention is not deliberately focused on linguistic rules (Tarone, 2018). The approach's somewhat weaker form, Task-Supported Language Teaching (TSLT), does support such a focus (Thomas and Reinders, 2013). In addition, Li et al. (2016) found that TSLT, where working on a task follows a phase of explicit instruction, yielded better learning outcomes than TBLT for grammar topics targeted in a cycle. It therefore makes sense to focus on this approach when integrating technology into classroom teaching. Following a Presentation-Practice-Production (PPP) model as backbone (Ur, 2018), TSLT explicitly teaches new concepts in the presentation phase, uses traditional form-focused exercises in the practice phase and more meaning-focused practice in the final task of the production phase. The practice phase serves the purpose of task planning to increase task-internal readiness of the target task by familiarizing the learner with the topic in terms of relevant learning targets and the task types (Gavin, 2014). When integrating TBLT into an ILTS, Colling et al. (2024) propose to explicitly assess this readiness through so-called diagnostic exercises.

Even though education in Germany is the jurisdiction of the individual federal states (Schecker and Parchmann, 2007), the workbooks used in German academic-track schools cover roughly the same grammar topics nation-wide (cf. Jones et al., 2020; Hanus et al., 2020; Harger and Niemitz-Rossant, 2014; Derkow-Disselbeck et al., 2008). The order in which they are introduced, however, differs considerably. The definition of these grammar topics is in some cases very general (e.g., simple past, questions), and in other cases rather specific (e.g., contact clauses, modal substi-

tutes). The scope of curricular restrictions therefore varies across states, and teachers require means to override the ILTS's automated navigational mechanisms.

2.1.3 Balancing Macro-adaptive and Curricular Aims and Scopes

The challenges of integrating adaptivity and curricular constraints into an ILTS differ depending on the macro-adaptive strategy and the institutional syllabus.

In structural syllabi, the structure of grammar topics is determined by the syllabus. In any other type of syllabus, the structure of grammar is independent of the syllabus structure within the units dictated by that syllabus. Macro-adaptivity is challenged most when the focus of adaptivity is identical to the focus of the syllabus. If the strategies and scopes of the adaptivity module and the syllabus clash, it is not possible to reconcile the two. Consider, for example, a task-based syllabus and an adaptivity strategy focusing on grammar practice. If the adaptivity module were the single organizing power, it would organize both high-level grammatical learning targets and low-level linguistic constructs within these learning targets based on how difficult they are for the learner. The task-based syllabus, on the other hand, would dictate non-linguistic constraints concerning the functional task, within which nearly any grammar topic can be targeted. Macro-adaptivity and task-based syllabi can thus be combined without difficulty.

We focus, however, on a structural syllabus integrated into TSLT, which constitutes the type of curriculum followed by most English textbooks in Germany (Mindt, 2014). In such a syllabus, high-level organizing units also constitute grammatical concepts. This puts certain constraints on potential grammar topics that can be considered for the practice material. Yet if these constraints are taken as a given, the adaptivity module still has sufficient freedom to organize grammar concepts and constructs within these boundaries. The curriculum typically defines `learning targets` on high-level language means that are usually broken down into more fine-grained `realizations` of these `learning targets` by the workbook used in classroom teaching. A `realization of a learning target` constitutes a unit that is necessary to achieve a functional goal of TSLT and cannot be substituted with another unit in order to do so. Multiple `learning targets` are grouped in a learning

unit that defines a pedagogical learning goal and in TSLT indicates the functional target task. Combining curriculum-based instruction and macro-adaptive exercise selection is therefore merely a matter of defining the scope of organizing sovereignty for syllabus and adaptivity. The resulting compromise consists in allowing teachers to assemble learning material at a higher level and allow the adaptivity algorithm to determine the specific material at a lower level. The question now arises whether an ILTS should mirror the workbooks' structures on the level of `realizations of learning targets` in its Graphical User Interface (GUI), or whether it is sufficient to offer an entry point for practice per `learning target` and structure their `realizations` adaptively.

Data In order to answer this question, we analyzed students' navigational patterns followed by learners in the ILTS FeedBook. Overall, four studies have to date been based on this system. For each of the studies, the FeedBook has been modified and extended in some particular aspects. The first project was called **FeedBook – ein interaktives Workbook für den englischen Fremdsprachenunterricht**[1,2]. The original system constituted a digital, web-based version of the Camden Town 3 workbook (Hanus, 2014) with the addition of intelligent feedback specific to a learner's error on an exercise item (Rudzewitz et al., 2017), as illustrated in Figure 2.1. The study of this project assessed the effects of such intelligent feedback. The project lasted from October 2016 to September 2019, with the RCT running from September 2018 to July 2019. The following two projects ran in parallel: **Interact4School (Außerschulisches individuelles Lernen und die Schnittstellen zum Schulunterricht: Effektives digitales Üben als Basis für den kompetenzorientierten Fremdsprachenunterricht**[3])[4] (Parrisius et al., 2022a,b) and **DigBindiff (Digital Differentiation: Integrating linguistic and cognitive measures to foster adaptive learning)**[5,6]. **Interact4School (I4S)** added motivational elements including task-based learning and a student dashboard.

[1]English translation: *An interactive workbook for English foreign language teaching.*

[2]This research was supported by the Deutsche Forschungsgemeinschaft (DFG).

[3]English translation: *Individualized learning outside of class and interfaces to classroom teaching: Effective digital learning as a basis for skills-oriented foreign language teaching.*

[4]This research was supported by the German Federal Ministry of Research and Education (BMBF; 01JD1905A and 01JD1905B) and developed in cooperation with the aim Heilbronn, financially supported by the Dieter Schwarz Foundation.

[5]http://digbindiff.de

[6]This research was supported by the Robert-Bosch-Stiftung.

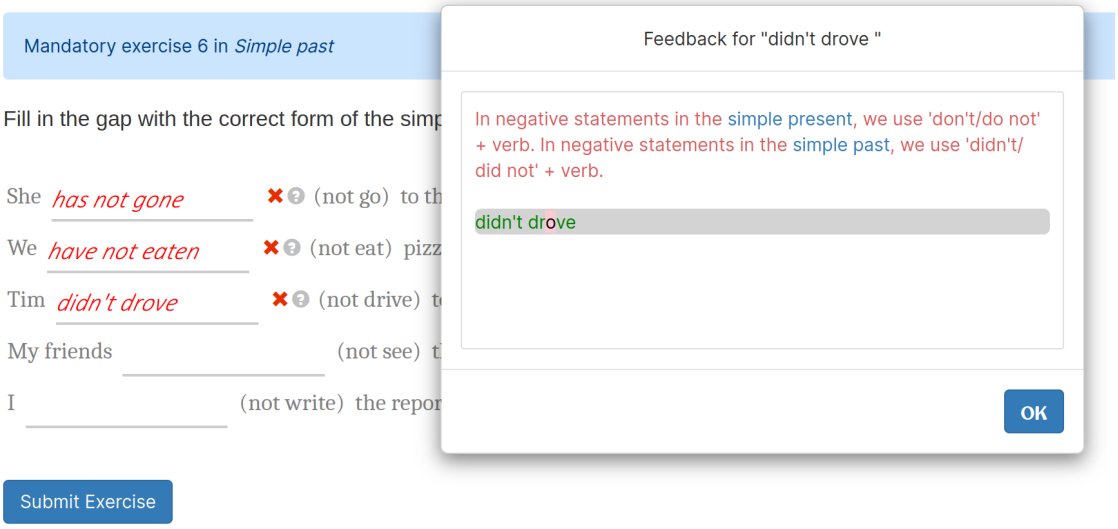


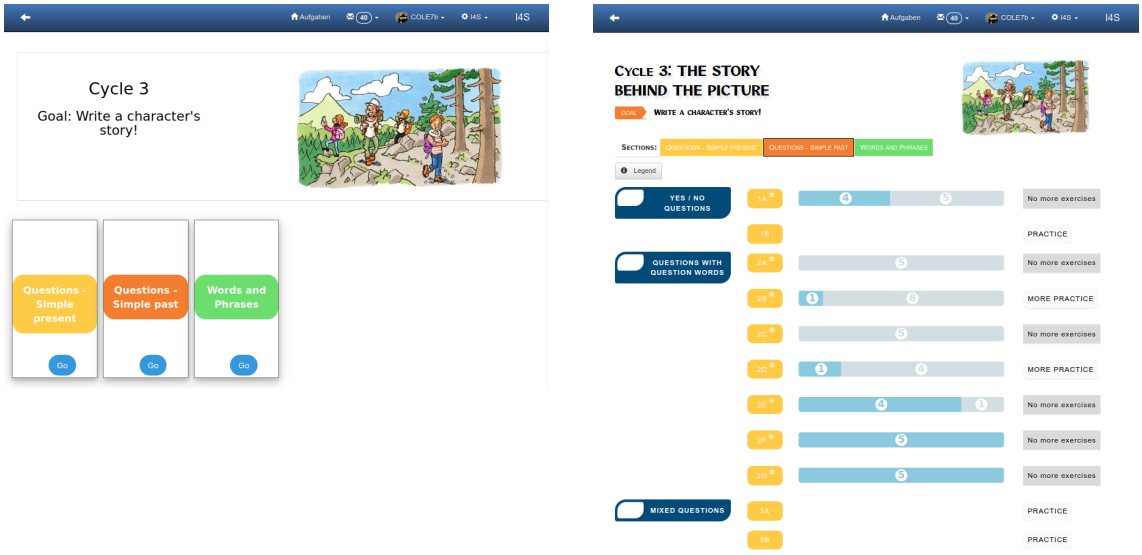
Figure 2.1: Intelligent feedback in the FeedBook.
Feedback is specific to a learner's error on an exercise.

The project lasted from April 2020 to December 2023, with the RCT running from August 2021 to June 2022. **DigBindiff (Didi)** added adaptive exercise sequencing based on current performance on practice exercises and on current working memory capacity assessed each time a learner logged into the system. The project lasted from October 2019 to March 2023, with the RCT running from July 2021 to September 2022. The most recent project is called **AI2Teach (Individual tutoring in an extended digital teaching-learning concept for foreign language classrooms)**[7] (Berens et al., 2024). It focuses on teacher training and a teacher dashboard. The project lasted from September 2020 to August 2026, with the RCT running from February to July 2024.

The data for the following evaluations constitutes a subset of the data collected from 7th grade learners of English in German secondary schools in the I4S study. The primary focus of the I4S project was on motivational aspects in a task-based setting. As illustrated in Figure 2.2a, the exercises in this version of the FeedBook were therefore organized into task-based cycles representing learning units each of which contains multiple linguistically and pedagogically motivated `learning targets`. The system was curriculum based and did not include any macro-adaptivity. Instead, learners

[7]This research was supported by the aim Heilbronn and developed in cooperation with the Zentrum für Schulqualität und Lehrerbildung Baden-Württemberg (ZSL).

had to select specific exercises that were visually organized within `realizations of learning targets` as displayed in Figure 2.2b. We analyzed navigational patterns of the submitted exercises as recorded during the study. Data was available from 619 students who provided written consent to participate in the study. After excluding `learning targets` with only a single `realization` or focusing on vocabulary practice or listening exercises, the dataset comprised 20,586 submissions for core exercises and 7,194 submissions for additional exercises from 593 students in overall 30 `realizations` distributed across 7 `learning targets`. Core exercises are exercises that learners could directly access via the student dashboard. Since the I4S version of the FeedBook did not integrate macro-adaptivity, all core exercises of a `realization` were listed and learners could open them by explicitly clicking on them. Additional exercises became only available once a core exercise had been submitted. All data on exercises and learner submissions was stored in a PostgreSQL[8] database and managed through Hibernate[9]. For the analyses, the data was extracted from the databases with utility scripts written in Java code using the Hibernate setup to access the data. For further processing, the extracted learner



(a) Task-based cycles. (b) Student dashboard.

Figure 2.2: GUI of the FeedBook version used in the I4S study. The content was organized in task-based cycles. Learners had to select the next practice exercise themselves in the student dashboard.

[8]PostgreSQL is available at http://postgresql.org.
[9]Hibernate is available at http://hibernate.org.

submission and exercise data was stored in CSV files. After thus pre-processing the raw database data into a format independent of the ILTS and enriched with meta-information, we implemented the analyses in Python.

In a first step, we looked at whether all students of a class did the same `realizations` irrespective of the order in which they were practiced. We found great variance within the classes. In no class did all students work on the same `realizations` across all `learning targets`. Zooming in on individual `learning targets`, Figure 2.3 illustrates that the percentage of classes within which all learners completed the same set of `realizations` ranged from 0% for *Comparatives/Superlatives* to 15.38% for *Questions in the Simple past*, with varying percentages for the remaining `learning targets`. This suggests that some teachers instruct their students to work on specific `realizations of a learning target` whereas in other classes, students make this decision more autonomously.

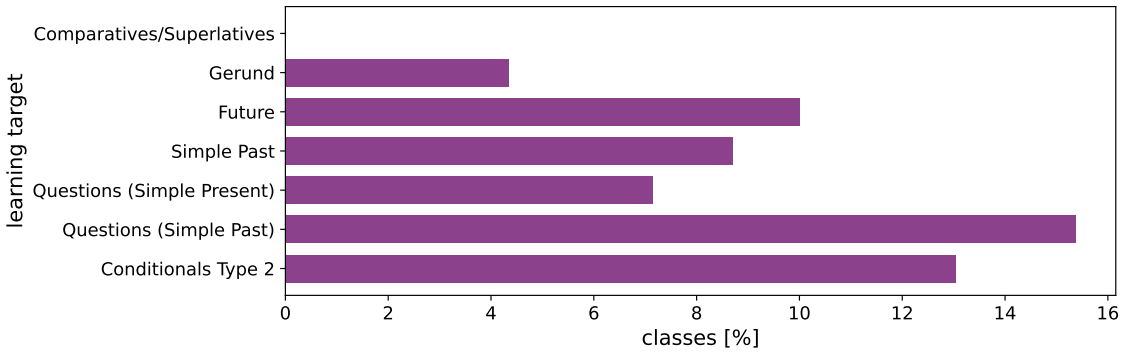


Figure 2.3: Prevalence of homogeneous classes with respect to `realization` practice.
In homogeneous classes, all students practiced the same set of `realizations`. In heterogeneous classes, the practiced `realizations` differed across the students.

In order to determine whether the same holds for the order in which learners work on the `realizations`, we next analyzed whether all learners of a class practiced the `realizations of a learning target` in the same order and, if so, whether this order was the default `realization` order. We determined each learner's `realization` order based on the first submission for each `realization`. Two learners were annotated to follow the same `realization` order if their `realization` sequences were identical, or if one learner practiced more `realizations` than the other but started with the same `realization` and covered the entire sequence of the other learner.

The default order was defined as the order of `realizations` from top to bottom as displayed in the web interface.

Figure 2.4 shows a very diverse picture across `learning targets`. While for *Simple Past* only 13.04% of the classes had identical `realization` sequences for all learners, the learners within 76.92% of all classes followed the same `realization` sequence for *Questions in the Simple Past*. For some `learning targets`, a few classes in which all learners followed the same `realization` sequence used the default order. There are, however, many more examples across all `learning targets` for classes where the commonly used `realization` sequence differs from the default order. It is therefore likely that not only the `realizations` covered by a learner, but also the order of the `realizations` relative to each other is sometimes coupled to the language course a learner follows.

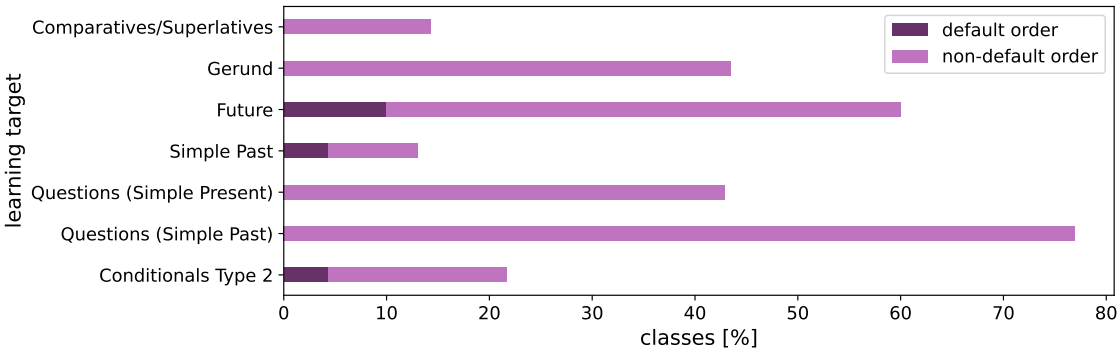


Figure 2.4: Prevalence of homogeneous classes with respect to `realization` sequences.
In homogeneous classes, all students practiced the `realizations` in the same order. In heterogeneous classes, the `realization` sequences differed across students. The color of the graph bar indicates whether the `realization` sequence followed by the learners it represents corresponds to the default order as displayed in the student dashboard from top to bottom.

In order to determine whether adaptive `realization` selection is desirable in those cases where the teacher or workbook does not prescribe a certain order, we looked at how many of the learners followed the default order. As can be seen in Figure 2.5, there is considerable variation across `learning targets`. However, a mean of $\mu = 86.19\%$ ($\sigma = 10.38$) of the students did the `realizations` in the default order. It is rather likely that many of those learners sticking to the default order merely did so out of low learner autonomy (Oxford, 2016). They would thus benefit from having

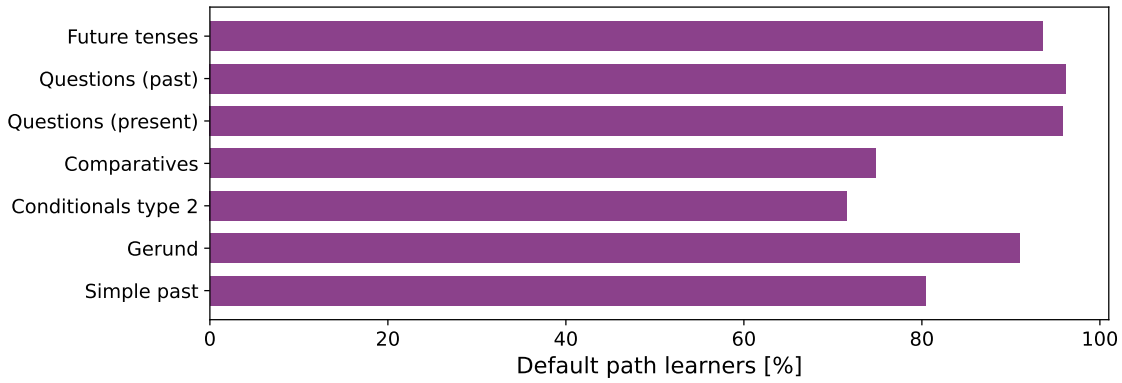


Figure 2.5: Prevalence of default path learners.
Default path learners followed the default `realization` order.

the selection of the `realization` managed by the system in a user-adaptive manner.

Overall, the results indicate that learners tend to practice various subsets of the `realizations of a learning target` offered in the ILTS in a variety of different orders. In some cases, this seems to be motivated by the teacher or the workbook that the learner follows. In other cases, where learners of the same class practice different `realizations` and/or in different orders, learner characteristics such as the amount of learner autonomy are more likely reasons for this diversity. Less autonomous learners might therefore benefit from adaptive content selection across `realizations`, whereas high teacher control requires adaptivity to be limited to exercise selection within a `realization`. The system should thus allow learners and teachers to practice specific `realizations` in a self-determined order, but on the other hand also provide an option for adaptive selection of a `realization` if this autonomy is not required. Dagger et al. (2005) resolve this conflict of interest by allowing teachers to determine the degree of adaptivity for the concepts to practice.

As illustrated in Figure 2.6, the scope of adaptivity thus varies depending on the teacher's configuration. An alternative solution instead permanently offers two entry points for adaptive exercise selection differing in the scope of adaptivity. This solution allows to on the one hand provide as much adaptive sequencing as possible whenever the instructional setting allows it, and to on the other hand empower teachers and students to take on a more active role in the navigation through the practice material.

For an adaptivity algorithm, this results in a two-step content selection where the

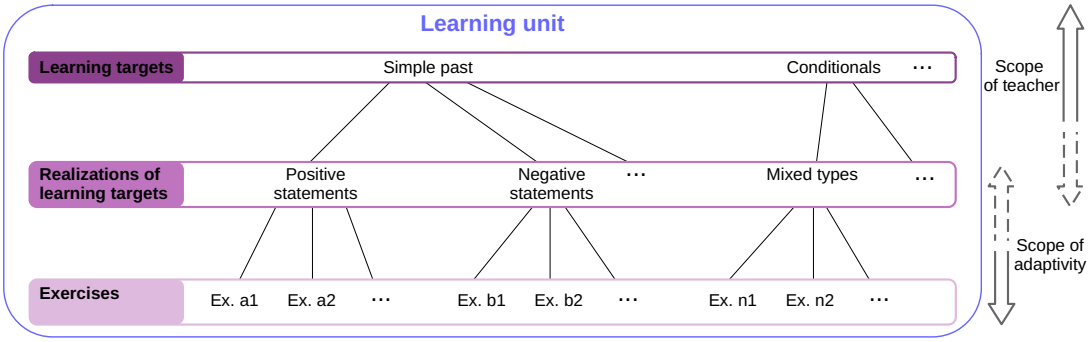


Figure 2.6: Scopes of responsibility for determining practice content.

Teachers are responsible for choosing the `learning target`. The adaptivity module is responsible for selecting a concrete exercise. The `realization` of the teacher-chosen `learning target` for which the adaptively selected exercise must have been designed may be determined by the teacher or by the adaptivity module.

first step may optionally be replaced by a user's navigation. This first step selects a `realization of a learning target`. The second step then selects a specific exercise within the scope of this `realization`.

We have defined the selection of the specific exercise instance to be within the scope of macro-adaptivity, not of teacher control. Addressing teachers' needs to assign specific exercises to their students therefore requires some additional consideration.

Having all learners complete the same exercises directly contradicts the goal of individualized practice through macro-adaptivity. Macro-adaptivity can, however, make sure that a specific exercise is sequenced in at the best possible timepoint, when it best matches the learner model. It is therefore possible to integrate a set of mandatory exercises into an adaptive sequence and increase the likelihood of them being selected by prioritizing them over alternative exercises. Since the adaptive sequence is dynamic, the algorithm can only determine whether an exercise is appropriate at the current point in time. It is not possible to determine where in the dynamically developing sequence an exercise might best fit in. The best compromise therefore consists in sequencing in mandatory exercises if they constitute a good fit at the current timepoint, or else when the learner masters the KC that the mandatory exercise practices if the exercise was not adaptively selected by that time due to poor fit with the learner model.

While such a compromise of adaptive selection of mandatory exercises can make sure that all learners receive all mandatory exercises at some point, that point may only be after a considerable amount of practice for weak learners. In order to ensure that learners with limited available time can attempt the mandatory exercises, these exercises should always be directly accessible.[10] This results in a hybrid approach of adaptive exercise selection whenever possible, and learner-controlled selection if curricular demands clash with individualized practice.

Another essential aspect of learning where TSLT approaches and mastery-based adaptivity might clash concerns the learning goal. Both concepts introduce means to determine whether the learning goal has been reached. In adaptivity-based Mastery Learning, the system maintains records of the learner's knowledge states that allow it to automatically determine mastery of a KC. In TSLT, task-internal readiness can be explicitly assessed through diagnostic exercises. When combining these two approaches, a learner should always be able to complete a diagnostic exercise if the adaptivity algorithm has determined mastery, provided that KCs and diagnostic exercises cover the same linguistic forms. Requiring learners to also pass such a diagnostic exercise makes the achievement of a milestone towards the functional target task more explicit in TSLT. Successfully completing a diagnostic exercise, on the other hand, does not necessarily imply mastery in terms of Mastery Learning. This inconsistency is introduced by the scoring of the diagnostic exercise, which should be tolerant enough to allow for accidental mistakes in diagnostic exercises when assessing readiness (Dušková, 1969). Learners may thus pass the diagnostic exercise even if they fail on individual KCs. A possible solution would be to only allow students to attempt a diagnostic exercise if the adaptivity algorithm has determined mastery on the practice exercises. However, some learners may have previous knowledge and would benefit from being allowed to bypass practice. Diagnostic exercises should therefore always be accessible. Learners with previous knowledge may then not have demonstrated mastery of all KCs of the `realization` when they are assessed as ready for the target task. In order to enable the adaptivity algorithm to correctly model the learner's knowledge state based on the evidence it has collected through learner interactions and at the same time make readiness salient to the learner, visualizations of progress in the GUI may therefore deviate from learner

[10] I would like to thank Leona Colling for introducing the requirement of making certain exercises accessible at all times and defining their visualization in the GUI.

model states.[11]

2.2 Integrating Interleaved Practice into Revision Learning Units

Workbooks used in schools typically integrate revision into a learning unit for one of two reasons (cf. Jones et al., 2020; Hanus et al., 2020; Harger and Niemitz-Rossant, 2014; Derkow-Disselbeck et al., 2008). (1) In order to facilitate the introduction of a new grammar topic, prerequisite grammar topics are revised shortly beforehand. (2) Grammar topics newly introduced in a school year are often revised again later in that year.

Many adaptive systems include the second type of revision exercises. However, they typically interleave revision exercises with exercises targeting new linguistic forms (Nagatani et al., 2019). A need for revision is detected when a learner's probability of correctly answering an exercise of a previously mastered KC is below the mastery threshold (Doroudi, 2020). This probability can be reduced either through misconceptions demonstrated in the form of negative evidence (Plass and Pawar, 2020), or through forgetting. Managing misconceptions requires a mapping of learner errors to linguistic forms for which the errors indicate negative evidence. Models of forgetting indicate when previously mastered KCs should be revised or, in the case of not yet mastered KCs, slow down progress towards mastery (Nagatani et al., 2019).

In order to determine forgetting, ILTSs have applied various approaches. Some Knowledge Tracing approaches such as BKT can directly integrate the concept of forgetting into the calculations of the learners' knowledge states (Pelánek and Řihák, 2017). Other approaches model forgetting separately (Zeng and Lin, 2011). Since Ebbinghaus (1885) first investigated forgetting curves based on nonsense syllables, researchers have applied this formula to a variety of material (Finkenbinder, 1913). Ebbinghaus (1913) approximated the form of the forgetting curve with the formula given in Equation 2.1, where he determines the percentage b of time savings for re-mastering an item relative to the time needed to first master it:

[11] I would like to thank Leona Colling for providing the learner-centered perspective and the visualization of mastery in the GUI.

$$b = \frac{100k}{(\log t)^c}, \text{ where } c = 1.25, k = 1.84 \tag{2.1}$$

In this formula, t indicates the time in minutes since the item was last practiced; the constants c and k were defined to fit Ebbinghaus's (1913) dataset. A value of b = 20% would indicate that the learner needs 20% less time to master the item than they needed when first learning it.

For vocabulary practice, incorporating forgetting has become a standard feature of ILTSs (Zhu, 2020). Zaidi et al. (2020) found that taking a word's complexity into account improves the prediction of forgetting. According to the experiments conducted by Scherer (1957), considering forgetting is even more essential for grammar practice. Based on tests assessing German language skills of native English college students before and after the summer break, his evaluations revealed higher loss of grammar knowledge than of vocabulary knowledge. Hildebrand and Scheibner-Herzig (1986) analyzed forgetting rates for passive, future, and gerund structures and found that they differed considerably between the structures. They suggest, in line with Zaidi et al.'s (2020) findings for vocabulary learning, that the forgetting rate is correlated with the structure's difficulty. De Groot and Keijzer (2000) recommend to practice the harder to learn – and more easily forgotten – non-cognates and abstract words more often than cognates and concrete words. In a real-world setting, Ridgeway et al. (2017) found that modeling forgetting for vocabulary, grammar and pronunciation was most predictive when applying a combination of power law and linear regression to the learner's study history.

While the use of forgetting curves does not clash with any practices concerning revision in instructional settings, the placement of revision exercises interleaving introductory exercises as opposed to in revision learning units does. Focusing on RQ 2, we investigate how these two approaches can be reconciled. To this purpose, we first zoom in on whether and how practice for initial mastery should and can be separated from revision practice. In a second step, we evaluate how revision targeting forgetting on the one hand, and revision targeting misconceptions on the other hand, fit in. We analyze whether these two motivators for revision exercises can be treated similarly, or whether they require different exercise types for best learning outcomes.

2.2.1 Disentangling Revision Practice from Mastery Learning

Revision of newly introduced `learning targets` involves `learning targets` of the same school year. The learner likely has already worked on that `learning target` in the ILTS before, so that learner data is available. For revision of prerequisite `learning targets`, on the other hand, learner data might not be available in the ILTS.

We therefore aim to investigate whether revision `learning targets` should be treated differently than introductory `learning targets`, and whether they should be treated differently depending on the motivation for including them. If learners submit only few incorrect answers in revision units but make more errors in introductory units, revision units require more difficult exercises to start with. If, on the other hand, the adaptivity algorithm picks up previous knowledge quickly enough, revision units do not require special treatment. In order to compare prerequisite revision units and introductory learning units, we evaluated learner submissions for revision units whose `learning targets` had already been introduced in the previous school year, and for learning units introducing new content.

Data The data for these evaluations was collected in the context of the AI2Teach project[12] between February and July 2024. The large-scale RCT of this project used a FeedBook version that combines all features of its previous versions. These include the original FeedBook and the I4S and Didi versions presented in Section 2.1.3. Extensive system development improved and extended the FeedBook in the course of this project. This system version re-uses parts of the old, monolithic system in a micro-service architecture as shown in Figure 2.7. The docker-based micro-services, each maintaining its own database, communicate via Hypertext Transfer Protocol (HTTP) requests with each other. A micro-service may keep a cache of aggregated data from another micro-service for performance reasons as this may reduce network traffic and/or avoid having to re-calculate aggregated metrics. All communication with the front-end and between the micro-services is authenticated. The database of the **materials micro-service** hosts all static in-

[12]https://www.iwm-tuebingen.de/www/en/forschung/projekte/projekt.html?name=AI2Teach

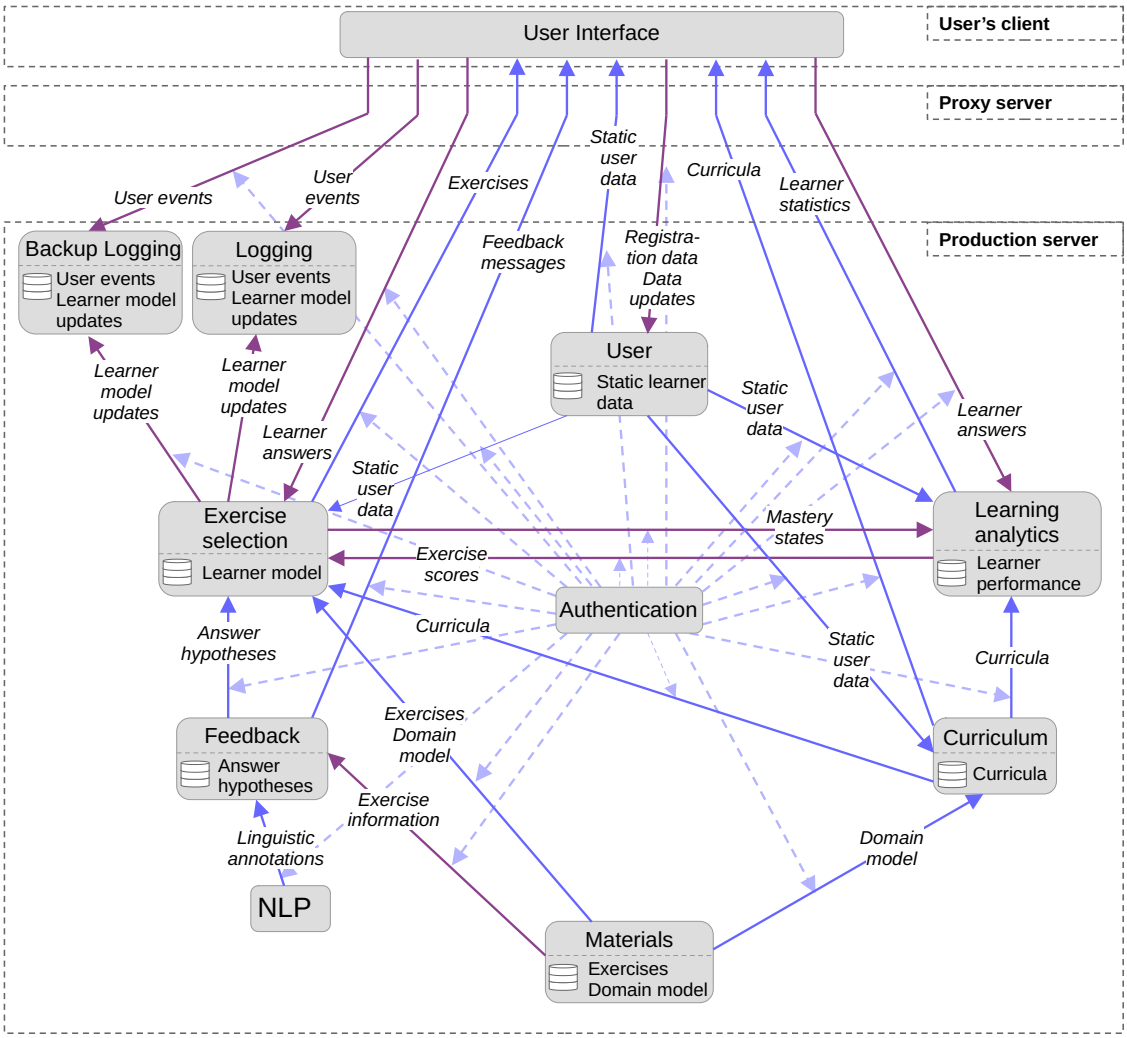


Figure 2.7: System landscape of the AI2Teach FeedBook.
The front-end user interface is connected to the back-end production server through a proxy server. The back-end micro-services communicate with each other as well as with the front-end. The arrows indicate the direction of the information flow. Blue arrows indicate that the information flow was requested by the recipient so that the direction of the information flow is inverse to that of the control flow. Purple arrows indicate that the information flow was initiated by the sender so that the direction of the information flow is the same as that of the control flow.

formation concerning the domain model and the exercises. While it can be updated in order to correct data, the content does in principle not change after the initial deployment and is entirely independent of user interaction and user data. If an exercise is corrected, the old version is kept in an archive with a reference to its new version so that the old version may be retrieved for analysis reasons as well as to be

displayed in the system if required. Exercises are generated semi-automatically in the **exercise generation** project based on manually created exercise specifications. The generated exercises are enriched with linguistic annotations based on the POLKE micro-service and with references to feedback hypotheses based on the feedback micro-service. The **POLKE micro-service** annotates all constructs of the EGP for a given input text. In addition, feedback is generated for all half-open exercises in the **feedback micro-service**. It generates answer hypotheses for actionable elements of exercises offline and diagnoses the most suitable answer hypothesis and corresponding feedback for a learner's answer based on Lucene[13] lookups online. All NLP used by the feedback generation is contained in the **NLP micro-service**. This micro-service is currently only used for feedback generation. Domain model elements are assembled into class-specific exercises in the database of the **curriculum micro-service**. Each curriculum references components of the domain model and adds additional information such as the assembly of domain model components into a learning unit. The database of the **learning analytics micro-service** hosts aggregated learner metrics required for the student and teacher dashboards. In addition, it scores the students' submissions based on the exercise type and learner answer. The **exercise selection micro-service** receives all requests for exercises from the front-end and, depending on the request, selects an exercise in a user-adaptive manner, within a pre-defined sequence, or with a specific ID. The database hosts the learner model used for adaptive sequencing. The database of the **user micro-service** hosts all static, user-specific data that does not usually change after the user has been registered in the system. It also manages authentication and deauthentication of a user's session. Authentication itself is handled in the **authentication micro-service**. It uses a Redis[14] distribution without further customization. Scientific logging in the **logging micro-service** is based on xAPI[15] with custom verbs and actions defined for the logging statements. A backup logging micro-service stores all logging statements as strings in an additional database. The front-end **user interface** is based on React[16]. The exercise and student dashboard structure were ported from the old Google Web

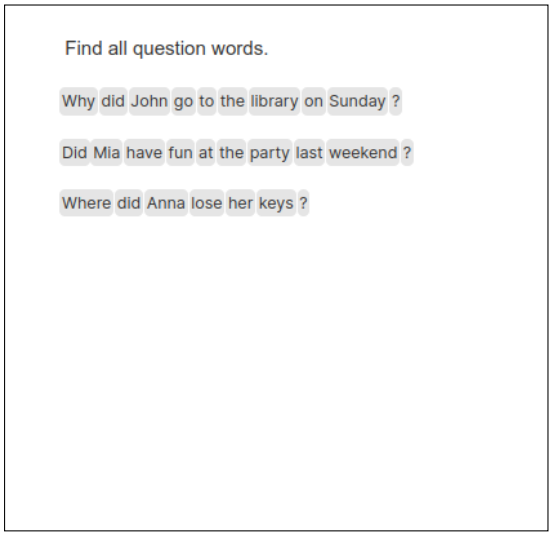
[13]Lucene is available at https://lucene.apache.org/.
[14]Redis is available at https://redis.io/de/.
[15]xAPI is available a https://xapi.com/?utm_source=google&utm_medium=natural_search.
[16]React is available at https://react.dev/.

Toolkit[17] implementation, with additional event logic and modularization added. The teacher dashboard is implemented in a separate project and integrated into the main front-end code as a distribution file.
With a focus on teacher training and a newly added teacher dashboard, the AI2Teach study involved 52 classes with overall 1,191 participating students providing written consent and who submitted at least one exercise. Participating students were from two different school types. *Gymnasium* constitutes an academic-track secondary school whose degree qualifies to go to college. *Gemeinschaftsschule* constitutes a comprehensive secondary school that does not apply achievement-based entry restrictions and qualifies for professional training after graduation. While 1,125 participants were 7^{th} grade Gymnasium students, 66 participants from 3 classes were 8^{th} grade Gemeinschaftsschule students. 733 Gymnasium students and 50 Gemeinschaftsschule students participated both in the pre- and in the post-test. All pre-tests were administered at the beginning of the study. Post-tests were administered within two weeks at the end of the school year. Gemeinschaftsschule students were only included in analyses looking explicitly at this school type.
Similar to the preceding FeedBook studies, the `learning targets`, distributed across two learning units, focused on the 7^{th} grade Gymnasium curriculum. The *Practicing Tenses* unit constitutes a revision learning unit and comprises the `learning targets` *Simple past*, *Present perfect*, and *Simple past vs. present perfect*. The unit *The School Trip* covers the revision `learning target` *Questions* and the introductory `learning target` *Conditionals*. It thus constitutes a mixed learning unit that contains `learning targets` revising content from previous school years as well as `learning targets` that introduce content that the students had not already acquired previously.

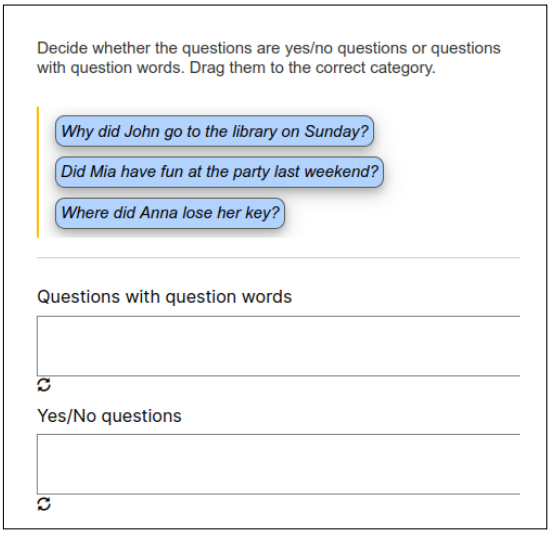
The exercise types supported in the FeedBook emerged in the Didi project with the intention of supporting different exercise complexity levels (Xiong et al., 2024). They therefore encompass four closed exercise types that rely on interactions such as drag and drop or mouse clicks, as well as two half-open exercise types that require typing-based input. The six exercise types, presented in Figure 2.8, were slightly adjusted for the AI2Teach study.

Mark-the-Words exercises are closed exercises. Learners have to identify sentence

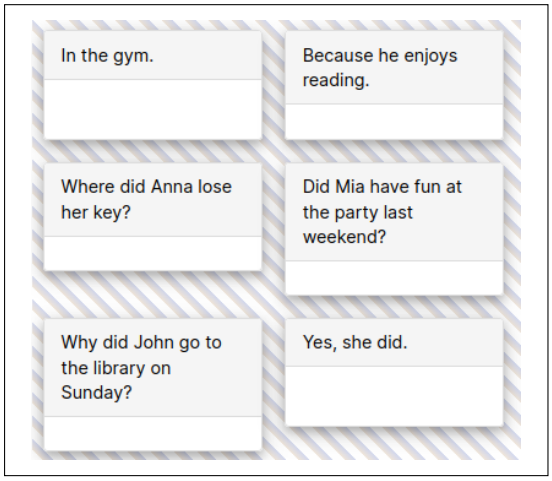
[17]GWT is available at https://www.gwtproject.org/.



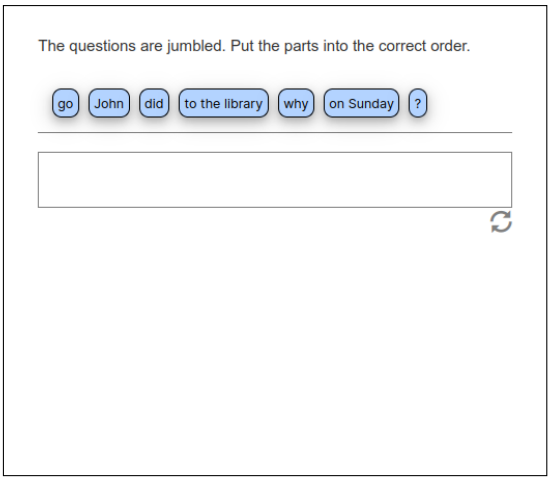
(a) Mark-the-Words exercise.



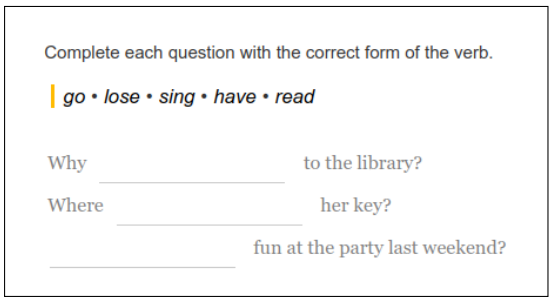
(b) Categorization exercise.



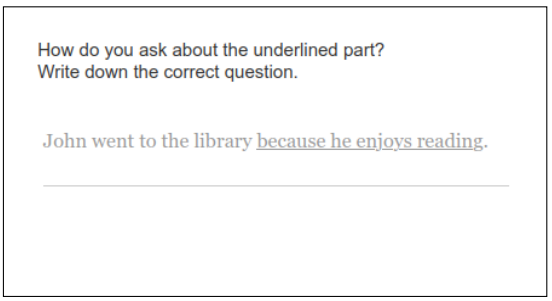
(c) Mapping exercise.



(d) Jumbled Sentence exercise.



(e) Gap-Filling exercise.



(f) Short Answer exercise.

Figure 2.8: Exercise types in AI2Teach.
The AI2Teach study supported 4 closed (2.8a–2.8c) and 2 half-open (2.8e–2.8f) exercise types.

constituents with specific meta-linguistic properties. They can select a constituent by clicking on it. **Categorization** exercises also fall into the category of closed exercises. They require learners to sort elements into meta-linguistic categories. Learners can do so using drag and drop functions. **Mapping** exercises also constitute a closed exercise type. Learners need to find pairs of elements that belong together. They can map cards by successively clicking on two cards. **Jumbled Sentence** exercises conclude the set of closed exercise types. In order to solve them, learners need to put the provided sentence constituents into the correct order. They can order the sentence chunks by clicking on them in correct succession. Clicking on a chunk moves it to the end of the chunk sequence in the drop area.
Gap-filling exercises are half-open exercises. They contain blanks within sentences that learners need to fill. **Short Answer** exercises constitute the second half-open exercise type. They each elicit an entire sentence as answer to a prompt.

In order to track mastery of KCs, the system relied on rule-based approaches. Mastery was determined based on the accuracy for a KC over the most recently submitted 5 elements (or less, if fewer were submitted) for that KC. If the accuracy was higher than a certain threshold, the learner was considered to have mastered the KC. The threshold depended on the exercise type. An actionable element constitutes a part of an exercise for which the learner needs to provide an answer and may receive a scoring point, for example a gap in a Gap-filling exercise. Considering that half-open exercises are more prone to typos, the accuracy threshold for Short Answer exercises was set to 60%. For Gap-filling – where input is short – and Jumbled Sentence exercises – a closed exercise type where each scored actionable element requires a number of user interactions –, the threshold was set to 80%. For the remaining – purely interactive – exercise types, where accidental mistakes are less likely, an accuracy of 100% was expected for mastery.

Note

The system could not employ any data-driven approaches to track mastery since no learner data was available before the start of the RCT. For lack of reference data in the literature, the mastery thresholds were set based on intuition. While our approach is subject to the criticism of arbitrariness in defining mastery criteria, it does not require learner data for training purposes.

An analysis of learner submissions for the two learning units *Practicing Tenses* and *The School Trip* revealed that accuracy was higher in the revision unit *Practicing Tenses* ($\mu = 89.31\%$, $\sigma = 14.46$) than in both the revision `learning target` *Questions* ($\mu = 85.43\%$, $\sigma = 15.84$) and the introductory `learning target` *Conditionals* ($\mu = 80.90\%$, $\sigma = 21.34$) of the learning unit *The School Trip*. We tested for statistical significance with a two-tailed t-test, applying the commonly used threshold of $p<.05$ for statistical significance and $p<.1$ for marginal significance. As can be seen in Table 2.3, the difference between the `learning targets` of all three practice states (`learning targets` of the revision learning unit, revision `learning targets` of the mixed learning unit, and introductory `learning targets` of the mixed learning unit) is highly significant according to this test.

`Learning targets`	**Accuracy**		**n exercises**	
	t-value	p-value	t-value	p-value
Revision_{Tenses} – $\text{revision}_{Questions}$	5.7303	.0000	9.5334	.0000
Revision_{Tenses} – $\text{introduction}_{Conditionals}$	10.4854	.0000	5.8080	.0000
$\text{Revision}_{Questions}$ – $\text{introduction}_{Conditionals}$	-4.1389	.0000	15.7138	.0000

Table 2.3: Performance on practice exercises across revision and introductory `learning targets`.
Accuracy on practice exercises and the number of exercises required to reach mastery per KC of the `realization` differed significantly between introductory and revision `learning targets`.

The number of exercises required to reach mastery per KC of the `realization` also differed significantly between the three practice states (see Table 2.3). Learners needed the largest number of exercises in the the introductory `learning target` ($\mu = 3.4128$, $\sigma = 2.5609$) and the lowest number of exercises in the revision `learning target` ($\mu = 1.2040$, $\sigma = .6375$) of the mixed learning unit. Interestingly, the revision learning unit ranges in-between these two `learning targets` ($\mu = 2.0317$, $\sigma = 1.2357$). The revision learning unit also served to familiarize learners with the ILTS, which could explain why learners needed more exercises to reach mastery of a KC in these `learning targets` than in the revision `learning target` of the mixed learning unit.

The suitability of the mastery criteria applied in our implementation is supported when considering Figure 2.9. The box plots compare performance on diagnostic exercises between `realizations` for which the learner had reached mastery and

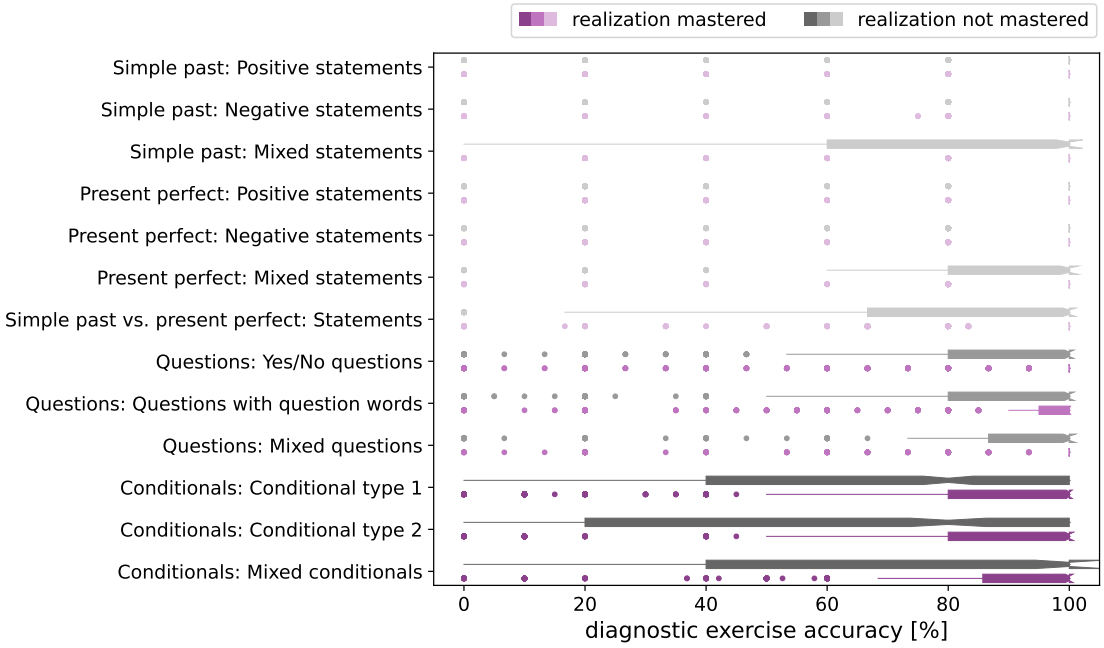


Figure 2.9: Accuracy on diagnostic exercises depending on automatically diagnosed mastery.
Learners performed better on diagnostic exercises if they had mastered the `realization` according to the adaptivity algorithm in introduction `learning targets` (dark color). In revision `learning targets` of the revision unit (medium light color) as well as in the mixed learning unit (light color), the effect is moderated by a ceiling effect with respect to learners' performance on diagnostic exercises.

those for which they had not. Performance on diagnostic exercises was generally better if the `realization` had been mastered. This is quite pronounced in newly introduced `learning targets`, whereas there is a ceiling effect in revision `learning targets`. This ceiling effect is particularly evident in the revision learning unit on tenses, but also noticeable in the *Questions* `learning target` of the mixed learning unit *The School Trip*.

In order to be able to treat learning units differently depending on whether they introduce or revise `learning targets`, adaptive exercise selection must be provided meta-information about the learning unit's purpose as revision or introductory unit. How exactly macro-adaptivity should adjust exercise selection for revision units remains an open research question. Options include higher initial mastery estimates, and greater increase of mastery estimates per positive evidence.

While revision of content taught in previous school years is typically integrated as individual learning units, this is rarely the case for content introduced earlier in the same school year. Yet, this knowledge fades over time as well. Introductory practice until mastery and revision practice may draw on the same pool of exercises, yet they differ in their learning goals. For introductory practice, the goal is to introduce and for the learner to master a topic. For revision practice, the goal is to target forgetting on the one hand, and misconceptions on the other hand. Forgetting is an important concept, but in a learning unit based system, integrating refresher exercises is not trivial. Providing them as additional practice material in the `learning target` where the knowledge was first acquired poses two problems. First, a `learning target` may be included in multiple learning units, so that an additional design decision is required to determine which one of them, or all, should host refresher exercises. Second, if students never re-enter already completed learning units, they will not practice the refresher exercises. Interleaving exercises in the `learning targets` of the current learning unit with refresher exercises for other `learning targets` also has considerable downsides. Weak students, in particular, already struggle to complete the exercises assigned to the `learning target` in focus. In addition having to complete refresher exercises might introduce too much variability in content, thus overloading and confusing the learners. For remedial exercises targeting errors a learner has made previously, these issues are even more pronounced as the learner has struggled with the practiced linguistic form before. While a learner still works on a `learning target` it is possible to prioritize remedial exercises created for the KC currently in focus. Note, however, that such prioritizations considerably restrict macro-adaptivity. Linguistic variability across thus selected exercises, in particular, is expected to be lower since remedial exercises cover previously practiced linguistic forms instead of introducing new variations of a KC.

These usability based reasons suggest that introduction of a `learning target` is best separated from revision practice. This could be implemented either as a separate revision unit for each `learning target` or as a single revision unit interleaving practice of the different `learning targets`. While common workbooks adopt the former approach, macro-adaptivity postulates the latter. We approach this conflict of interests from a user-centered perspective relying on empirical data. Our goal is to determine whether learners in school settings tend to practice exercises across

various `learning targets` in an interleaved manner for revision practice or rather to revise the `learning targets` sequentially. To this purpose, we analyzed the data collected in the I4S study described in Section 2.1.3. Focusing only on the additional practice exercises, we analyzed whether learners tended to navigate between `realizations` of a `learning target`, and also between different `learning targets` for revision practice.

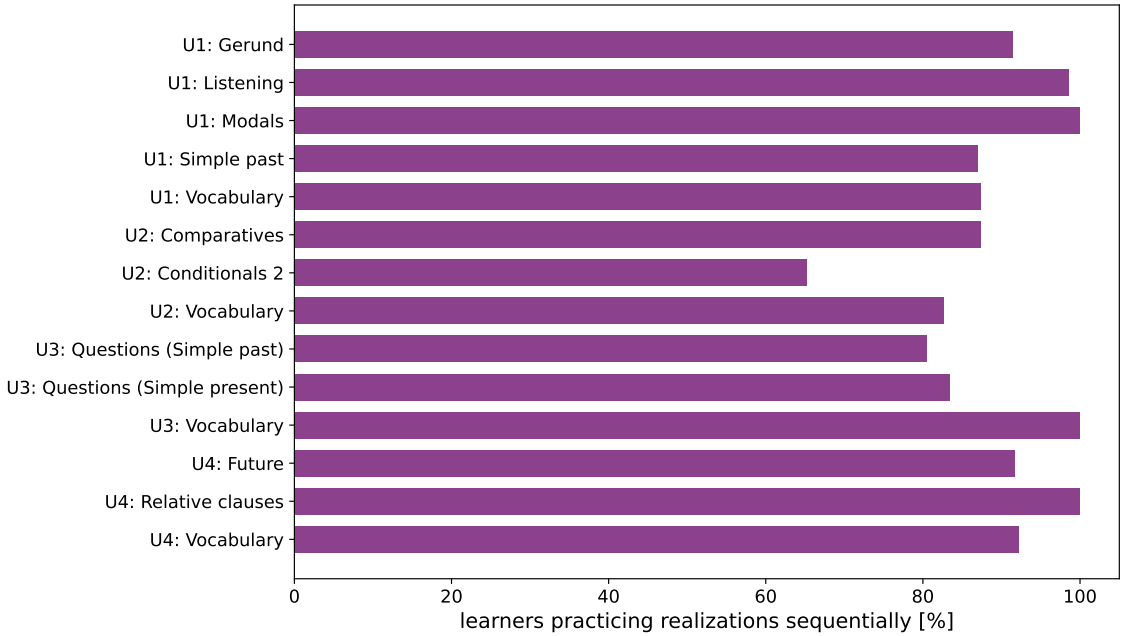


Figure 2.10: Prevalence of sequential learners.
Most learners practiced one `realization` after the other in additional practice.

54.30% of all learners completed all additional exercises of each `realization` in a row. As illustrated in Figure 2.10, the numbers are even higher when looking at exercise sequences within each `learning target`: Averaging over all `learning targets`, a mean of $\mu = 89.11\%$ ($\sigma = 9.59$) of the learners practiced the `realizations` successively per `learning target`. They did, however, navigate considerably between the `learning targets`. Figure 2.11 illustrates that this usually involved no more than two `learning targets` at a time. In addition, after switching to another `learning target`, learners usually practiced several exercises within that `learning target`. They thus tended to not practice different grammatical `learning targets` in an interleaved manner.

A possible interpretation of the results suggests that the learners participating in

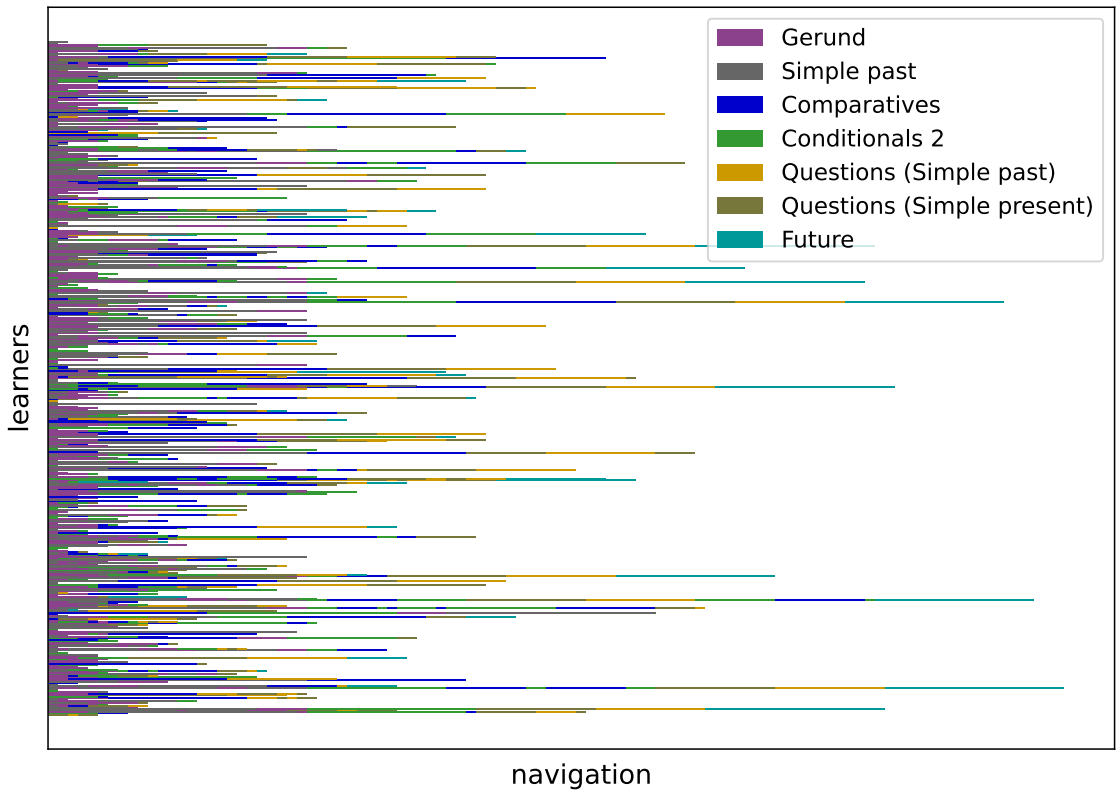


Figure 2.11: Navigational patterns for additional practice.
Learners did not frequently navigate between more than two `learning targets` at a time and stayed within each `learning target` for several exercises.

the I4S study preferred to not interleave multiple `learning targets` in revision practice. An alternative interpretation, however, assumes that learners found it too cumbersome to navigate between `learning targets`. Since the system did not offer easy access to interleaved practice, we further investigated whether a global revision unit would be accepted if offered. To this purpose, we combined revision for all `learning targets` in a single revision learning unit called Fitness Center that we included in the FeedBook version of the AI2Teach study. Contrary to other learning units, the Fitness Center specifically targets remedial and refresher exercises. It selects single exercise items from exercises that are similar to exercise items that learners answered incorrectly, or items that practice a KC that needs refreshing, depending on which of the two is targeted.

In order to select revision exercises for either learning target, the algorithm relies on a learner model. Lukas (1997) refers to the learner model data used with refresher

exercises as knowledge state and to the data used with remedial exercises as misconceptions. While his approach attempts to integrate both types of information into a single model, an approach to adaptivity only benefits from such integration efforts if it makes use of both kinds of information in the same exercise selection process. Since this is not the case in our implementation, we maintain the two types of information in separate models: an overlay model for knowledge states, and a bug model for misconceptions.

Refresher exercises targeting forgetting are based on KCs that the learner has previously mastered and not practiced for some time. Exercise selection relies on the overlay model to track and provide this information. Whether mastery has expired is re-calculated on each learner login based on the time elapsed since the learner has last practiced a KC. Since no forgetting curves are available for grammar knowledge and our learner population, the expiration time span is initially set to 7 days in the AI2Teach FeedBook. It is gradually adjusted as the learner works with the system. If the learner refreshes a KC for which forgetting has been diagnosed and gets more than 80% correct, the time span is increased to double its duration. If accuracy is below 80%, the time span is reduced to half its duration. If multiple KCs have expired, all exercise items practicing any of these KCs are considered.

Remedial exercises are selected based on misconceptions stored in the bug model. This model maintains misconceptions in the form of pairs of expected and provided linguistic constructs where the two constructs are not identical. For exercise selection, the misconceptions are ranked by the relative frequency of negative over positive evidence. A second resource required for exercise selection consists in a mapping of exercise items to potential misconceptions. In the FeedBook, the module providing intelligent feedback generates and maintains such mappings (Rudzewitz et al., 2018). This module anticipates the most common correct and incorrect learner answers, henceforth referred to as answer hypotheses, and matches them to error diagnoses (Meurers, 2012). It automatically generates these answer hypotheses for each actionable element when a new exercise is created. Each answer hypothesis is associated with a misconception. The exercise selection algorithm queries the feedback database for answer hypotheses matching the most frequent misconception stored in the bug model, and retrieves the associated exercise items. If the bug model has not yet been populated due to low interaction with the system, the al-

gorithm determines the KC for which the learner has made the most errors, if any. The algorithm then selects exercise items practicing that KC.

If multiple items are found, one is selected at random. If no exercises are found, the Fitness Center does not provide any.

The evaluations of whether a global revision learning unit is accepted by and beneficial for learners are based on the dataset from the AI2Teach study. In addition to the Fitness Center, our implementation offered the option to continue practicing inside a learning unit after reaching mastery. We compared learners' preferences of using the Fitness Center over continued practice within a specific learning unit in terms of frequency of usage. Frequency of usage was defined as the number of submitted exercises either in the Fitness Center or in other learning units if the respective `realization of the learning target` had already been mastered. Mastery was defined as having passed a diagnostic exercise.

According to the results, the Fitness Center was generally poorly frequented, with only 203 out of the 1,191 study participants submitting any exercises in this learning unit. The number of exercise submissions amounted to 2,566. The alternative of continued practice within a learning unit showed slightly higher usage with 284 learners and overall 14,728 submissions. Acceptance of the global revision unit was thus poor. Most learners preferred revision of a specific `learning target` inside the introductory learning unit.

The most relevant outcome in the context of revision are long-term learning gains. Since the post-test was administered at the end of the school year in our study, most participants had passed the diagnostic exercises of the learning units several weeks before the post-test. Figure 2.12 illustrates that this is particularly true for the revision learning unit on tenses. The `learning targets` of this learning unit were thus practiced recently only in revision exercises. This study setup makes it possible to observe potential effects of revision practice on learning gains between the pre-test and the post-test.

In order to determine the potential benefits of a global revision unit over continued practice within an introductory unit, we evaluated whether learning gains increased with increased use of the Fitness Center relative to more practice within a learning unit. This analysis contrasts learners using predominantly the Fitness Center against

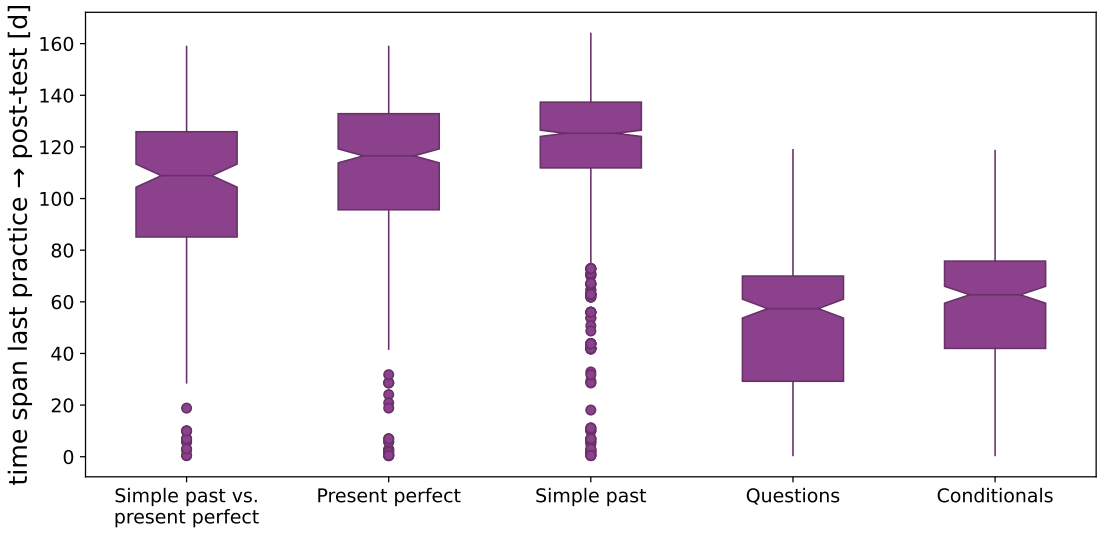


Figure 2.12: Time span between mastering a `realization` and post-test. Most learners passed the diagnostic exercises of the revision learning unit 80 to 140 days before the post-test. For the second learning unit, the elapsed time ranges between 30 and 80 days.

those predominantly continuing practice within the learning unit. Learning gains were operationalized as improvement between pre- and post-test. Significance of the effect was determined through mixed effects analyses controlling for the school class and the number of submitted revision exercises. In addition, we report effect sizes as Hedge's g, where small effect sizes range from .2 to .5, medium effect sizes from .5 to .8, and large effect sizes from .8 to 1.0.

As can be seen in Figure 2.13, continued practice within a learning unit was slightly favorable relative to using the Fitness Center with respect to learning gains. The effect was, however, not significant (t=.483, p=.629; g=.0691). Pearson's correlation between the ratio of continued practice in a learning unit and practice in the Fitness Center with learning gains is also very low ($\rho = .0369$).

The overall low usage of the Fitness Center confirms our initial assumption that a general revision learning unit is not well applicable in a school based setting, although our findings about its effectiveness are inconclusive. In order to support teachers in revising specific `learning targets` throughout the school year, a more practicable approach thus needs to offer configurable revision learning units where teachers can activate their own choice of `learning targets` and `realizations`

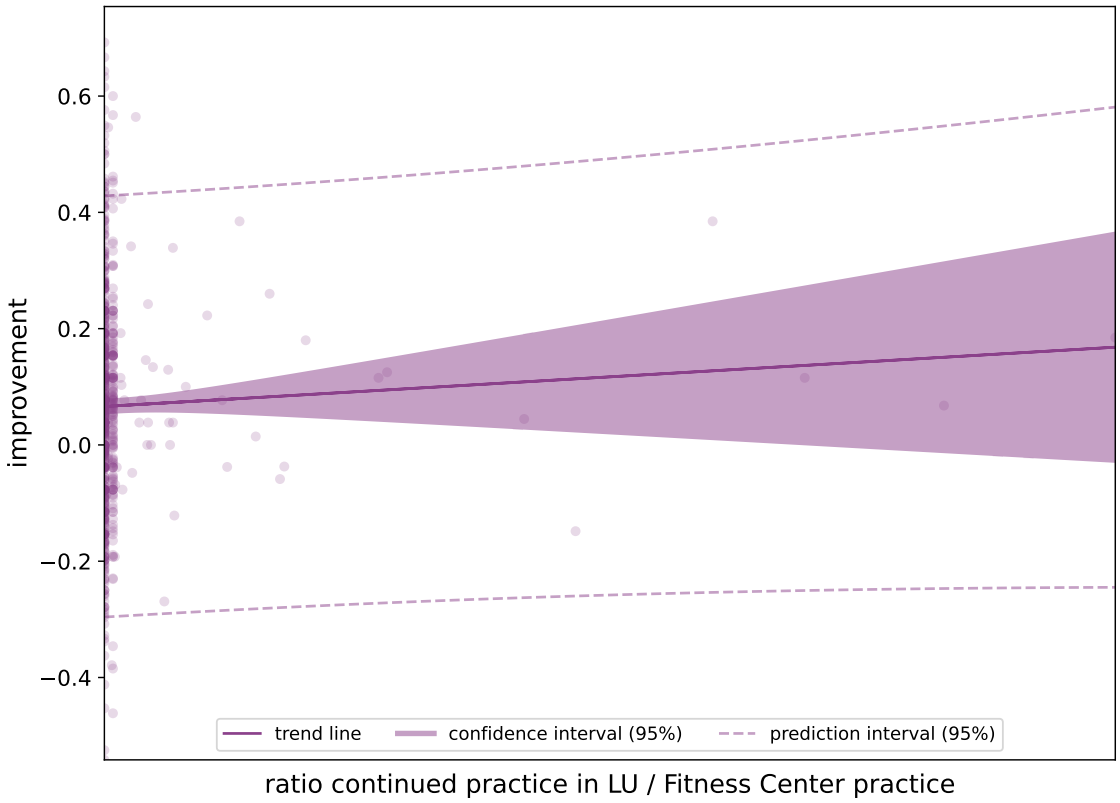


Figure 2.13: Learning gains depending on Fitness Center usage.
Learning gains between pre- and post-test were slightly higher for learners using continued practice within a learning unit (LU) rather than the Fitness Center, but the effect is not significant.

thereof. Within these revision `learning targets`, exercises can be adaptively selected to focus, for instance, on remedial practice targeting a learner's misconceptions.

2.2.2 Different Exercise Types for Different Aspects of Revision

Supporting language learners to progress in their ZPD requires exercises at different complexity levels (Shabani et al., 2010). While input enhancement for implicit exposure to linguistic forms through visual highlighting can foster receptive skills at the lower end of the complexity range (Meurers et al., 2010), open exercises that elicit production of entire sentences constitute the other extreme (Becker and

Roos, 2016). Closed exercises allow learners to advance from one extreme to the other (Melvin and Stout, 1987) by acquiring the constructions relevant for language production in a controlled way.

When revising a grammar topic, however, learners have already been introduced to and have potentially mastered the targeted linguistic forms. Revision may be relevant for one of two reasons: (1) refreshing forgotten knowledge or (2) remedying misconceptions. Refresher exercises require active sequencing that aims to achieve a learning goal in terms of re-mastery of the KC. Remedial exercises require passive sequencing in reaction to an observed misconception (Brusilovsky, 1999). While it often makes sense to address both aspects of revision jointly, it is rather likely that the linguistic forms they practice differ in the learner's mastery thereof. The degree of mastery, in turn, influences the exercise complexity suitable for a learner.

We therefore hypothesized that refresher exercises can be more difficult as they target an already mastered KC whereas remedial exercises target linguistic forms with which the learner has had trouble in the past. According to this hypothesis, optimal remedial exercises are of closed exercise types. Optimal refresher exercises are of half-open exercise types, which require production of the target forms in entire sentences, but restrict the space of possible answers in such a way that all correct learner answers can be anticipated. In order to test this hypothesis, in the AI2Teach study we introduced two study conditions for the Fitness Center revision unit presented in Section 2.2.1. The two conditions differed in the openness of the exercises the learners received in the Fitness Center. The control group received exercises of arbitrary types. The intervention group received half-open or closed exercises aligned with the learning target of remedial or refresher exercises, as proposed above. Only in the case that the bug model had not yet been populated, openness was not aligned with the learning target. In the fallback scenario of selecting exercises that were similar to submitted, incorrectly answered exercises, exercises of all types were considered in order to not further restrict this fallback exercise selection.

Note

In order to focus on the effect of aligning exercise openness with the revision learning target, we did not introduce further selection mechanisms. However, the selection of revision exercises could also integrate all macro-adaptive exercise selection criteria of active exercise sequencing that are applied in other learning units.

Data The evaluations of whether learners benefit from such an alignment are based on the dataset described in Section 2.2.1. 549 Gymnasium students were in the intervention condition and 576 in the control condition. Of the Gemeinschaftsschule students, 30 students were in the intervention condition and 36 in the control condition.

In order to determine whether adjusting exercise types to the type of revision affected the acceptability of the Fitness Center, we first compared the two study conditions with respect to learners' preferences for using the Fitness Center over continued practice within a specific learning unit. The results show a more or less equal distribution of the 203 students using the Fitness Center across the study conditions. However, the students in the intervention condition used the Fitness Center slightly more frequently than those in the control condition. 100 students were in the control group and submitted overall 1,165 exercises, and 103 in the intervention condition with a total of 1,401 exercise submissions.

As can be seen in the box plots in Figure 2.14, the picture is similar for continued practice within a learning unit. Slightly more students in the intervention condition (146) used this means of revising mastered `realizations of learning targets` than students in the control condition (138), and the control group submitted overall 6,923 exercises while the intervention group provided 7,805 exercise submissions. According to a mixed effects analysis controlling for the school class, the difference between control and intervention group in using the Fitness Center rather than continued practice within the specific learning unit was not significant (t=.526, p=.599; g=.0483), slightly favoring the intervention condition.

Next, we examined whether exercise difficulty in the Fitness Center was more adequate for the intervention condition than for the control condition. To this purpose, we evaluated learners' accuracy scores on revision exercises with a focus on

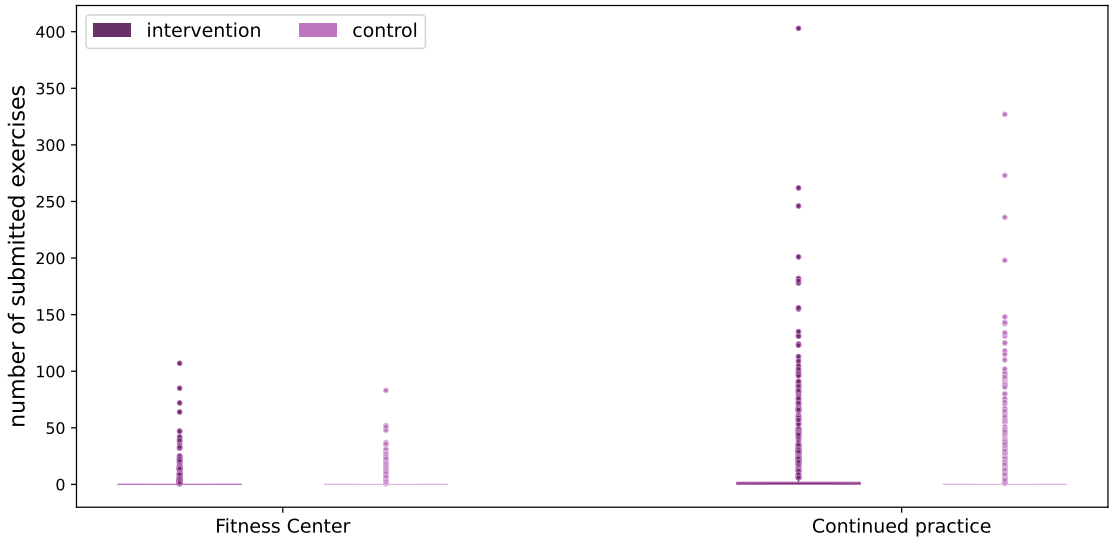


Figure 2.14: Revision practice preferences.
Learners in the intervention condition used both the Fitness Center and continued practice within a learning unit slightly more frequently than learners in the control condition.

the difference between the two study conditions. The literature on language learning does not define optimal accuracy on practice exercises. In Mastery Learning, however, a mastery threshold of 80-90% is applied (Block and Burns, 1976). Assuming that practice accuracy should be slightly lower, be define adequate practice accuracy as 60-80% correct. Although learners could correct their answers after receiving feedback, accuracy was determined based on the first attempts on each actionable element. Significance of the difference between the conditions was tested by means of a two-tailed t-test. In these analyses, the intervention condition showed significantly (t=3.5361, p=.0006; g=.6238) higher accuracy on remedial exercises ($\mu = .9865; \sigma = .0607$) than the control condition ($\mu = .8379; \sigma = .3306$). On refresher exercises, on the other hand, the intervention condition had significantly (t=-4.5566, p=.0000; g=.7148) lower accuracy scores ($\mu = .8296; \sigma = .2701$) than the control condition ($\mu = .9707; \sigma = .0619$). This makes sense considering that remedial exercises of the intervention group, who performed better on these exercises, were of closed exercise types and thus on average easier than those provided to the control group, who received arbitrary exercise types. For refresher exercises, this is reversed since the intervention group received only half-open exercise types. On the other hand, the plots presented in Figure 2.15 show a ceiling effect for both

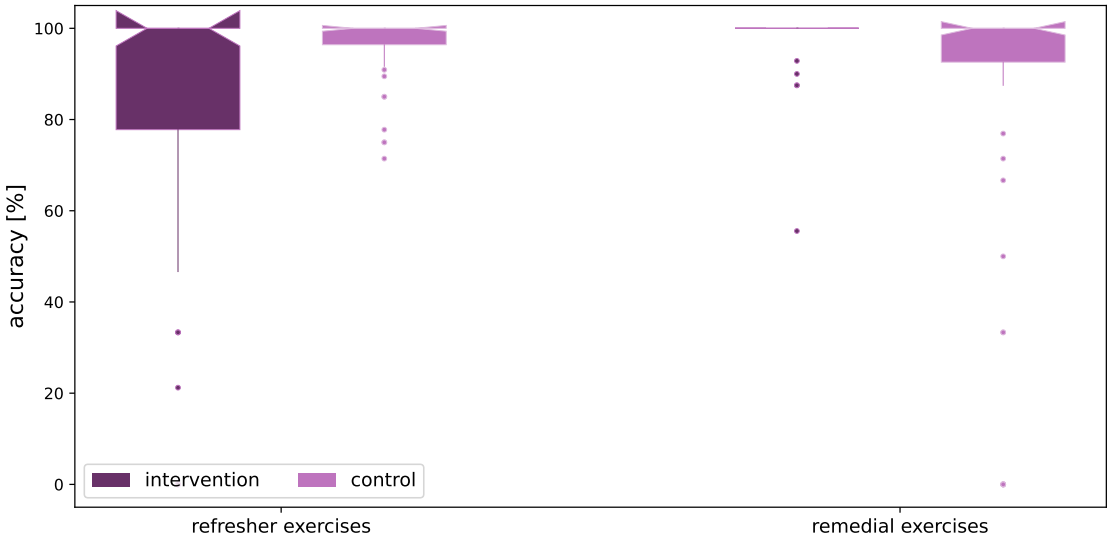


Figure 2.15: Accuracy scores on revision exercises.
Both study conditions had high accuracy scores for both refresher and remedial exercises. The intervention condition scored higher than the control condition on remedial exercises, but lower on refresher exercises.

revision `learning targets`. Learners of both conditions had overall very high accuracy scores above the 60-80% threshold we set for adequate practice accuracy. The complexity of exercises selected for revision practice should therefore generally be higher.

Lastly, we determined whether there were differences in learning gains between the study conditions. Due to the low usage of the Fitness Center, results on the beneficial effects of the Fitness Center study conditions are inconclusive. Table 2.4 provides an overview of the results of a mixed-effects analysis controlling for the school class and previous knowledge in terms of pre-test scores. Improvement was slightly higher

`Learning target`	**t-value**	**p-value**
Tenses	.008	.993
Questions	-.183	.855
Conditionals	-.480	.631
Overall	-.654	.513

Table 2.4: Relative learning gains depending on Fitness Center study conditions.
Positive t-values indicate a benefit of the treatment when contrasted to the control group. Learning gains between pre- and post-test were higher for the control group than for the intervention group in almost all `learning targets`, but none of the effects are significant.

for the control condition in almost all `learning targets`, yet none of the effects are significant.

Considering the high practice performance scores in the Fitness Center for both study conditions, a restriction of exercise types is not necessary for remedial exercises. Since only few students used the Fitness Center, it is, of course, possible that these were the most motivated students who might also be the most proficient ones. As can be seen in the histogram in Figure 2.16, most learners using the Fitness Center had high to intermediate English grades. Future research therefore needs to examine whether low-proficient learners would benefit from restricting the openness of the selected exercises depending on their purpose as refresher or remedial exercises. For more proficient learners, it does not make sense to make a distinction between remedial and refresher exercises. In light of research advocating a slight challenge for best learning outcomes, including Krashen's (1982) input hypothesis demanding input matching the learner's current proficiency level i with an additional increase $1\ (i + 1)$, we expect greater learning gains for more proficient learners when exercise difficulty is increased for both types of revision exercises. If restrictions are applied for these learners, a focus on half-open exercises for both refresher and remedial exercises might be more beneficial in terms of learning gains.

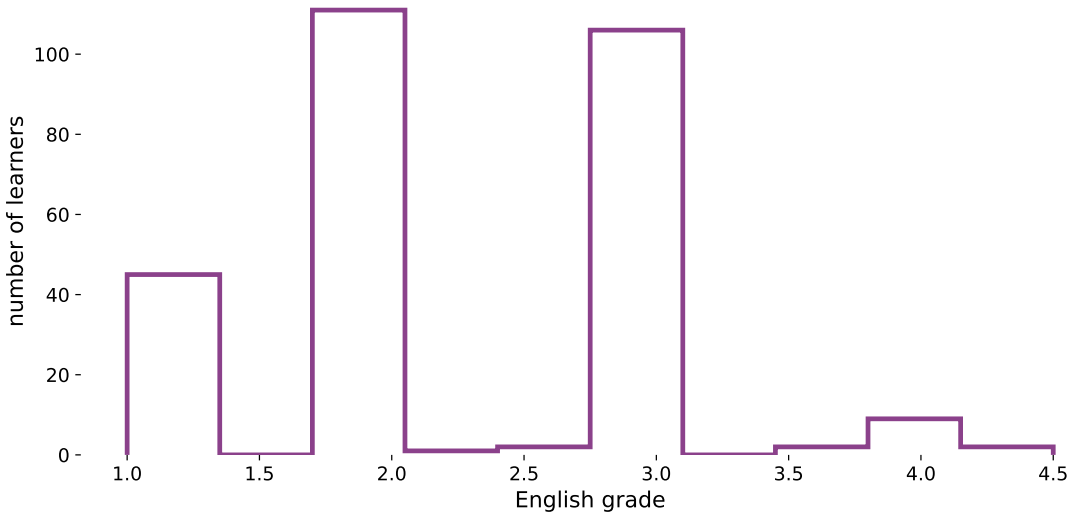


Figure 2.16: Distribution of English grades across learners using the Fitness Center.
Learners using the Fitness Center had very good to medium grades.

2.3 Summary

In this chapter, we looked at the integration of macro-adaptive ILTSs into school settings. We identified the goal of macro-adaptivity as selecting exercises matching learner characteristics, typically including the learner's proficiency level. Implementations of Mastery Learning primarily determine when to move on to the next KC. Which KC that is most often depends on the difficulty of the KCs in relation to the learner's proficiency.

Language classrooms, however, usually follow a curriculum that determines the general grammar and vocabulary contents. These may differ from one language class to another so that there is no single set of learning material and practice order suitable for all language learners. Within a class, the teacher typically determines the general content of each lesson and their order.

Reconciling adaptive exercise selection with human-orchestrated teaching therefore needs to consider the scopes within which each must operate. Our evaluations showed that the boundaries of these scopes should be flexible in order to support different levels of learner autonomy and teacher involvement. An ILTS should therefore offer multiple entry points at two distinct levels for learners to practice exercises. These two levels should differ in the scope of included exercises constituting candidates for adaptive sequencing, encompassing either all exercises of a `learning target` or only the exercises of a `realization of a learning target`. In order to also allow teachers to assign specific exercises, for example to discuss certain homework exercises in class, an ILTS should in addition offer mechanisms to specify and directly navigate to such mandatory exercises. This adds a third level of entry points to exercise selection with the most narrow scope for macro-adaptivity. Implemented this way, we can answer RQ 1 in the affirmative: The prototypical curricular organization of `learning targets` can be made compatible with adaptive sequencing implementing Mastery Learning.

In order to address forgetting of the acquired knowledge, language classrooms and macro-adaptive tools also tend to apply differing strategies. While language classrooms typically use designated revision learning units, macro-adaptive algorithms track fading knowledge and interleave revision exercises into exercise sequences of the current learning unit. A separate revision learning unit encompassing all `learning`

`targets` and within which macro-adaptive sequencing could provide revision exercises when faded knowledge has been detected does not provide the sought-for compromise. In our study, such a learning unit was neither well accepted by learners nor did it result in better learning gains. Instead, traditional revision learning units per `learning target` are better suited in school contexts. These may encompass revision of content from previous school years or of content introduced earlier in the same year. We showed that macro-adaptivity needs to apply different selection strategies for such revision units than for introductory units in order to provide sufficiently challenging practice material.

Within revision learning units or when mastery has been reached for a `realization`, we examined the suitability of different exercise types for remedial exercises targeting a learner's misconceptions and refresher exercises targeting forgetting. We saw that performance on revision exercises is generally very high for intermediate to high-proficient learners so that these exercises should be overall more complex than introductory exercises in order to be sufficiently challenging, both when targeting forgetting and misconceptions. It therefore makes sense to restrict exercise types for these learners to more open types not only after the learner has reached mastery, but also in revision learning units regardless of their purpose as remedial or refresher exercises. For low-proficient learners, further research is required. In answer to RQ 2, we therefore conclude that in school settings, adaptive ILTSs should comply with the typical setup in these contexts of designated revision units, yet may interleave refresher with remedial exercises in such units.

Chapter 3

Linguistic and Skill Variability Impacting Exercise Difficulty

Parts of this chapter are based on Publications 3, 5 and 7.

In order to select exercises fitting a learner's abilities, adaptive approaches generally try to match exercise difficulty to a learner's proficiency level. Tutoring systems should, however, also consider peculiarities of the domain they aim to teach (Yazdani, 1986). In language learning, this includes variability both in terms of practiced skills and language itself. Practicing varied skills enables learners to more broadly apply the taught linguistic knowledge (DeKeyser, 2018). Linguistic variedness is on the one hand a property inherent to language and therefore to the practice material (Yildirim et al., 2013), and on the other hand might in addition have beneficial effects on learning outcomes.

When considering further exercise parameters in addition to exercise difficulty in exercise selection, it is important to understand interaction effects between all selection parameters. If, for instance, increased linguistic variability increases exercise difficulty, exercise selection might need to prioritize one over the other if no exercise is available that perfectly matches the learner's profile in both parameters.

In this chapter, we therefore first analyze challenges to exercise difficulty calibration in order to assess whether it is possible to in addition consider further exercise parameters in exercise selection. We next evaluate whether linguistic variability is

beneficial for learning and should thus be deliberately introduced in practice material (RQs 3 and 4). Finally, we examine the impact of linguistic and skill variability on exercise difficulty (RQs 5-7) in order to be able to come up with appropriate macro-adaptive strategies.

Research Questions

RQ3 Does knowledge of a linguistic form transfer to other linguistic forms with similar properties? Is this transfer conditioned by varied practice?

RQ4 Is varied practice beneficial for proceduralization? Does it result in more varied learner productions?

RQ5 Does exercise difficulty vary across linguistic forms of similar properties?

RQ6 Does linguistically varied practice increase practice difficulty? If so, is this difficulty desirable?

RQ7 Does co-text complexity impact the difficulty of form-focused grammar exercises?

3.1 Challenges of Exercise Difficulty Calibration

Exercise difficulty constitutes the probability of a learner answering the exercise correctly (Pelánek et al., 2021). It is closely related to exercise complexity, yet there are important distinctions. Exercise complexity constitutes a property of an exercise that is operationalized based on exercise parameters such as the length of the co-text in terms of the base text into which actionable elements are embedded, the number of concepts it involves, and interactions between these parameters. While measuring exercise complexity is not always trivial, difficulty is usually more easily measured. Operationalized for example as failure rate, response time, hint usage, and number of attempts, it represents learners' performance based on their interactions with the exercise.

The Desirable Difficulty Framework (DDF) uses a slightly different terminology for related concepts. It introduces three dimensions to exercise difficulty (Suzuki et al., 2019): (1) Linguistic difficulty is an absolute complexity value inherent to the target construct to be acquired. This dimension is encompassed in exercise complexity. (2) Learner-related difficulty considers learner characteristics that affect a construct's difficulty depending on the specific learner. This dimension corre-

sponds to exercise difficulty. (3) Context-related difficulty considers the impact of the practice condition, such as time constraints or interleaved vs. blocked practice, on exercise difficulty. This additional dimension is mentioned as confounding factors in Pelánek et al.'s (2021) categorization, but assigned its own dimension in the DDF. In fact, the DDF uses this dimension to intentionally manipulate exercise difficulty. The goal of such manipulations is to create practice conditions where learners may have lower – but successful – performance during practice, but higher long-term learning gains thanks to the increased effort required by higher difficulty (Suzuki, 2023). Yet this increased effort only results in desirable difficulty if the learner's abilities match the increased processing demands. Otherwise, it results in undesirable difficulty (Bjork and Bjork, 2011).

In the following, we consider the trinity of complexity used in the DDF, but extend linguistic complexity to exercise complexity considering all factors inherent to an exercise independent of the specific learner or the environment. We therefore use slightly different terminology when referring to the three dimensions: (1) General exercise difficulty comprises learner-independent, static parameters such as linguistic complexity of the textual material and characteristics of the exercise types. (2) Learner-dependent exercise difficulty includes, for example, cognitive abilities and personal experience. (3) Context-dependent exercise difficulty refers to external factors affecting exercise difficulty that do not originate in learner characteristics.

An essential prerequisite for exercise selection consists in calibrating the difficulty of the available exercises. Difficulty estimates are obtained either through human ratings or by means of data driven approaches (Wauters et al., 2012). Human rating based approaches are subjective in nature and require manual effort. Data based approaches are more objective, yet they rely on large amounts of learner answers in order to provide reliable difficulty estimates. Aiming to overcome this shortcoming, multiple strategies have been explored to determine exercise difficulty based on a range of exercise parameters instead.

Approaches estimating exercise difficulty primarily rely on general exercise difficulty measures. They consider mostly one-time estimations of exercise difficulty so that difficulty estimates are not specially adjusted to the target group of users. In order to take the target group into account to some degree, Ravi and Sosnovsky (2013) propose to re-calibrate exercise difficulty once enough training data is available.

More dynamic implementations adjusting the exercises' difficulty scores as learners work on the exercises have also been explored (Koutsojannis et al., 2007).

Using data-based approaches to determine an exercise's difficulty requires learner answers for that specific exercise. We have previously discussed IRT in the context of macro-adaptivity. Dichotomous IRT models, which are based on binary learner answers, provide a popular means to determine a difficulty score for an exercise item (Ravi and Sosnovsky, 2013). In order to account for dependencies between items in the same exercise, Eckes (2011) use a polytomous IRT model with possible difficulty scores between 0 and the number of items in the exercise. Wauters et al. (2012) explored and compared alterative approaches to determine item difficulty. Their results identified Elo rating, which we also have previously discussed in the context of adaptive sequencing, as a good and simpler alternative. Using the ratio of correct answers to all answers provided by learners as difficulty estimates yielded rather reliable results as well. These approaches, however, offer no solutions to determine the difficulty of newly created exercises for which no learner data is available. For these cases, the authors found subjective learner feedback to work best. Expert ratings, on the other hand, did not yield reliable difficulty estimates in their evaluation. As a compromise, Labaj and Bieliková (2014) use a combination of expert and learner ratings to address the cold-start problem through expert ratings, and dynamically adjust the difficulty scores based on learner ratings and their proficiency scores at the time of submitting the exercise.

While expert ratings avoid the cold start problem, they constitute cumbersome, manual tasks. Researchers have therefore explored various approaches to determine the difficulty of exercises based on exercise parameters.

3.1.1 Exercise Parameters Impacting Exercise Difficulty

Hartig et al. (2012) point out that the relevant parameters vary depending on the skill targeted by the exercise so that the set of parameters likely differs between content domains focusing on fact learning, and domains like Maths or language learning with a focus on proceduralization. For language exercises, most work so far has focused on cloze exercises with a particular emphasis on C-tests. This test format

consists of short texts with blanks. It requires test takers to fill the blanks in order to complete the partially give tokens. In an early approach, Wilson (1994) used co-text readability as a single determining feature of exercise difficulty, acknowledging the need to yet establish its correlation with general exercise difficulty. Others have identified a range of linguistic features on the word, sentence, and text levels that impact exercise difficulty (e.g., Galasso, 2018; Beinborn et al., 2014; McCarthy et al., 2021; Settles et al., 2020; Brown, 1989). Word difficulty features include the linguistic (e.g., morphology, spelling) and psycholinguistic (e.g., age of acquisition, frequency) domains and the candidate space. Sentence difficulty features focus on syntactic complexity and sentence length. Text difficulty features comprise readability parameters ranging from lexical and morpho-syntactic to semantic and discourse features (Pallotti, 2019; Beinborn et al., 2015; Svetashova, 2015). The observed effect of exercise format parameters such as gap size, deletion pattern, and deletion frequency on exercise difficulty varies across studies (Sigott, 1995; Lee et al., 2019; Kamimoto, 1993). Beinborn et al. (2015) point out that most of the parameters relevant to C-tests are not applicable to the closed type Single Choice exercises. The authors instead present promising results for the use of context fitness of Single Choice options. Exercises are more difficult if one of the distractors receives a higher fitness score than the target answer. While Holzknecht et al. (2021) found that Single Choice reading comprehension exercises were more difficult when the correct option was in the last position, studies on Single Choice exercises in other domains found exercises with the correct option in the first or last position (Attali and Bar-Hillel, 2003), or next to the most attractive distractor (Shin et al., 2020) to be harder. Not focusing on language exercises, Swanson et al. (2006) explored the number of distractors, finding that more distractors resulted in more difficult exercises. 8 out of 9 studies included in Haladyna and Downing's (1989) review found longer correct answers to be easier, whereas one study found no length effects. In their comparisons of syntactically, paradigmatically, and not related distractors, Hoshino (2013) found that the syntactically related ones were the most difficult distractors, yet only in exercises that require semantic parsing of the co-text. Kubinger and Gottschall (2007) concluded that Multiple Choice exercises, which allow multiple correct options, are similar in difficulty to free response exercises and more difficult than Single Choice exercises with a single correct option. Abraham and Chapelle (1992) explored different input types and found drop-down selection to be

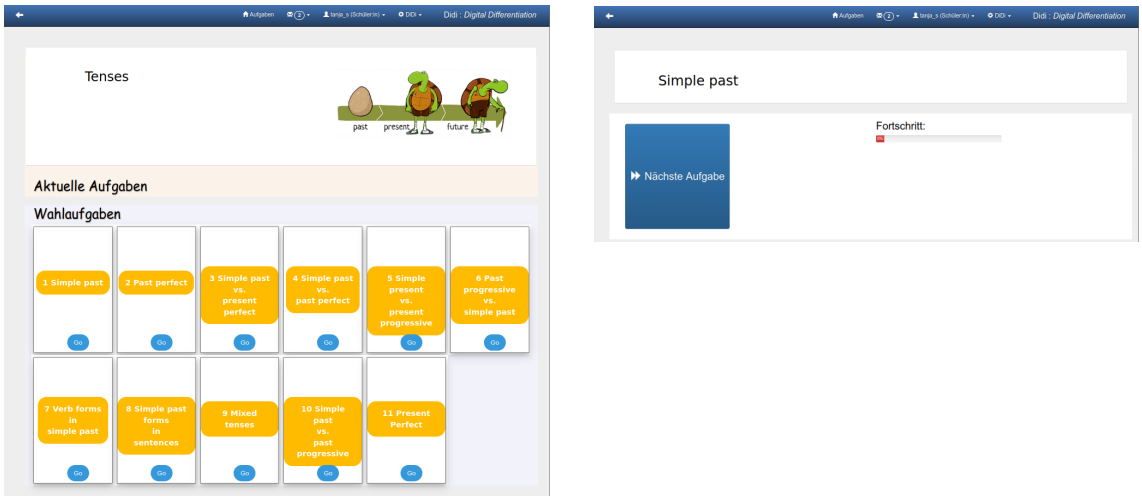
easier than text input. Oller (1972) identified reading comprehension Single Choice exercises as more difficult than Jumbled Sentence or Single Choice grammar exercises, and dictation exercises as even more difficult. Also in the context of reading comprehension, Kobayashi (2002) analyzed properties of the target words in terms of their influence on the difficulty of cloze exercises for Japanese learners of English. According to the author's evaluations, function words are easier than content words. For specific parts-of-speech (POSs), she attributes differences in difficulty mostly to differences between the native and target language. More frequent target words, either by general word frequency or by occurrences within the co-text, tended to be easier. Items were more difficult the more context was required in order to solve them. Also considering the item co-text, Hartig et al. (2012) found that an item's grammatical and lexical complexity had less impact on exercise difficulty than the distribution of relevant information across the text related to global text comprehension. De Kuthy and Meurers (2021) mention additional parameters in their survey that have been explored in past research in a variety of domains: The degree of abstractness of the target concept, cognitive processes involved in solving the task, subject-specific terminology, and complexity of the information have been found to affect exercise difficulty.

When language exercises are automatically generated, their complexity is sometimes already determined and controlled for during generation (Kurdi et al., 2020). In this line of work, Pilán et al. (2017) considered only the co-text complexity of their Single Choice exercises for vocabulary practice. Generating the same type of exercises, Susanti et al. (2017) in addition used semantic similarity between the correct option and the distractors, as well as the word-level complexities of the distractors in order to control exercise difficulty.

Comparing the effect of exercise parameters provides merely relative difficulty values in relation to other instantiations of the same parameter (Lorge and Kruglov, 1953). In order to match exercise difficulty to learner abilities, however, adaptive systems usually rely on absolute difficulty scores. In addition, interaction effects between different exercise parameters make it impossible to rank exercises by their difficulty based on independently assessed exercise parameters (Pandarova et al., 2019). In order to obtain an aggregate, absolute difficulty score from exercise parameters, more recent approaches usually take available difficulty scores for similar exercises as a

reference. A number of approaches generating Single Choice reading comprehension exercises applied machine learning and statistical methods to generate predictors of exercise difficulty from the text, the question, and answer options (Liu et al., 2021; Huang et al., 2017; Loukina et al., 2016; Benedetto et al., 2020). In the medical domain, Xue et al. (2020) rely on embeddings, providing both the question and Multiple Choice options as input. Ha et al. (2019) found that jointly using embeddings and linguistic complexity features yielded best results. Fang et al. (2019) in addition consider images contained in the exercises in the input to their multimodal embedding based classifier. Very little research has focused on grammar exercises. A noticeable exception constitutes the approach by Pandarova et al. (2019), which examines the effect on exercise difficulty of various linguistic properties of the gap, item, and text of Fill-in-the-Blanks exercises practicing tenses. Parameters of all three scopes were found to impact exercise difficulty. Similar to the approach employed by Benedetto et al. (2020), this implementation determines the final difficulty scores by means of a regression model.

The number of exercise parameters impacting exercise difficulty is vast, and considering them all when determining general exercise difficulty scores poses various challenges. Manually annotating all parameters is extremely time-consuming, so



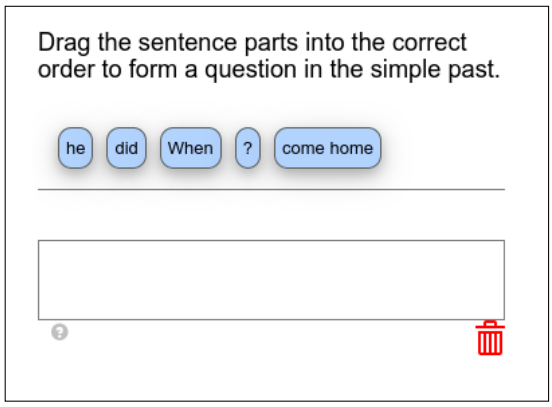
(a) Grammatical `learning targets`. (b) Adaptive exercise sequencing.

Figure 3.1: UI of the FeedBook version used in the Didi study.
The content was organized in grammatical `learning targets`. The system selected the next practice exercise in a user-adaptive manner.

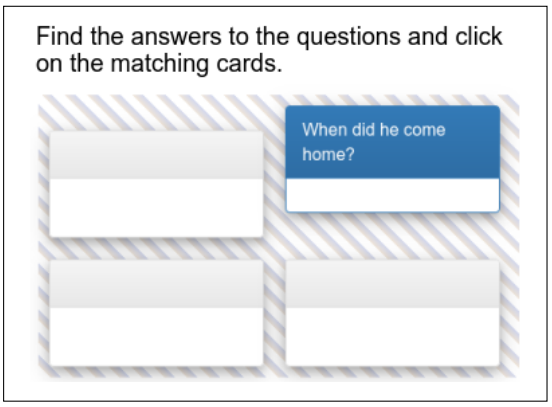
automatic processes are necessary. These, however, need to be reliable in order to yield reliable difficulty estimates. In addition, considering many exercise parameters considerably increases the amount of necessary training data in order to prevent overfitting (Liu et al., 2017a). We therefore aimed to identify those exercise parameters that are most predictive of exercise difficulty.

Data To this purpose, we analyzed exercises and learners' submissions collected in RCT studies of two projects: I4S and Didi (see Section 2.1.3). Both studies were conducted in the school year 2021/22 and with a learner population of 7^{th} grade Gymnasium students. They differed from each other in their research focus which was on motivational aspects in I4S, and on user-adaptive exercise sequencing in Didi, as depicted in Figure 3.1b. As can be seen in Figure 3.1a, the FeedBook version used in the Didi study therefore groups exercises only according to `learning targets` without the additional meta-level of cycles available in I4S. For form-focused grammar exercises, the two studies covered the seven exercise types illustrated in Figure 3.2: Fill-in-the-Blanks, Single Choice, Jumbled Sentence, Categorization, Memory, Short Answer, and Mark-the-Words exercises. Each of the 201 exercises in the I4S study – excluding listening exercises – that were attempted by at least one learner was submitted by a mean of $\mu = 136.13$ ($\sigma = 112.58$) learners. They were grouped into four learning units and contained a total of 1,140 actionable elements. An actionable element could be the blank of a Fill-in-the-Blanks or Single Choice exercise, a sentence of a Jumbled Sentence exercise, a clickable chunk in a Mark-the-Words exercise, an element to sort in a Categorization exercise, a Memory pair, or an answer to a Short Answer exercise. In the Didi study, a mean of $\mu = 29.19$ ($\sigma = 46.00$) learners attempted each of the 470 exercises with overall 2,003 actionable elements. These numbers differ considerably from those of the I4S study as the macro-adaptive focus of the Didi study resulted in a more varied practice environment adapted to the individual learner. The exercises were distributed across eleven distinct `learning targets` ($n_{I4S} = 9$; $n_{Didi} = 4$). While revising and re-submitting an exercise was possible in the studies, the evaluations only consider first submissions ($n_{I4S} = 153,596$; $n_{Didi} = 120,431$). There is no overlap of learners or practice exercises between the two studies.

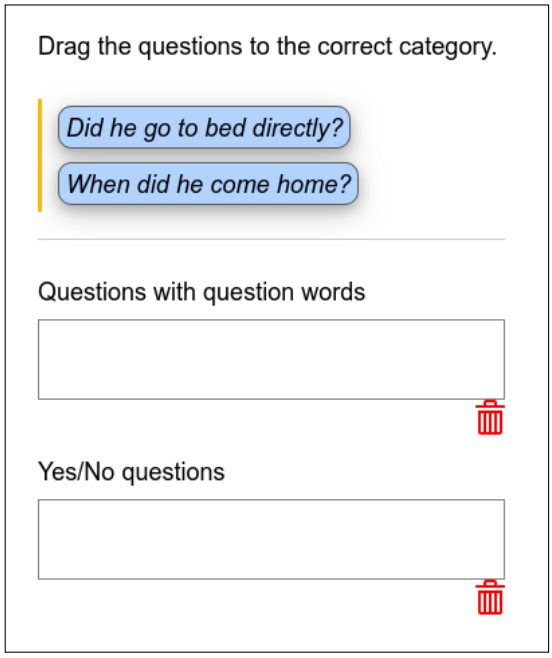
For this dataset, we trained classification and regression models from the Python



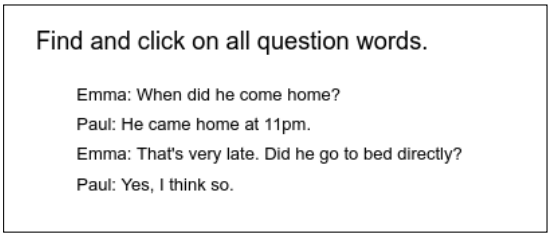
(a) Jumbled Sentence.



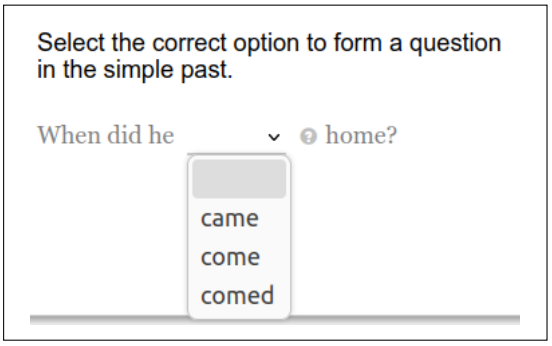
(b) Memory.



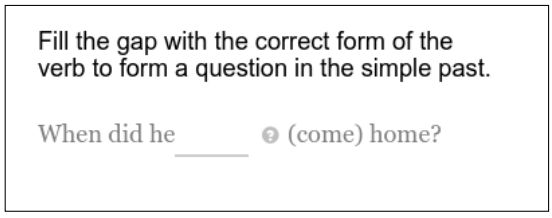
(c) Categorization.



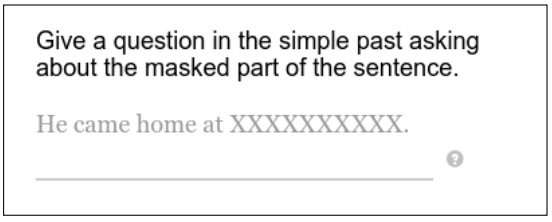
(d) Mark-the-Words.



(e) Single Choice.



(f) Fill-in-the-Blanks.



(g) Short Answer.

Figure 3.2: Exercise types in I4S and Didi.
Both studies supported seven exercise types in the ILTS FeedBook.

scikit-learn library[1] with a train-test split of 75-25% to predict exercise difficulty based on a selection of exercise parameters. These parameters cover (1) generally applicable parameters including the number of actionable elements in the exercise and the length of an actionable element, (2) exercise type specific parameters such as the number of distractors of Single Choice exercises, the number of chunks of a Jumbled Sentence exercise, the number of categories of a Categorization exercise, the number of pairs of a Memory exercise, or whether a Fill-in-the-Blanks exercise requires the learner to determine the correct lemma in addition to transforming it into the correct form, and (3) the exercise type itself. All predictors were encoded as numerical features. Difficulty of the actionable elements was operationalized as item difficulty scores obtained from an IRT model[2]. While the continuous scores served

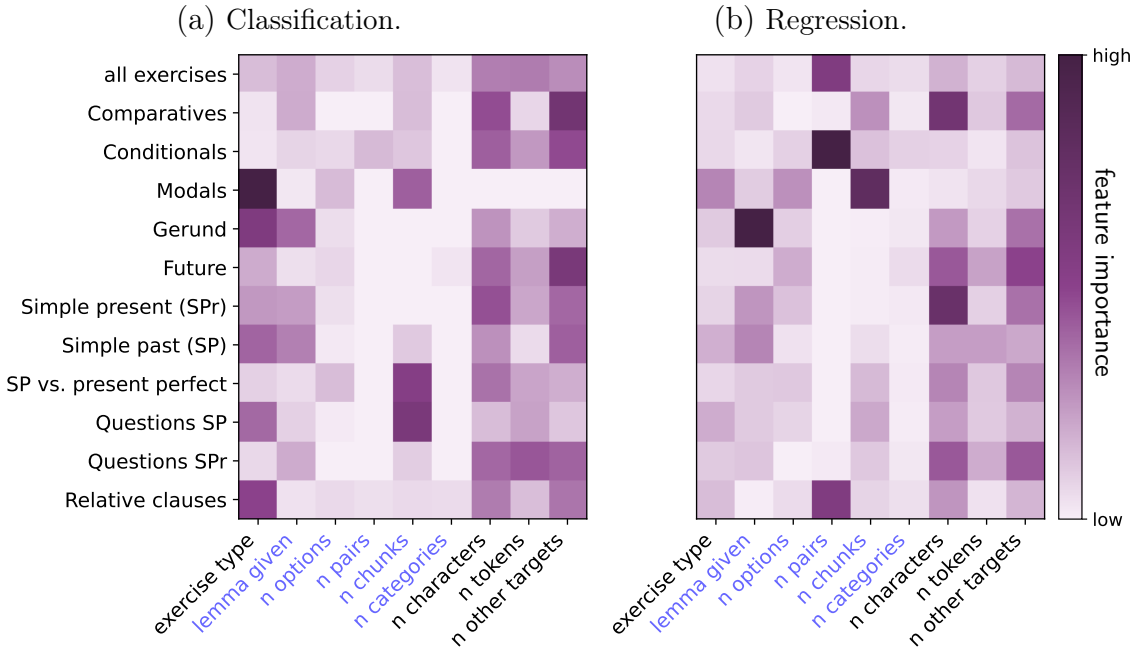


Figure 3.3: Feature importances in statistical models predicting exercise difficulty.
The exercise type is an important feature for classification to predict exercise difficulty in most `learning targets` (y-axis). In regression models, type-specific parameters (blue) have the highest predictive power.

[1] scikit-learn is available at https://scikit-learn.org.

[2] IRT difficulty values b were determined for all actionable elements based on the Rasch model of the `TAM` package for R. Since the datasets from the two studies constitute discrete sets with no overlap in learners or exercises, we determined the difficulty values independently for each dataset. For performance reasons, the data in addition needed to be split by `learning targets`.

directly as outcome variables for the regression models, they were transformed into categorical values for the classification models. Since the FeedBook distinguishes between three proficiency levels for learners, we applied the same number of exercise difficulty levels. Analogous to Katz et al. (2021), the thresholds between the three levels were determined through K-means clustering. All employed machine learning models support feature ranking, which allows to easily determine their most predictive features. In order to identify the overall most predictive features, we calculated the mean of the predictor ranks from all regression and those from all classification models, thus obtaining overall rankings for regression and classification, respectively.

The heatmaps in Figure 3.3 assign colors of increasing darkness to combinations of parameters (x-axis) and `learning targets` (y-axis). The darker the color, the more important the parameter is for that `learning target`. They show that for generally applicable parameters, the rankings are rather similar across `learning targets` for both the regression and the classification models. Generally applicable parameters all occupy ranks in the middle ranges, indicating that while they do not constitute the most informative of the evaluated parameters, their predictive power is rather constant and reliable across exercises. The exercise type specific features show considerably more variation across `learning targets` especially with regression, as peaks appear at both extremes of the ranking. The exercise type overall emerges as rather important for classification, whereas it ranks among the least predictive

	Lemma given	n options	n pairs	n chunks	n categories	n characters	n tokens	n other targets
Fill-in-the-Blanks	X					(X)	X	(X)
Single Choice		X				(X)	(X)	(X)
Short Answer						(X)	X	(X)
Mark-the-Words						(X)	(X)	X
Jumbled Sentence				X		(X)	(X)	(X)
Categorization					X	(X)	(X)	(X)
Memory			X			(X)	(X)	(X)

Table 3.1: Applicability of exercise features for exercise types. Generally applicable parameters are applicable to all exercise types. Type-specific exercise features are only applicable for one exercise type.

features for regression. Type-specific parameters hold more predictive power with those models. This is in line with Svetashova's (2015) results where item level features were most predictive for regression and item and text level features for classification.

Table 3.1 presents a mapping of exercise parameters to exercise types, illustrating that generally applicable features can be determined for all exercise types whereas type-specific features are applicable to only one exercise type.

The regression results per exercise type, illustrated in the heatmap in Figure 3.4, highlight that the parameters that are applicable to only a particular exercise type indeed are more important for the difficulty of the respective exercises.

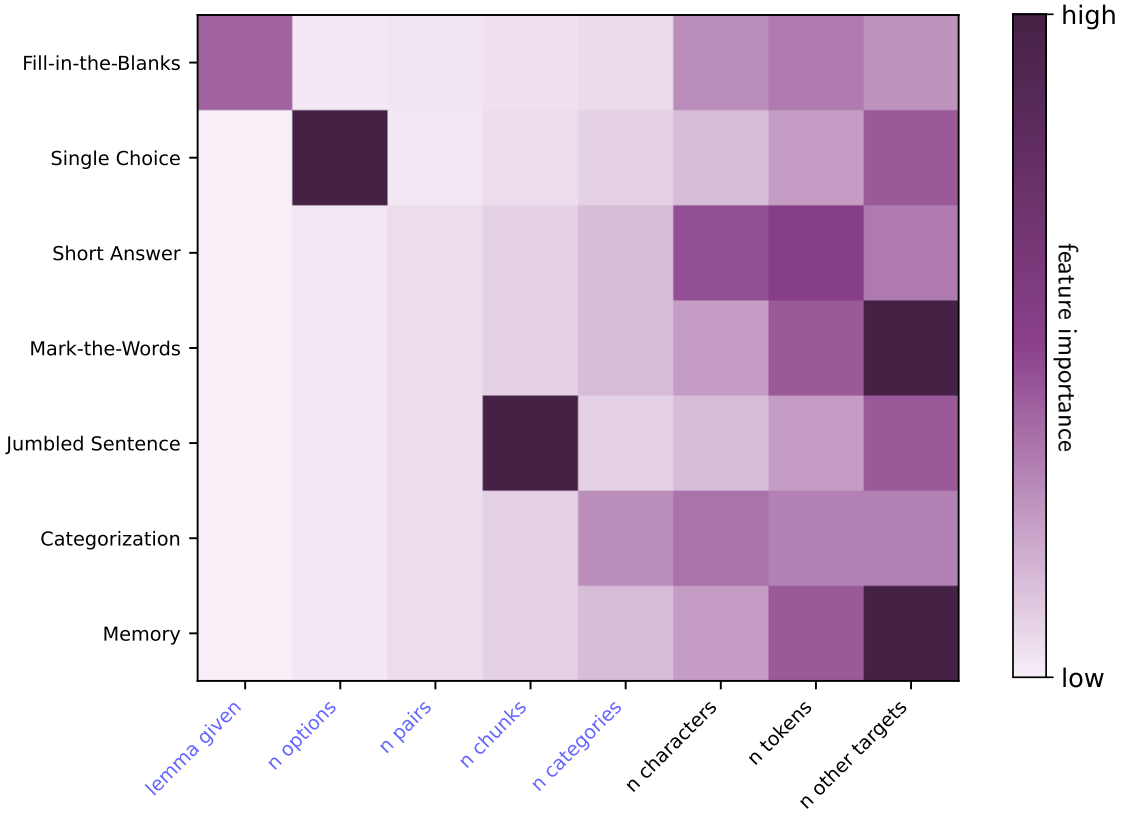


Figure 3.4: Feature importances for regression predicting exercise difficulty per exercise type.

Darker colors in the heatmap indicate greater importance of the feature when predicting exercise difficulty for the corresponding exercise type with the regression model. Exercise type-specific features (blue) are more important for their respective exercise type than for other types.

More global features thus seem to be more predictive for classification whereas regression relies on local features. This might indicate that the exercise type can inform coarse-grained difficulty estimations, while fine-grained distinctions require more detailed, type-specific parameters.

3.1.2 Validity of Difficulty Metrics

Since exercise difficulty constitutes the predominant discriminative exercise parameter for most macro-adaptive implementations, it is important that its operationalization yields valid difficulty estimates. Research has so far paid little attention to comparing operationalizations of exercise difficulty such as answer correctness or response time. An exception constitutes Wauters et al.'s (2012) comparison of a number of manual and data-driven approaches. Yet the authors look at overall exercise complexity without considering the effect of a specific operationalization on the difficulty scores obtained for individual exercise parameters. Another issue with general exercise difficulty estimates consists in that they are applied to all learners so that their validity is compromised if the difficulty parameters are learner dependent but the adaptive system fails to take corresponding learner characteristics into account. IRT aims to address learner dependence by jointly considering learner abilities and exercise difficulty. However, in order to obtain binary exercise evaluation scores as input for IRT based difficulty estimation, most approaches determine whether a learner got the answer correct at the first try but do not consider alternative operationalizations of exercise difficulty. We therefore compared the operationalization *correctness at first try* with the alternative operationalization *number of attempts*, and with learner-sensitive *IRT estimates* with respect to their difficulty estimates for the exercise parameter *exercise type*.

The data described in Section 3.1.1, including the IRT difficulty scores, constitutes the basis for these evaluations. Since absolute difficulty scores cover different ranges depending on the operationalization of exercise difficulty, they need to be normalized into comparable ranges. The resulting scores, however, then depend on the normalization approach, so that absolute difficulty scores might differ depending both on the operationalization of exercise difficulty and the normalization approach. We therefore merely ranked the exercises according to their difficulty scores obtained with the different metrics.

The box plots in Figure 3.5 display the distributions of differences between the metrics when comparing exercise types. The rankings obtained through the three metrics resulted in relatively similar ranking positions for most exercises. There are, however, edge cases where one operationalization yields higher results for one exercise type and lower results for another type than an alternative operationalization. This may even result in different relative difficulty for exercise types depending on the operationalization, as with Jumbled Sentence and Categorization exercises.

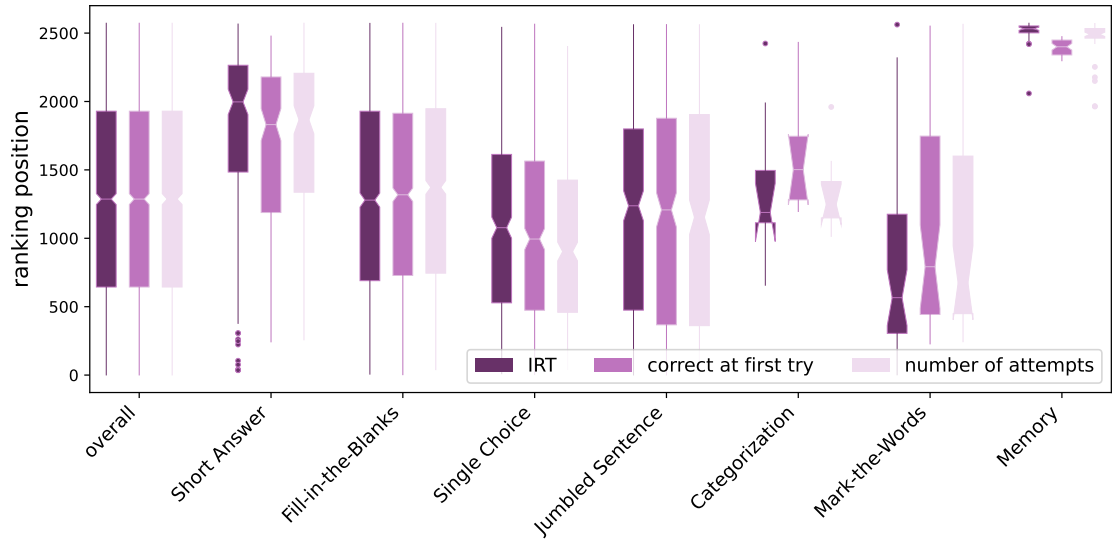


Figure 3.5: Exercise difficulty ranking positions per exercise type across difficulty metrics.
The difficulty ranking positions assign each exercise a difficulty value relative to the other exercises for each learner individually. The ranking positions constitute incremental integers. They are based on mean difficulty values over all submitted exercises of the corresponding type. The difficulty value calculation depends on the difficulty metric.

Obtaining reliable difficulty estimates for exercise parameters is thus a challenging task as relative difficulty scores of some exercises are substantially influenced by the concrete operationalization of difficulty and are highly learner dependent in some cases. Obtaining absolute difficulty scores that validly represent an exercise parameter's difficulty for a given learner can be expected to be even more challenging. Such an endeavor requires to carefully consider which dataset of exercises and learner answers should be used to determine the estimates. It should also consider learner characteristics in conjunction with the exercise parameters. As learner models are dynamic, they need to be factored in at the time of selecting an exercise (Kunichika

et al., 2002). In addition to putting stress on the system's performance through these online calculations, taking learner characteristics into account also introduces an additional cold start problem when new learners start interacting with the system for the first time. If all three dimensions of the DDF are to be considered, the system must in addition be able to consider dynamic, context-dependent exercise difficulty. Contextual influences such as previously attempted exercises or scaffolding are manifold (Pelánek et al., 2021). Accounting for all relevant context-dependent exercise difficulty parameters is thus extremely challenging especially if external restrictions such as lack of sensors or data privacy regulations prevent access to the required data sources (Hasanov et al., 2019).

3.1.3 Cold Start Problem for Learner-dependent Exercise Difficulty

Since the learner model is empty for new learners interacting with the system for the first time, considering learner-dependent exercise difficulty poses a serious challenge (Kerres et al., 2023). Extensive placement tests intended to cover all `learning targets` introduce a tedious entry barrier to the system (Al-Jadaa et al., 2017). Moreover, such tests do not offer any solution for `learning targets` that are later on introduced into the system. We therefore analyzed the suitability of C-tests as alternative placement tests to initialize the learner model.

C-tests are widely used to assess general language proficiency and have been shown to reliably and validly do so (Klein-Braley, 1996; Sarapuu and Alas, 2016). However, more recent critical evaluations have yielded mixed results, ranging from high (e.g., Lei, 2008; Rasoli, 2021) to very low (e.g., Farhady and Jamali, 2006; Rouhani, 2008) validity for English. These discrepancies might stem from differences in the participants as Rouhani (2008) found C-tests to only be reliable for certain proficiency groups.

We therefore conducted a range of analyses with the aim of determining whether, for our target group of German 7$^{\text{th}}$ grade Gymnasium learners of English, C-tests provide useful ability estimates to initialize a learner model.

To this purpose, we enriched the dataset presented in Section 3.1.1 with additional metadata concerning co-text complexity, as well as learners' C-test scores. In order

to determine co-text complexity of the exercises in the dataset, we extracted the text material from all exercises. This includes prompts as well as all actionable elements and surrounding co-text. It does not include instructions or any support texts such as hints or feedback messages. For lack of gold standard values in our dataset for text complexity, we operationalized co-text complexity for each of the extracted texts as the mean value of normalized[3] readability scores obtained from the readability formulas outlined in Table 3.2. Readability formulas constitute easy-to-use measures of linguistic complexity thanks to their numerical output scores and therefore serve the purpose of obtaining approximate estimates of co-text complexity. Although IRT scores were estimated separately for the `learning targets`, normalized readability scores were determined together for all exercises of the full dataset as text complexity should be independent of exercises and learners.

Note

We acknowledge that readability formulas do not cover the entire spectrum of linguistic properties relevant to linguistic complexity. Additionally, more sophisticated approaches could differentiate between different scopes of the features since for some exercises only the linguistic constructs in the sentence of the actionable element might impact exercise difficulty, whereas in others, the linguistic constructs of the entire co-text might be relevant.

In addition to submissions of practice exercises, both studies collected performance data on C-tests. These tests were conducted once at the beginning and once at the end of the studies, thus framing the practice exercises. The C-tests used at both test timepoints and in both studies are identical and consist of six parts in a fixed order that each comprise a short text with blanks. Of the 1,360 learners consenting to participate in the studies, 1,102 completed the first and 774 the second C-test. 553 learners completed both C-tests. The C-test scores were determined for each part as the learner's accuracy in terms of the ratio of correct answers to the number of actionable elements of the part. Learners may start working on a new `learning target` at random times when interacting with an ILTS. Since the learner model lacks data not only upon first interaction with the system but

[3]We used the `StandardScaler` of the Python scikit-learn package for scaling of the readability scores of each formula, and the `MinMaxScaler` of the same library to scale the mean readability scores into the range [0,1].

	Formula (Wang et al., 2022)	**Variables**	**Output**
Flesh Reading Ease	$206.835 - 1.015\frac{w}{s} - 84.6\frac{syl}{w}$	syl: n syllables w: n words s: n sentences	score [100;0]
Flesh-Kincaid Grade Level	$0.39\frac{w}{s} + 11.8\frac{syl}{w} - 15.59$	syl: n syllables w: n words s: n sentences	grade level
Coleman-Liau Index	$0.0588 * c - 0.296 * s - 15.8$	c: n characters s: n sentences	grade level
Automated Readability Index	$4.71 * \frac{c}{w} + 0.5 * \frac{w}{s} - 21.43$	c: n characters w: n words s: n sentences	grade level
Dale-Chall Readability Score	$0.1579 * \frac{100hw}{w} + 0.0496 * \frac{w}{s} + 3.6365$	w: n words hw: n hard words s : n sentences	score [0;10]
Linsear Write Formula	$\begin{cases} \frac{\frac{ew+3hw}{s}}{2}, & \text{if } \frac{ew+3hw}{s} > 20 \\ \frac{\frac{ew+3hw}{s}}{2} - 1, & \text{otherwise} \end{cases}$	hw: n hard words ew: n easy words s: n sentences	grade level
Gunning FOG Formula	$0.4 * (\frac{w}{s} + 100\frac{hw}{w})$	w: n words hw: n hard words s: n sentences	grade level

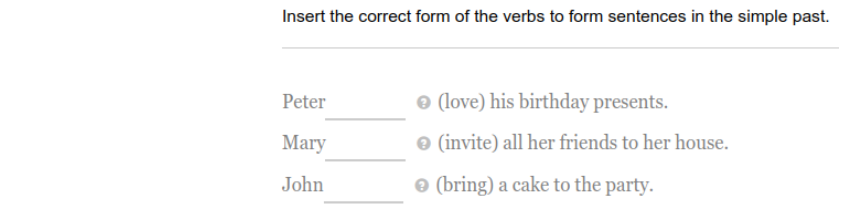
Table 3.2: Readability formulas.
The output represents either a score within a given range form easy to difficult, or a grade level.

also for any newly started `learning target`, we considered both C-tests as means to initialize the learner model for a new `learning target`. This allowed us to also investigate whether a C-test taken upon first interaction with the system still

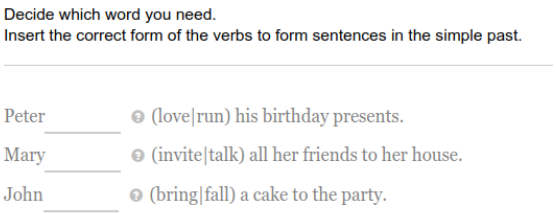
yields representative ability estimates for `learning targets` started after a period of interaction with the system.

Since we assumed that general language proficiency might correlate stronger with exercise difficulty for some `learning targets` and exercise types than for others, we extracted subsets of exercises for isolated analyses. Each combination of exercise type and `learning target` resulted in a distinct subset of exercises. In addition, some data might not be representative due to the low number of submissions for an exercise. A further subset of key exercises therefore is based on the number of learner submissions for the exercises. It encompasses all exercises that were submitted by at least 50% of all learners in the respective study. The next three subsets control for learner proficiency. To this purpose, learners with very low, intermediate, or very high C-test scores were extracted. In order to maximize the number of learners per subset while minimizing the range of proficiency scores within each subset, intermediate proficiency was assigned if the learner's C-test score of the entry C-test deviated from the median value by no more than 1%. For low and high proficiency, we determined the same number of learners with the lowest and highest C-test scores respectively. Each subset contains only the submission data for exercises attempted by the learners of the respective proficiency group, and not all submissions were included in one of these three subsets. Similarly, three additional subsets control for exercise difficulty. They consist of exercise items with similar IRT difficulties in the low, intermediate, and high difficulty ranges. Each item in the subsets represents a single actionable element. In the intermediate difficulty subset, we again included exercises that deviate from the median value by no more than 1%. For the low and high difficulty subsets, we used the same number of exercise items with the lowest and highest difficulty scores respectively. Fill-in-the-Blanks exercises require special attention. They support two possible codings, as illustrated in Figure 3.6: (1) Specifying the required lemma in parentheses behind the blank (3.6a) results in mechanical drill exercises. (2) Giving the lemmas as bags of words for the entire exercise (3.6c) or providing an additional distractor lemma in parentheses (3.6b) requires top-down skills in the form of parsing the co-text (Frehan, 1999) in order to successfully answer the exercise. We extracted the second type into an additional subset of co-text sensitive exercises.

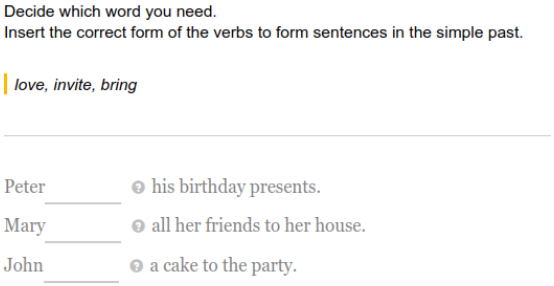
Applications in the field of readability assessment tend to align a text's linguistic



(a) Lemma in parentheses.



(b) Distractor lemma in parentheses.



(c) Lemmas as bag of words.

Figure 3.6: Codings of Fill-in-the-Blanks exercises.
Hints for Fill-in-the-Blanks exercises may differ from one exercise to another. Their coding constitutes an important parameter of this exercise type.

complexity with a learner's general language proficiency (Chen and Meurers, 2019). We therefore expected the difficulty of co-text sensitive exercises to correlate stronger with general language proficiency than that of exercises not requiring top-down skills. However, it only makes sense to attribute an impact on exercise difficulty to co-text sensitivity if the nature of the learners' errors is indeed related to incorrect semantic processing. We therefore determined for co-text sensitive exercises whether learner answers incorporated incorrect lemmas. We used an automatic approach to determine the nature of the learners' errors and mistakes to this avail. Learner answers for which the edit distance to the correct lemma or to the correct form was shorter than the edit distance to the distractor lemma or to one of the lemmas in the bag of words were annotated as errors not related to semantic processing; otherwise, we annotated the error as related to semantic processing. The results, although only an approximation due to the automatic annotation of the nature of learner errors, suggest that roughly half of the errors and mistakes (55.26%) made in those exercises are indeed related to incorrect semantic processing. Failing to correctly process an exercise's co-text therefore truly constitutes a problem with

co-text sensitive exercises.

In order to evaluate the suitability of assessing general language proficiency through C-tests for our target group, we determined the distributions of the C-test scores based on histogram plots. Although Daller and Phelan (2006) point out that C-tests are not necessarily normally distributed, we expected similar distributions for all C-test parts. As a reference point, we determined the overall distribution of C-test scores for both C-tests of the dataset, which was found to have a curved shape with a maximum more or less at median accuracy scores. Figure 3.7 shows that out of the six parts of each C-test, only the second, third and fourth part reflect this form while the other three parts have monotonically increasing distributions. The meta information available for the C-tests confirms that these parts do indeed not provide representative data: The first part constitutes an example exercise with which the test instructors explained the task. The last two parts were attempted by

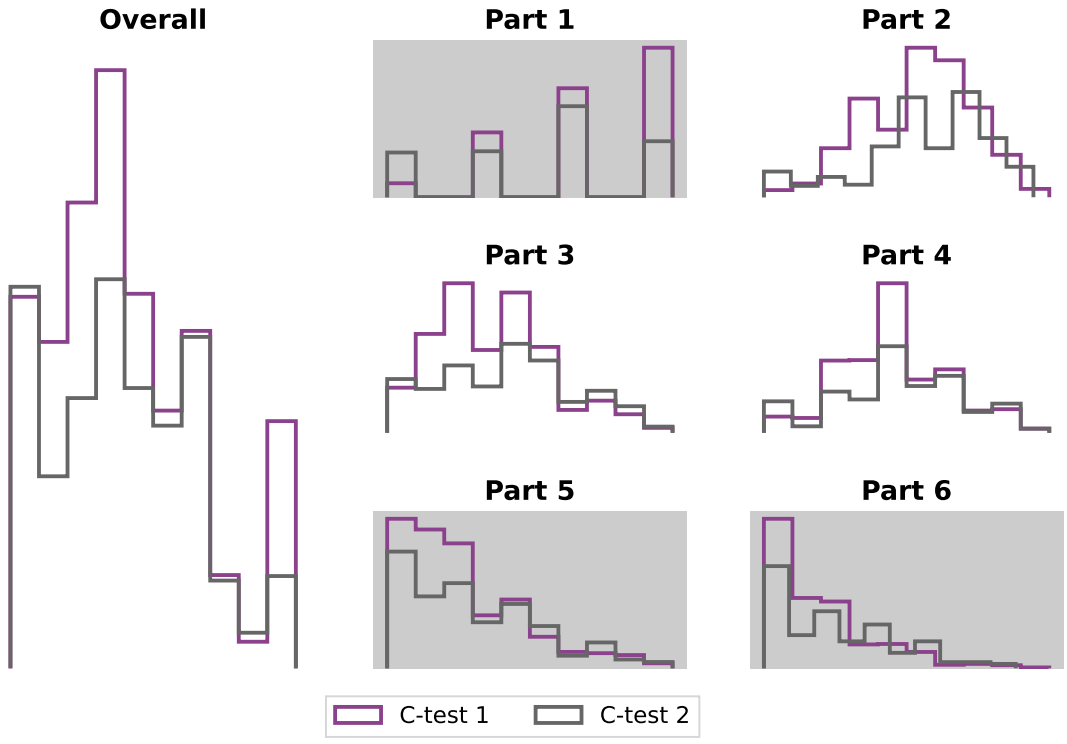


Figure 3.7: Distributions of C-test scores.
The scores for all C-test parts combined are distributed along a curved shape with maxima around the median of the accuracy scores (x-axis). This shape is similar for parts 2 to 4, but not reflected in parts 1, 5 and 6.

only a small number of learners who managed to complete them within the given time frame. Since higher language proficiency cannot be disentangled from better typing skills in the dataset, greater numbers of answered test items cannot reliably be attributed to higher language proficiency. In the subsequent evaluations, we therefore only used the answers of the second, third, and fourth part.

In order to verify the validity of adjusting exercise difficulty to C-test scores, we next assessed whether the C-tests effectively yielded test results that varied across learners. Only then can C-tests be indicative of varied performance on practice exercises. In order to verify that our dataset covers learners of diverse proficiency levels, we determined the range of accuracy scores obtained on the C-tests across all learners. The values were similar for both C-tests with minimum scores of .00 and the highest observed accuracy at .62. Excluding the learners who did not correctly answer any item ($acc = .00$), the lowest score amounted to .01. The study participants thus indeed comprised learners of very low English proficiency who nevertheless made an effort to complete the C-tests. The dataset therefore covers learners with overall English language proficiencies ranging from very weak to moderately strong.

After establishing the validity of the C-tests in themselves, we can effectively use them to operationalize a learner's general language proficiency. This learner characteristic can only impact exercise difficulty if there is any relationship between the operationalizations of both. In order to determine whether this is the case for our dataset, we calculated Pearson's correlation ρ between the learners' performance on the C-tests (operationalizing general language proficiency) and that on practice exercises (operationalizing exercise difficulty). C-test performance was defined as the mean accuracy score of C-test parts 2 to 4. Practice performance was defined as the ratio of correct first attempts of all submitted exercises to the number of submitted actionable elements. In addition to global correlation, we also looked at the correlations within the subsets representing combinations of exercise types and `learning targets`. This allowed us to determine whether general language proficiency impacts exercise difficulty only for certain exercise types or `learning targets`. Table 3.3 gives an overview of the results. For the first C-test, the Pearson correlation reveals only a weak relationship between C-test accuracy and practice accuracy ($\rho = .2811$). It does not increase when only considering key exercises

Exercise set	ρ_{c1}	ρ_{c2}
All	.2811	.4070
Key	.2821	.3641
Co-text sensitive	.2887	.3882
Fill-in-the-Blanks – *Simple past vs. Pres. perf.*	.4688	.3890
Single Choice – *Conditionals*	.4101	.4392

Table 3.3: Correlations of the C-tests with practice performance.
Pearson's correlations of the C-test from the post-test with practice performance (ρ_{c2}) is generally higher than that of the C-test from the pre-test with practice performance (ρ_{c1}), although there is considerable variation across exercise subsets.

($\rho = .2821$), and only marginally for co-text sensitive exercises ($\rho = .2887$). This suggests that the data for the overall exercise pool reflects the performance of our target group and that general language proficiency is not more relevant in exercises that require processing of the text material. When looking at the different `learning targets` and exercise types separately, correlations are higher for a number of subgroups covering almost all exercise types and `learning targets`. The highest – although weak – correlation ($\rho = .4688$) is for Fill-in-the-Blanks exercises on *Simple past vs. Present perfect*. The Gap-filling exercise types Fill-in-the-Blanks and Single Choice, as well as the occasional Jumbled Sentence exercise type, have the highest correlations for a number of `learning targets`. Of these, there is no pattern indicating that any `learning target` generally has higher correlations between C-test and practice performance than others.

Interestingly, the scores of the second C-test correlate much stronger with the learners' practice performance, although the relationship is still weak ($\rho = .4070$). When looking at the subsets, the pattern is similar to that with the first C-test: Key exercises ($\rho = .3641$) and co-text sensitive exercises ($\rho = .3882$) have comparable correlations. The combination of exercise type and `learning target` with strongest correlation between the first C-test and practice performance again shows a weak correlation for the second C-test ($\rho = .3890$), and is only surpassed by one other combination. The correlation between performance on this C-test and practice performance is highest for Single Choice exercises on *Conditionals* ($\rho = .4392$). The patterns for specific exercise types and `learning targets` are similar to those from the first C-test.

In order to better compare the relevance of the two C-tests with respect to their

predictive power for practice performance, we generated a partial dependence plot based on an AdaBoost classifier[4] trained to predict from the C-test scores whether an actionable element is answered correctly. As the probability increases, the coloring turns from light to dark purple. For the plot given in Figure 3.8, the color changes progressively on the vertical axis representing the second C-test, but not on the horizontal axis representing the first C-test. This illustrates that while for the second C-test, the probability of a learner answering an element correctly increases with higher test scores, this is not the case for the first C-test. The C-test administered upon the learners' first interaction with the system did thus not yield scores that can be used to initialize a learner's general language proficiency.

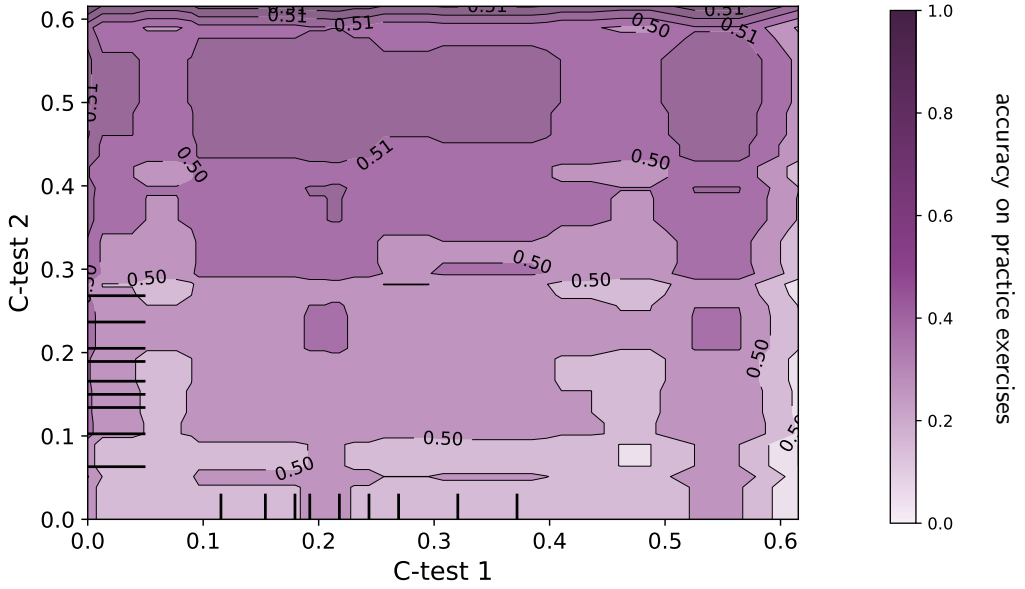


Figure 3.8: Partial dependence plot for the C-tests when predicting the correctness of a learner's answer.
For C-test 2, the colors progress from dark to light on the y-axis, indicating increasing probability of correctly answering a practice element with increasing C-test scores. For C-test 1, there is no progressive change in colors on the x-axis, thus not showing a relationship between performance on the first C-test and practice exercises.

In the next analyses, we considered not only learners' C-test scores but also the exercises' co-text complexity values. We trained various classifiers to predict practice performance in terms of whether a learner answers an actionable element correctly.

[4]The classification was based on the `scikit-learn` implementation for Python and used a train-test split of 75-25%.

The classifiers encompassed a Decision Tree, a Random Forest, and an AdaBoost classifier from the Python `scikit-learn` library, all of which provide predictor rankings. For the baseline model, we used only simple exercise features as predictors such as the exercise type, the number of tokens in the target answer, and the number of other targets in the exercise. We then conducted an ablation study for various subsets of the data in order to determine which features most impact model performance of a classifier predicting a learner's performance on an exercise. The full feature set includes IRT difficulty, co-text complexity, and C-test scores of the first and second C-test. While IRT difficulty scores can be expected to be the most indicative exercise parameter in terms of practice performance, this feature – similar to learner model metrics for new learners – is unknown for new exercises. In the ablation study, we therefore successively excluded features of interest from the feature set in order to determine whether they may inform learner performance predictions. All features were encoded as Integer values; features that were not applicable for an exercise in this case received the value *zero*. We determined precision, recall, and F_1-scores as performance metrics for all model variants in order to evaluate whether adding certain features improves model performance. Accuracy does not constitute a suitable metric for our dataset as it is imbalanced with respect to the number of correct and incorrect submissions. Precision, recall and F_1-scores were comparable for all three classifiers, although the AdaBoost classifier slightly outperformed the others for most subsets and input feature combinations. For the entire dataset, precision and recall were almost always identical and mirrored the F_1-scores. For our purposes, recall and precision are equally important in a model's performance. We therefore report only F_1-scores, which combine precision an recall scores, of the AdaBoost classifier. The scores are summarized in Table 3.4.

The baseline model already achieved a high F_1-score of .7238, which increased to .7665 when adding co-text complexity scores, IRT difficulties, and C-test scores of both C-tests. This model including the full feature set achieved the second highest performance. Removing the scores of the second C-test (c_2) from the full feature set only slightly decreased model performance ($F_1 = .7659$) and removing the scores of the first C-test (c_1) even resulted in a slight increase in model performance ($F_1 = .7667$). Removing both C-test scores also resulted in a slight decrease in model performance ($F_1 = .7615$). When removing co-text complexity (txt) from the full feature set, there was almost no decrease in performance ($F_1 = .7660$). Ex-

Predictors / Set of exercises	all	-*b*	-txt	-c_1	-c_2	-c_1-c_2	-txt -c_1-c_2	base-line	μ	σ
All	.7665	.7256	.7660	.7659	.7667	.7615	.7597	.7238	.7545	.0186
Core	.7775	.7630	.7798	.7785	.7756	.7612	.7612	.7503	.7684	.0109
Co-text sensitive	.7498	.7126	.7505	.7512	.7407	.7372	.7393	.7056	.7359	.0175
b_{low}	.9475	.9472	.9477	.9469	.9469	.9450	.9450	.9450	.9464	.0012
$b_{intermediate}$	.7441	.7433	.7450	.7429	.7416	.7458	.7454	.7374	.7432	.0027
b_{high}	.8509	.8560	.8502	.8502	.8531	.8538	.8538	.8553	.8529	.0023
txt_{low}	.6837	.6837	.6837	.6531	.6939	.6122	.6122	.6122	.6543	.0368
$txt_{intermediate}$	.7677	.7677	.7677	.7475	.7475	.7172	.7172	.7172	.7437	.0235
txt_{high}	.8305	.8390	.8305	.8475	.8559	.8729	.8729	.8729	.8528	.0186

Table 3.4: F_1-scores of classifiers predicting exercise correctness.
The full feature set encompassed simple exercise features and additional features comprising *IRT difficulty (b)*, *scores of the first C-test (c_1)*, *scores of the second C-test (c_2)*, and *co-text complexity (txt)*. In the ablation study, the additional features were successively excluded from the input feature set. The model that excluded IRT difficulty from the feature set performed almost as low as the baseline model.

cluding C-test scores and co-text complexity from the feature set also only slightly decreased model performance ($F_1 = .7579$), indicating that these features do not hold much predictive power on their own, nor in combination with each other. However, removing IRT difficulties from the full feature set resulted in low model performance ($F_1 = .7256$) that was only marginally higher than that of the baseline model ($F_1 = .7238$). Thus, C-test scores and co-text complexity cannot successfully replace IRT difficulty scores when predicting a learner's performance on an exercise.

In addition, we examined the feature importances provided by the classifiers, illustrated in the heatmaps in Figure 3.9. The feature importances allow to estimate the relevance of the individual features to the models' predictions. Not surprisingly, IRT difficulty is the overall most predictive feature for practice performance. The feature rankings for the analyzed features – IRT difficulty, co-text complexity, and C-test scores – are similar for all subsets of exercises in terms of relative rankings, although absolute values vary. Differences in the rankings concern mostly the simple exercise features of the baseline model and are quite pronounced between the different exercise types. In terms of exercise types, the relevance of C-test scores also varies considerably from one exercise type to another. According to the predictor rankings, general language proficiency is highly relevant – even more relevant than IRT difficulty – with Memory and Categorization exercises, and less so with Jumbled Sentence, Single Choice, Short Answer, Mark-the-Words, and particularly Fill-in-the-Blanks exercises. This holds for both C-tests.

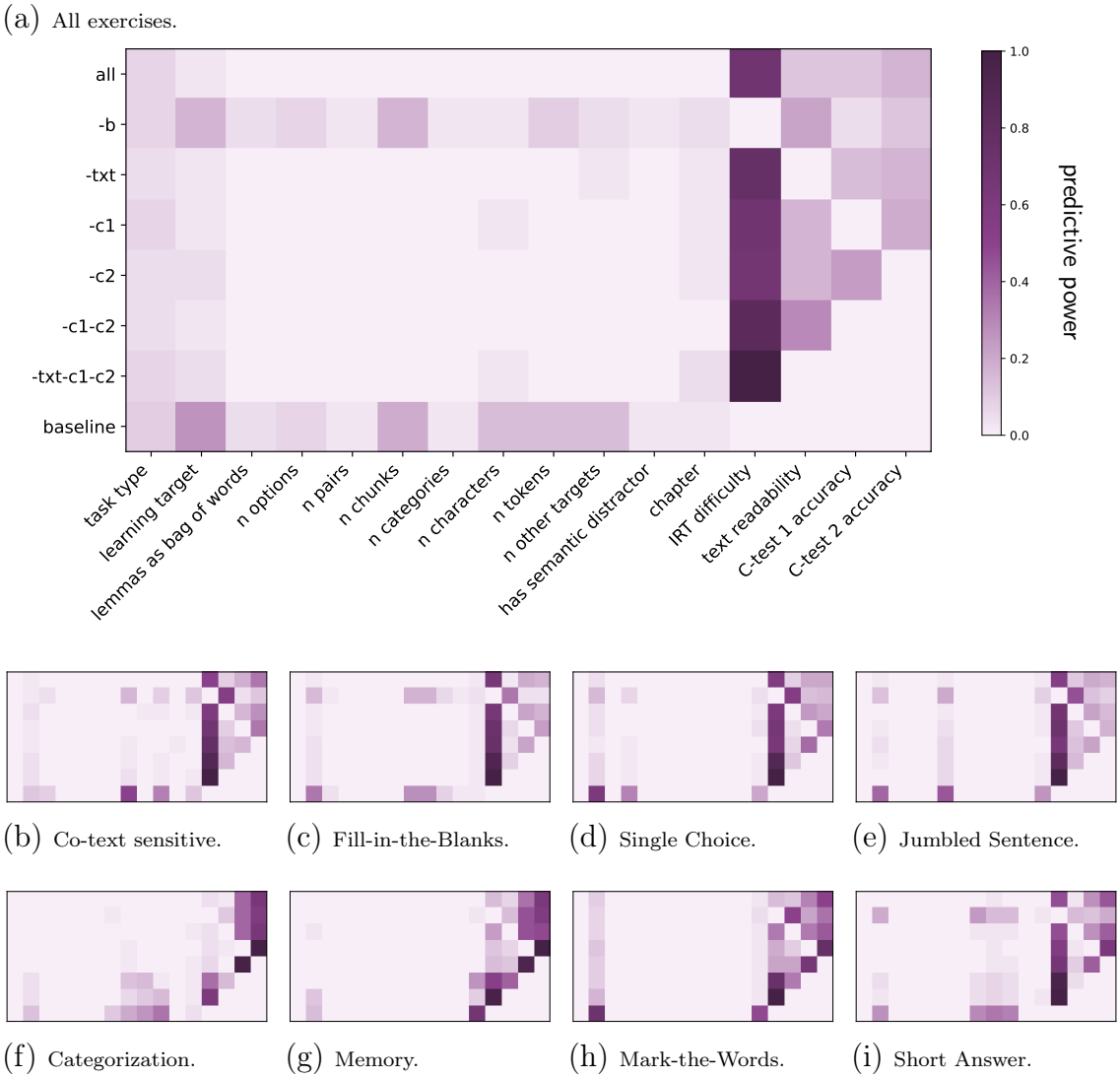


Figure 3.9: Feature importances in a classifier predicting practice performance. Darker colors indicate greater predictive power of the feature on the x-axis for the predictor set on the y-axis. C-test scores are particularly predictive with Categorization and Memory exercises.

When looking at the development of the learners' C-test scores from one test timepoint to the other, the scatter plot given in Figure 3.10 reveals that for a considerable number of learners, located in the shaded area underneath the first bisector, the scores do not show the expected increase, but instead decrease over time. This also results in an only moderate correlation ($\rho = .5260$) between the two tests. Considering the previous findings that the scores of the second C-test correlate better with practice performance than those of the first C-test for all `learning targets`, this

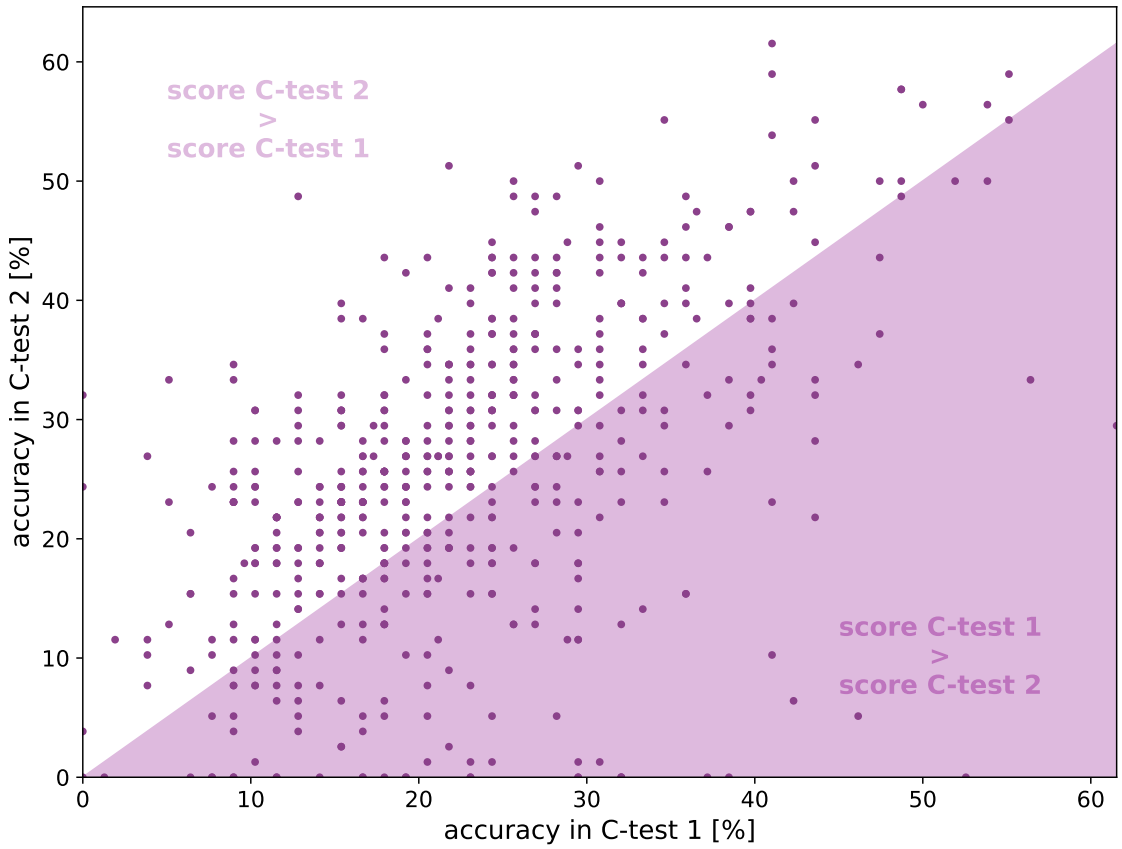


Figure 3.10: Development of C-test scores between test timepoints. Many learners performed better on the second C-test than on the first C-test (white area). However, a considerable number of learners performed worse on the second C-test (purple area).

could indicate that C-tests taken during a learner's first interaction with the system are not entirely representative of their general language proficiency, possibly due to the novelty of the system and the test setup. A tentative conclusion assumes that tests are more representative if learners are already familiar with the test platform. If this is the case, then administering a C-test after a short period of familiarization might yield the desired scores after all. In this case, the question remains whether this score can be used to initialize the learner model for new `learning targets` even after longer periods of time. As a learner's general language proficiency may change during their involvement with the system, the validity of the initially elicited proficiency score might decrease over time. In order to determine whether this is the case for our learner population, we trained an AdaBoost classifier individually

for each of the four learning units of the ILTS to predict a learner's performance on an exercise, using the C-test scores and co-text complexity as input. The learning unit index represents an exercise's relative practice timepoint. The development of the Gini importance of the two C-tests as input features relative to each other over the sequence of succeeding learning units can therefore give insights into whether recency of a C-test influences the predictive power of general language proficiency for exercise difficulty.

Learning unit	**Gini importance**			**Relative impact**
	c1	c2	c1-c2	
1	.16	.12	.04	$1 > 2$
2	.04	.10	-.06	$2 > 1$
3	.02	.08	-.06	$2 > 1$
4	.14	.10	.04	$1 > 2$

Table 3.5: Predictiveness of the C-tests.
The predictive power of the first C-test (c1) is higher than that of the second C-test (c2) for the first and last learning unit, but lower for the second and third learning unit.

While the classifier's feature rankings – outlined in Table 3.5 for the entire dataset – indicate varying priority of one of the two C-tests over the other, the importance of none of the C-tests increases monotonically with its temporal proximity to the practice unit. This is similar for all data subsets as illustrated in Figure 3.11, which displays the difference in feature importances between the first and second C-test depending on the learning unit. Monotonically decreasing lines would indicate that the first C-test loses importance with later learning units while the second C-test's importance increases. However, this is not the case for any of the subsets. It is also not the case that the predictive power of the second C-test is always higher than that of the first C-test in the last `learning target`. It is therefore unlikely that the unreliability of the first C-test is the reason for the non-monotonic development of the relevance of the C-tests. From this we conclude that the test timepoint does not seem to play a substantial role in the predictive power of C-tests, indicating that C-tests do not lose validity over time, at least not within the course of a school year.

Overall, these results indicate that C-test scores have no or only weak relationships with performance on practice exercises. Especially for low- and high-difficulty exercises, the relationship of general language proficiency with practice performance is very weak. C-tests are, however, more predictive of a learner's performance on

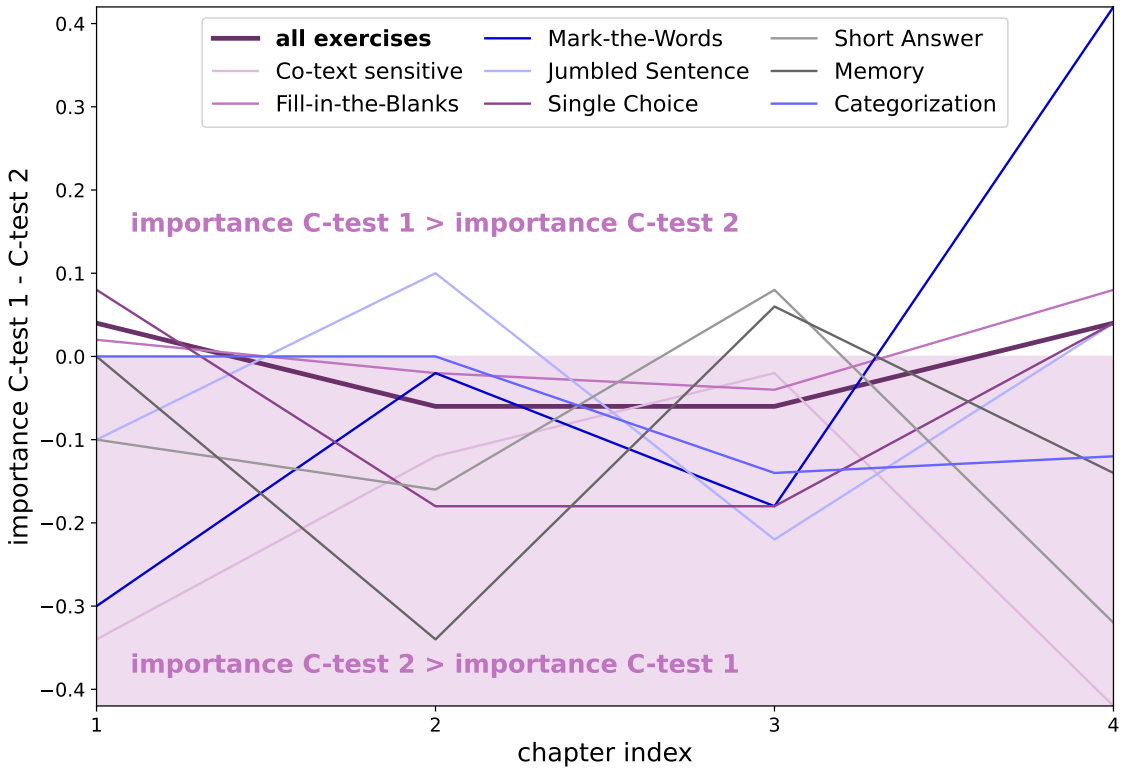


Figure 3.11: Temporal development of the predictive power of the C-test scores. The predictive power of C-test 2 relative to C-test 1 does not increase monotonically with increasing temporal proximity of the practice exercises to the test administration timepoint.

practice exercises when taken after a period of familiarization with the system. Although the C-test scores of the available dataset cover a considerable range, learners in our setup might still constitute a more homogeneous group than in other ILTSs where learners do not follow the same curriculum and workbook. In such systems, the suitability of C-tests to initialize learner proficiency may be re-considered.

Since the initial assessment through C-tests turned out not to be very reliable, C-tests cannot help to overcome the cold-start problem. Macro-adaptive systems therefore need to rely on alternative placement tests – potentially administered at the beginning of each `learning target` as introductory exercises (Theis and Wackermann, 2020) –, or should be optimized to quickly react to updates of the learner model.

3.2 Knowledge Transfer and Proceduralization through Variability

In order to be able to apply knowledge gained in practice sessions to new contexts, practice must be transfer-appropriate. The definition of KCs largely depends on the transferability of the knowledge they represent. Each KC can cover the scope within which transfer is possible. Where transfer is not possible, a distinct KC is required. Different strands of research have analyzed characteristics of and theories about transfer-appropriate practice. The beneficial effects of varied practice for transfer and generalization of knowledge have been extensively discussed and examined in multiple domains in the context of implicit learning. Empirical research is, however, scarce with respect to whether these results apply to school contexts, where grammar rules are introduced before a phase of explicit practice. In fact, variability is often neglected in instructional settings, where the primary goal is to prepare the students for exams based on the limited set of variants assessed in the tests (Wahlheim et al., 2012). We therefore focus on this aspect in RQ 3 and consider implications for proceduralization in RQ 4.

3.2.1 Theoretical Grounding of Transfer-appropriate Practice

Depending on the field of research, terminology revolving around transfer as well as the factors and mechanisms presumed to contribute to this transfer differ.

In the field of applied linguistics, (Stratton, 2022) illustrate how **intentional learning** can be beneficial for transfer across languages. Hulstijn (2012) provides similar findings for transfer across tasks. He suggests that learners store knowledge in a transferable form if made aware of the need for transfer before practice.

Transfer-Appropriate Processing (TAP) focuses on cognitive processes involved in retrieving items (Lightbrown, 2007). Originated in cognitive psychology, it is based on the assumption that retrieving an item from memory is more successful if the cognitive processes involved in the retrieval process are similar to those involved in the encoding process when learning the item. It should be noted

that similarity hereby relates to cognitive processes rather than to surface features (Nikouee, 2021). The more contexts an item has been encountered in before, the greater the pool of cues becomes that a learner can consult when retrieving the item. On a similar notion, skill acquisition theory postulates that learning is skill specific so that transfer of procedural knowledge between tasks is very limited (DeKeyser, 2018). In language learning, knowledge thus does not automatically transfer from comprehension practice to production and vice versa (DeKeyser, 1997). Procedural knowledge is less flexible and thus less transferable to new contexts than declarative knowledge (DeKeyser, 2018). Yet, the primary goal of practice in language learning is proceduralization of declarative knowledge. Depending on the level of abstraction, representations of knowledge in memory allow for near transfer of procedural knowledge to very similar contexts, or far transfer via declarative knowledge to less similar contexts. In order to be able to rely on declarative knowledge, however, transfer-appropriate practice needs to make learners aware of the declarative knowledge. Study results suggest that productively acquired knowledge transfers more easily to receptive knowledge than the other way round, although varied practice is most effective (Webb, 2009).

In their review of the effect of variable practice, Kim et al. (2021) attribute this to the contextual interference effect grounded in the benefits of **interleaved practice** compared to blocked practice. This strand of cognitive psychology has mainly focused on motor skill practice (Schorn and Knowlton, 2021). While the forgetting-reconstruction position grounds the benefit of interleaved practice in the need to constantly re-construct an action plan, the elaboration position attributes the benefit to more elaborate representations of the practiced skill as the learner assesses the differences and similarities of the varied skills. This corresponds to the position the Varied Practice Model represents, in that variable practice fosters the identification of relevant, shared information, as well as discriminative features (Apfelbaum et al., 2019). In Apfelbaum et al.'s (2019) studies, varied practice material and contexts resulted in better retention of knowledge and higher recall flexibility of reading skills. More recent research sees the reason for the beneficial effect in increased attention, decision making and response demands through interleaved practice that result in more resilient representations of the practiced content Kim et al. (2021). However, the benefit of interleaved practice is controversial in Second Language Acquisition (SLA) research, where different studies have yielded contradictory results (DeKeyser,

2018). While no or even negative effects of interleaved practice were found in studies using oral tasks (Carpenter and Mueller, 2013), Nakata and Suzuki (2019) found decreased accuracy during practice, but increased learning gains through interleaved practice in particular for learners of lower proficiency. Kim et al. (2021) stress that all experiments so far combined variable practice with interleaved practice and used the two terms interchangeably. They highlight that it is detrimental to establish the effect of variable practice targeting varied content in itself, regardless of scheduling.

Characteristics of transfer-appropriate practice have thus been identified to consist in transfer-awareness and variability. In terms of variability, the focus of TAP as well as of blocked vs. interleaved practice lies on the activity (Suzuki, 2022). Grammar should be learned in different task types and skill modalities in order to support transfer. Merriënboervan (1997, p. 188) suggests that varying any task dimension that is not the focus of learning encourages generalization of knowledge. In the same vein, Suzuki (2023) encompasses, among others, linguistic structures in the dimensions impacting transfer.

Raviv et al. (2022) outline that this kind of variability has been the focus of a number of studies in different domains including categorization, motor learning, and also language acquisition. Here, the authors of the review article identified research demonstrating the positive effects of variable practice in several strands such as phonology, vocabulary acquisition, and grammar learning. They do, however, call for caution with respect to the employed terminology. While much research revolves around *variability*, this term is used to refer to the number of practiced examples, the type frequency, situational diversity, and scheduling. As we have seen with contextual and scheduling variability, there may be links between these different types of variability. In this thesis, however, we focus on variability in terms of type frequency, which indicates the number of distinct examples in the learning input.

Empirical findings suggest that variability in the practice material slows down learning, but increases long-term learning gains by improving generalization of the practiced knowledge (Raviv et al., 2022). Using artificial languages to investigate the effect of variable practice on the acquisition of non-adjacent dependencies, Gómez (2002) found a positive effect for both adults and children of 18 months. The author concludes that the learners resorted to more distant clues when variability in the more local clues proved them to be unreliable. Also employing an artificial language,

Wonnacott et al. (2012) identified a positive effect of variable practice on the robustness and generalization of syntactic constructions for five-year-olds. Eidsvåg et al.'s (2015) study on Russian noun gender also found variable practice to be beneficial, which the authors attribute to an increased saliency of the invariant properties in the input when other properties are more varied. In the context of category learning, Perry et al. (2010) taught multiple words to 18-month-old children, finding that low within-category variability resulted in over-generalization, whereas high variability resulted in the correct discrimination of the different categories. Several studies confirmed both a beneficial effect of low variability for practiced examples and a positive effect of high variability for novel items, thus indicating better generalization. Examples of such studies investigated the effect of variability on children learning the semantics of Japanese postpositions (Viviani et al., 2022), on young school children when teaching to infer underlying structural patterns of verb-argument constructions (Schulz, 2024), on the discrimination of English dialects in the field of phonological training when comparing a group that practiced with a single speaker per dialect to a group that practiced with multiple speakers per dialect (Clopper and Pisoni, 2004), and on the classification of birds (Wahlheim et al., 2012).

Theoretical grounding supporting the beneficial effect of variability can be found in First Language Acquisition: The Competition Model framework advocates that language users rely on cues to determine a mapping between a form and its function (MacWhinney, 1997). Cue competition allows learners to identify features that help to distinguish between forms. It thus serves to develop more generalized concepts based on combinations of cues that are relevant to identify the function of a form. The more cues a learner has encountered during practice, the better they can identify those that reliably predict the function (Schulz, 2024). The mapping between form and function in addition relies on the strengths of the cues. Strength of a cue is in turn determined by its availability and frequency (Stafford et al., 2012). When frequency of exposure is low, the learner freely varies the forms mapped to the competing cues (MacWhinney, 1987). A lack of variation in linguistic forms of a linguistic construction might provide insufficient or inadequate opportunities for cue competition during practice. This, in turn, does not allow learners to differentiate between relevant and irrelevant cues and might thus result in incorrect or incomplete mappings of forms to functions (Raviv et al., 2022). Learners' mappings of cues to linguistic forms might then lead to incorrect rule application for unknown forms.

The Competition Model in addition suggests that varied practice can result in high-level mappings that rely on shared features of multiple subordinate forms instead of on features specific to the subordinate form. To illustrate this, consider the exercise in Example 3.1a. If the learner has only seen a sentence like Example 3.1b during practice, the only available cue for the formation of the verb constitutes in that wh-questions need do-support. If, on the other hand, the learner has only seen a question like Example 3.1c, the available cue would indicate that wh-questions require a conjugated verb. Obviously, neither of the two constitute complete cue mappings, even though the learner's production would be correct in the second case. This representation would, however, not suffice to correctly answer the exercise given in Example 3.1d. Having seen both sentences, on the other hand, introduces cue competition that allows a learner to identify additional cues in the specific question word and the presence or absence of a subject in the question. If the Competition Model can be applied at the level of linguistic forms, such mappings allow learners to generalize – and thus to transfer – knowledge of a linguistic form to other forms of a linguistic construction.

3.1 a. What _____ (dry) in the sun?
b. What does the caretaker repair?
c. Who eats all the cake?
d. Who _____ (Peter (*subj.*), invite) to the party?

Alternative theoretical grounding can be found in cognitive science research in the context of Schema Theory. According to this theory, learners form schemas of categories by abstracting stimuli from members of the category into more general concepts (Schmidt, 1975). Within this framework, however, narrow reading, in contrast to varied reading, has been argued to enable students to develop background knowledge that promotes text comprehension (Carrell, 1984). Yet, studies on motor skill practice have provided positive evidence showing that varied practice fosters transfer of knowledge (Kerr and Booth, 1978). According to connectionist models, the connections between features and categories grow stronger with each practice item (Ellis, 2002).

However, insights from various domains on the effectiveness of high-variability training call for caution. Contrary to initial findings, an increasing number of studies

are producing results indicating that high variation in training of classes with internal variability does not result in better generalization than low-variability training (Brekelmans et al., 2022; Hu and Nosofsky, 2024). Most studies focused on speaking variations in phonetic training (e.g., Brekelmans et al., 2022). However, dot patterns (Hu and Nosofsky, 2024), different fonts for the recognition of Chinese characters (Enns, 2018), different temporal durations for non-word sentences in a speech-motor task (Kaipa et al., 2017), and varied nouns in a word learning task (Brown et al., 2022) also yielded negative results.

Reasons for these discrepancies in empirical findings may be grounded in differences in learner characteristics such as previous knowledge and in whether variability involves the target construct or non-target properties of the input (Raviv et al., 2022). Although Brown et al.'s (2016) study included only 7-year-old children, they acknowledge that their findings of a negative effect of variability on generalization when teaching gender classes might be restricted to the initial stages of learning. The authors assume that the negative effects of variability on generalization stem from their learners being overwhelmed by the increased complexity of the input. Likourezos et al.'s (2019) study in the domain of Maths included adults with varying prior knowledge. The authors found that while low variability was beneficial for learners with little prior knowledge, learners with high prior knowledge benefited from high variability during practice. They suggest that learners with low prior knowledge did not have sufficient working memory capacity to meet the increased cognitive load induced by high variability. Madlener (2016, 2015) obtained similar results in experiments on learning syntactic rules in listening comprehension tasks with adult learners. She concludes that initial low variability and the thus increased repetition rate of each item allows to strengthen and automatize the newly acquired knowledge and may facilitate pattern detection through structural alignment, whereas high variability confuses them. In later stages of learning, on the other hand, high variability enables learners to generalize these patterns and to develop productivity in applying them as it provides sufficiently varied evidence. The results of these studies are in line with Wiener et al.'s (2020) findings that high-variability training is beneficial for learning Chinese tone productions only in combination with prior explicit instruction. Concerning the type of variability, Raviv et al. (2022) suggest that learners with low prior knowledge benefit only from variability that does not involve the target construction, whereas more experienced learners benefit from

both types of variability. An additional factor potentially impacting the effect of variability concerns the degree of variability. While Twomey et al. (2014) found a beneficial effect for moderately varied practice over low variability, highly varied input hindered the correct categorization of even the practiced vocabulary items. In a similar vein, related research has found positive evidence for a beneficial effect of skewed practice, which uses few, prototypical items (e.g., Casenhiser and Goldberg, 2005). This strand of research focuses, however, on the initial acquisition and has paid little attention to generalization (Goldberg et al., 2004). In addition, Madlener (2016) points out that the effect of skewed practice is conditioned by the type of construction, the practice context, and the amount of practice.

According to theories from cognitive science, variability increases cognitive load (Merriënboervan , 1997, p. 189). However, since the cognitive resources are thereby concentrated on the learning process in our exercises, it provides, in principle, the means to maximize learning effects (Corbalan et al., 2006). It should, however, be ensured that learners are not cognitively overloaded. This directly connects to the DDF introduced in Section 3.1. Variability only constitutes a parameter of desirable context-related difficulty if it improves long-term performance and transfer. In conclusion to the findings of varying effects of variable practice depending on the learner's previous knowledge, researchers have therefore proposed to keep variability low during initial practice and increase variability in later stages of learning when the learner is familiar with the target construction (Likourezos et al., 2019; Bybee, 2008). Wiener et al.'s (2020) study constitutes a rare example investigating whether providing explicit instruction on the patterns before practice results in sufficient prior knowledge to enable all learners to benefit from variable practice. To the best of our knowledge, this effect has not yet been examined in the context of written grammar exercises.

While the effect of variability on learning and on the generalization of knowledge has been studied extensively in the context of implicit learning, settings where instructions and learning are explicit have received far less attention. It therefore needs to be determined at what, if any, linguistic level and for what learner population linguistically varied practice fosters transfer of knowledge across linguistic variations of a language form when provided prior explicit rule explanations.

3.2.2 Knowledge Transfer Across Linguistic Forms

When determining KCs in the domain model it is straightforward to create a new KC for each linguistic form that relies on grammar rules that are not targeted by any other KC. For instance, conditional type 2 sentences require two distinct KCs for different clause orders, the first KC practicing conditional sentences with the main clause before the if-clause such as Example 3.2a, and the second KC practicing sentences with the if-clause first such as Example 3.2b.

3.2 a. We will go hiking if the sun shines tomorrow.
b. If the sun shines tomorrow, we will go hiking.

The decision whether a form justifies a distinct KC is less clear when it comes to lexical variations of closed types. For irregular verbs where the formation needs to be practiced for each verb individually similar to vocabulary practice, it makes sense to attribute a KC to each verb. Modal verbs, on the other hand, all follow the same rules for sentence formation and do not require inflection, but the challenge lies in deciding whether a verb constitutes a modal verb. On a more abstract level, modal verbs support different interpretations depending on their modality. While epistemic modality expresses possibility, deontic modality conveys an obligation (Winiharti, 2012). Similar to modals, question words elicit the same syntactic structure and differ on the lexical level. In addition, learners need to decide which question word is required. Question words also introduce an additional challenge in that the question words *who* and *what* support both object and subject questions. In addition, question formation differs depending on the verb type in terms of do-support verb, modal, or the copula *be*. Since wh-questions thus have multiple distinguishing properties at different levels, we take a closer look at knowledge transfer across these properties.

The distinguishing properties of wh-questions include the specific question word and the required syntax depending on whether it is a subject or an object question and includes a modal verb, the verb *be*, or another verb. As illustrated in Figure 3.12, each combination of distinguishing properties, henceforth referred to as `property`, in turn comes with a range of possible concrete instantiations, which we will call `linguistic constructs`. We consider `linguistic constructs` to be linguistic

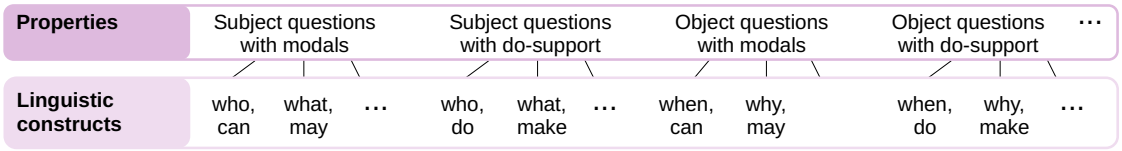


Figure 3.12: Instantiations of properties of linguistic constructs.
Each `property` may be instantiated by various combinations of `linguistic constructs`. Each combination of `linguistic constructs` is associated with a single `property`.

entities that have certain linguistic annotations. If construction rules are similar for all `linguistic constructs` of a `property`, learners need to first determine the properties of an instantiation and then apply the correct grammar rule to form a question. Since the grammar rules to construct the `linguistic constructs` are independent of the concrete modal verb or question word it is, in principle, possible to understand and produce a never seen `linguistic construct` that differs from practiced constructs in the specific verb and the question word, if a learner has seen other instantiations of the `property` before.

If, on the other hand, a distinguishing property requires application of a different grammar rule, we hypothesized that transfer does not occur. This is for example the case with subject and object questions, which require different syntax for questions with do-support verbs. We therefore analyzed for multiple linguistic levels including syntax, lexis, and semantics whether linguistic knowledge transfers across wh-questions that differ on that level.

Data In order to collect data with the aim of answering RQs 3 and 4, we conducted a small-scale RCT in two German Gymnasium schools. The study design received governmental as well as ethical approval (see Appendix A.1). The participants encompassed 7^{th} and 8^{th} grade students that were recruited through teachers involved in the AI2Teach project. The scale of the study was determined prior to recruitment based on a power analysis. In terms of related research, Schulz (2024) reports an effect size of Cohen's d=.87 on a post-test without delay. For the significance level $\alpha = .05$, a power of .8, and an effect size of d=.87, this results in a necessary sample size of n=44. Expecting that not all students would consent to participating in the study and that some data would have to be excluded from the analyses, we recruited 5 school classes of approximately 25 students in order

for the collected data to meet the requirements of the power analysis. The study participants providing written consent for participation in the study and use of the collected data, signed by themselves as well as by a legal guardian, amounted to 79 (62 7^{th} grade students, 17 8^{th} grade students). Poor internet connection in the schools accounted for some additional data loss so that not all learners could be considered in all data analyses. All study participants were randomly assigned to one of two study conditions. Randomization was applied within each class. The study took place in the schools during regular English classes of 90 minutes. Each participating class offered one such session towards the end of term.

The grammar exercises focused on the linguistic learning target of wh-questions. Within this learning target, we distinguished between the four `properties` *questions with ' be'*, *subject questions*, *questions with modals*, and *questions with do-support*. The participants completed all parts of the study on tablets using a version of the FeedBook system specially prepared for the study. The material consisted of six parts: (1) pre-test, (2) grammatical explanations of the learning target, (3) interim test, (4) practice exercises, (5) working memory test, and (6) post-test exercises. Study participants received corrective feedback while working on the practice exercises, but no feedback while doing the test exercises. All six parts were conducted sequentially with strict time limitations. These were enforced by the test conductors who introduced each part and made sure that all participants navigated to the next part. In addition, the implementation made each part of the study accessible only

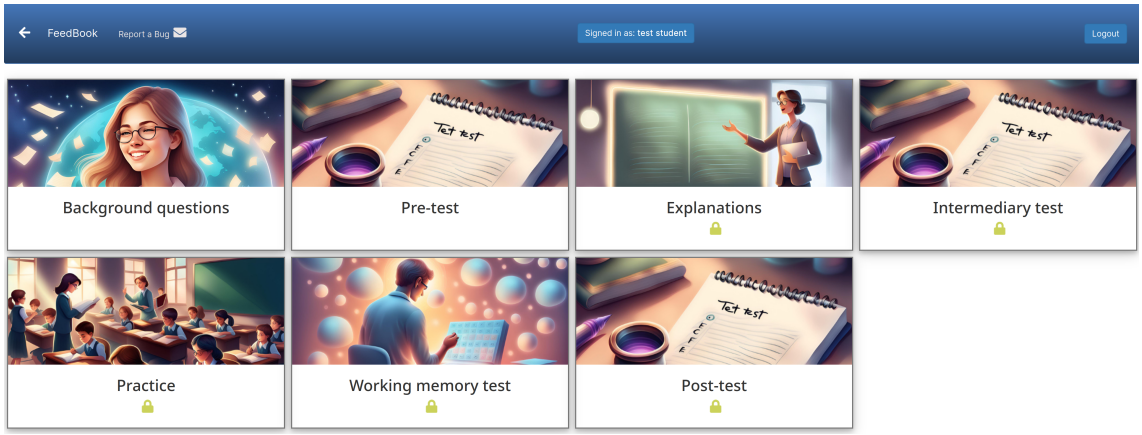


Figure 3.13: User interface of the FeedBook used in the variability study. Submitting exercises in a part of the study automatically locked previous parts and unlocked the succeeding part.

once the learner had started on the previous part. Preceding parts were locked at the same time. Figure 3.13 illustrates how this was reflected visually in the GUI through lock symbols. Following Ellis's (2005) setup, we determined time allocated for each part of the study based on time needed by a proficient speaker, with an additional 20% recommended by the author to account for slower processing by learners. Durations granted for the individual parts are given the Table 3.6. In order to be able to attribute learning gains to proceduralization, during the post-test participants should not be able to rely on knowledge retained in the working memory. According to Newlin and Moss (2020), working memory effects in the post-test are removed if participants are prevented from rehearsing the stimuli or responses, regardless of the length of the delay between practice and post-test. We therefore placed the cognitively challenging working memory test, which prevented the participants from rehearsing content from the practice phase, in-between practice and post-test.

Part	**Number of instances**	**Duration [minutes]**
Organizational matters		**10**
Introduction and explanations		10
Pre-test		**10**
Elicited questions	4	10
Reading of explanations		**5**
Interim test		**10**
Elicited questions	4	10
Practice exercises		**15**
Gap-filling	30	15
Working memory test		**10**
N-back task	3	10
Post-test		**25**
Elicited questions	4	10
Gap-filling	30	15
Total		**85**

Table 3.6: Schedule for the variability study.
The time limits were enforced by the test conductors by asking the students to navigate to the next part.

The pre-test covered only elicited questions. With them, we aimed to determine varied, pro-active usage of the constructs. In such eliciting exercises as suggested by Mak and Xiao (2018), the space of possible productions is limited and more comparable across study participants. Participants were given a sentence and asked to write as many questions as they could think of that the sentence might answer. The pre-test did not cover any closed exercises so as not to provide practice opportunities before the practice phase (Hartley, 1973).

Explanations were provided as a distinct part of the study in order to establish similar conditions to those found in real-world settings of using the system accompanying a language class. All explanations were provided in German language. They contained the rules for formation of questions with the four `properties` given above. A list of the modal verbs used in the exercises was also provided. Participants could always navigate back to the explanations after they had been introduced.

The interim test was structured similarly to the pre-test and elicited the same question words and verb classes, but used different prompts.

The administered practice exercises consisted of Gap-filling exercises with a single gap for the question word, the subject (or object in subject questions), and the verb. The control group received only exercises on object who- and what-questions with the modal *can*, as well as questions with the modal substitute *have to*, do-support verbs, and with *be*. The intervention group received wh-questions with a variety of questions words, modals and modal substitutes including subject questions with *who*, and questions with do-support verbs and with *be*. Each group received a maximum of 30 exercises. The exercises were identical for both groups except for the concrete modal or modal substitute of the questions and the question word. For subject questions, the co-text and the question word were kept identical and the exercises differed between the conditions merely in whether they asked about the subject or the object. Thus, the exercises of the two study condition groups targeted the same topic, task type, skill modality, and pragmatic intention. By thus controlling for all dimensions mentioned by Suzuki (2023) to have an impact on transfer except for the linguistic context, we aimed to investigate the effect of varying the dimension of linguistic context on transfer. The linguistic forms were presented in interleaved order since most studies involving variability used this scheduling approach (Kim et al., 2021).

The working memory test was based on the implementation of an n-back task used

in the I4S study (Parrisius et al., 2022a). It constitutes an adapted version of the implementations by Pelegrina et al. (2015) and Schleepen and Jonkman (2009). First introduced by Kirchner (1958), the task presents sequences of stimuli. Test takers are required to react depending on whether the stimulus n positions back fulfills certain conditions. Typical implementations are 0-back, 1-back, 2-back and 3-back tasks, often administered in succession. Our implementation comprised a 1-back, 2-back and 3-back task in which letters were displayed on the screen one after the other for a couple of seconds each. Learners had to press a key if the currently displayed letter was identical to the letter displayed 1, 2, or 3 positions previously.
Similar to Madlener's (2015) experiments on the effect of variability, we aimed to assess correctness and variedness of the produced constructions after practice through closed exercise types, and generalization and productivity through more open production exercises. The post-test therefore included elicited questions that prevent learners from drawing upon controlled processing through their more open nature, thus allowing to assess proceduralization effects (Fakharzadeh et al., 2014). The elicited questions were similar to those presented in the pre-test, but with different linguistic contexts. With those questions, we also aimed to determine whether presenting constructs in different linguistic contexts encourages varied, pro-active usage of the constructs in varied contexts, thus addressing one of the pillars of language complexity (Ellis, 2003; Chen and Meurers, 2019). The more closed exercises allowed to in addition measure transfer to new linguistic forms regardless of skill transfer to different task types. These exercises were of a similar nature to that of the practice exercises, only introducing new linguistic contexts. They were completed after the more open, elicited post-test exercises in order to not provide additional exposure to varied constructs before the elicited questions assessing variedness of learner productions. The closed post-test exercises covered all practiced question words, modals, modal substitutes, some do-support verbs, and *be*. In addition, they covered the new forms of subject what-questions, the modals *could*, *shall*, *might*, and the modal substitute *be allowed to*, which neither of the two groups had seen before. New constructs were presented first, then constructs that only the control group had seen, followed by constructs that both groups had received in the practice phase.
The full study material is provided in Appendix A.2.

The thus collected learner answers served as data for a range of analyses. The closed Gap-filling exercises allowed us to enforce the use of specific `linguistic`

`constructs`. By comparing the accuracy on these exercises between the study conditions we thus aimed to investigate whether linguistic knowledge practiced in the practice exercises transferred to other `linguistic constructs` on the one hand and `properties` on the other hand. Since most learners finished neither all practice nor all test exercises, it was not possible to simply consider the set of exercises of the post-test that covered `linguistic constructs` not covered in the practice exercises. Instead, we determined for each study participant the constructs they had practiced based on the submitted practice exercises, as well as the constructs targeted by the post-test exercises. In terms of `linguistic constructs`, we focused on annotations specific to wh-questions: The specific wh-word, object vs. subject question, the verb category ('be', specific modal, specific modal substitute, or other do-support verb), and the modality (deontic, epistemic or circumstantial) of modals and modal substitutes. Two annotators with backgrounds in computational linguistics independently labeled all exercises with these annotations. Multi-label Inter-Annotator Agreement (IAA) was calculated as Krippendorff's Alpha at $\alpha = .7351$. Discrepancies were resolved manually. We analyzed learner performance depending on whether they had practiced constructs with the annotations of the currently evaluated exercise with four distinctions: (1) At least one practice exercise covered all annotations of the current exercise (*all practiced together*). (2) No single practice exercise covered all annotations of the current exercise, but all annotations of the current exercise were covered by at least one practice exercise (*all practiced*). (3) Some annotations of the current exercise were covered by at least one practice exercise (*some practiced*). (4) No annotation of the current exercise was covered by any practice exercise (*none practiced*).

Annotation practice	**Accuracy C**		**Accuracy I**	
	μ	σ	μ	σ
All practiced together	.8	.2	.7	.1414
All practiced	.5667	.2233	-	-
Some practiced	.5325	.2563	.5094	.2864
None practiced	.6374	.2732	.3844	.2488

Table 3.7: Accuracy on post-test Fill-in-the-Blanks exercises depending on annotation practice.

Performance was highest for both the control (C) and the intervention (I) group if all annotations had been practiced together before, and lowest if some annotations or all but not together had been practiced.

We first determined the effect of whether the learner had practiced the targeted `linguistic constructs` on the learner's accuracy in post-test exercises. None of the effects of the mixed-effects regression controlling for class level and English grade were significant with p-values ranging from .102 to 1.0. Table 3.7 outlines that for both study conditions, performance was highest if all annotations had been practiced together before. This is reflected in large effect sizes calculated as Hedge's g when comparing submissions where all annotations had been practiced together with submissions of any of the other practice states, as can be seen in Table 3.8. If some annotations or all but not together had been practiced before, performance was even lower than if none had been practiced. This pattern emerged regardless of whether practice was successful. Successful practice was defined as eventually having submitted a correct answer to an exercise during the practice phase.

Annotation practice	**Hedges's g**
All practiced together - all practiced	.8786
All practiced together - some practiced	.8984
All practiced together - none practiced	.9324
All practiced - some practiced	.1681
All practiced - none practiced	.2290
Some practiced - none practiced	.0672

Table 3.8: Effect sizes between submissions of different practice states.
Effect sizes for accuracy differences between submissions of different practice states were high between submissions where all annotations had been practiced together and the remaining practice states. Effect sizes between other practice states were low.

When comparing post-test performance between the study conditions, the control condition performed better than the intervention group, although the effect was only significant with a moderate effect size when no annotations had been practiced (t=2.045, p=.044; g=.6204) and marginally significant also with a moderate effect size when all constructs had been practiced together (t=1.886, p=.062; g=.4344). A possible explanation assumes that having been exposed to the same `linguistic constructs` at a higher frequency, the control group was able to form representations of annotation groups in memory that they were able to retrieve in the post-test. For the annotations that they had not seen before, both groups had to resort to declarative knowledge. However, the intervention condition was exposed to the annotations in a variety of constellations so that high cue competition and incorrect or insufficient abstraction may have led to confusion during the post-test. Cue

validity and cue strength thus probably could not be established in the intervention group as there were too few occurrences per `linguistic construct` to learn from during the practice phase (Bates and MacWhinney, 1987).

In order to determine at what linguistic level KCs should be defined, we next analyzed learners' performance on post-test exercises practicing certain instantiations of a linguistic characteristic depending on how many different instantiations of that characteristic they had seen in practice exercises. We looked at characteristics whose instantiations differ on a lexical, syntactic, or semantic level. To this purpose, we extended the exercise annotations with the lexical verb type such as *come*, *buy*, or *read*, and the interrogative type classifying wh-words into interrogative pronouns and interrogative determiners. The verb category differentiates between *do-support*, *be*, and *modal* verbs, which all require different syntax for question formation. Two characteristics group verb forms expressing the meaning of a necessity and of a possibility, respectively. Wh-word type, modality, and subject vs. object questions were defined in analogy to the annotations used previously. This resulted in the two lexical characteristics *verb type* and *wh-word type*, the three syntactic characteristics *interrogative type*, *subject vs. object question*, and *verb category*, and the three meaning-focused characteristics *modality*, *'necessity' meaning*, and *'possibility' meaning*. For each instantiation of these characteristics, we determined all learners who had not seen the same instantiation in any practice exercise. For each of these learners, we then determined the accuracy across all test submissions targeting this instantiation of the characteristic, as well as the number of distinct other instantiations of the same characteristic practiced in the practice phase. Table 3.9 gives the results of a mixed effects analysis of the effect on test accuracy of having practiced any other instantiation of the characteristic before, controlling for class and grade.

Note

The dataset was too small to allow to control for additional variables. Future research should, however, also consider the number of practiced exercises, the exact number of distinct practiced instantiations of the characteristic, and the specific instantiation.

The analysis indicated significant effects and moderate to high effect sizes for all characteristics except *'necessity' meaning*. When looking at the effect of varied

Characteristic	Effect for practice state			Effect for practice variability		
	t	p	g	t	p	g
Verb type	2.467	.037	.5398	2.467	.037	.5398
Wh-word type	2.845	.021	.6426	2.918	.018	.6479
Interrogative type	3.037	.004	.7267	-	-	-
Subject vs. object question	3.996	.0	1.2083	-	-	-
Modality	3.312	.002	1.4579	3.312	.002	1.4579
'necessity' meaning	1.516	.155	.4707	1.959	.052	.5258
'possibility' meaning	2.68	.009	.5639	-	-	-.3399

Table 3.9: Mixed effects analysis results for accuracy depending on practice states.

The dependent variable constitutes the accuracy in post-test exercises practicing a not practiced instantiation of the characteristic. The independent variable constitutes the practice state in terms of having practiced another instantiation of the same characteristic against having practiced no other instantiation, or the practice variability in terms of having practiced at least two other instantiation of the same characteristic against having practiced less distinct instantiations. Effects are significant with moderate to high effect sizes across all linguistic levels. Positive t-values indicate higher accuracy when having practiced other instantiations and for varied practice, respectively.

practice in terms of having practiced at least two other instantiations of the characteristic, the effect was similar for the same characteristics except for *'possibility' meaning*, for which the model did not converge. The syntactic characteristics *interrogative type* and *subject vs. object question* only have two instantiations so that varied practice in terms of practicing more than one instantiation that in addition differs from the test instantiation is not possible in those cases. The box plots in Figure 3.14 illustrate that accuracy on test exercises was higher when learners had practiced multiple instantiations of the characteristic, although there does not seem to be an additional benefit to having practiced more than two or three distinct instantiations.

We therefore conclude that although transfer might be possible on the lexical, semantic, as well as the syntactic level, the syntactic level in most cases has only a very limited number of instantiations per characteristic so that the number of distinct instantiations required for varied practice already covers all instantiations. It therefore makes most sense to define KCs as syntactic `properties`, and cover variations of lexical and semantic characteristics, between which transfer is possible, within each KC. It is not necessary to exhaustively practice all variants at these

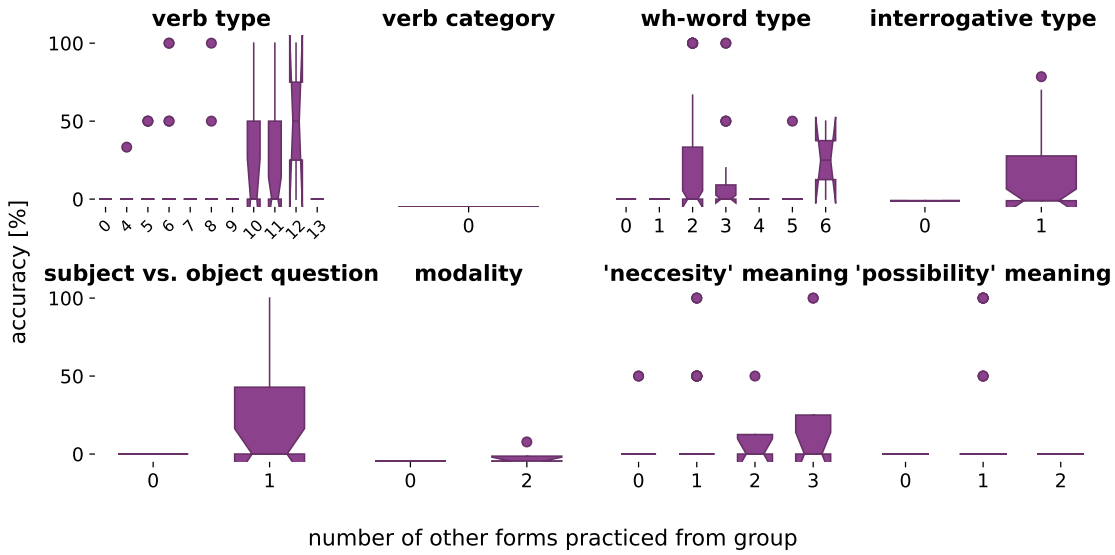


Figure 3.14: Accuracy on post-test exercises depending on practice variability. Accuracy on test exercises was higher when learners had practiced multiple instantiations of the characteristic. There does not seem to be an additional benefit to having practiced more than two or three distinct instantiations.

levels. KCs are thus defined as closed-type characteristics so that the space of KCs to master becomes feasible.

3.2.3 Proceduralization Effects of Linguistically Varied Practice

Language learning is different from content domains in that grammar practice aims at proceduralization of knowledge, whereas content domains aim at introducing and contextualizing concepts in their domain. In this respect, language learning is more similar to Maths where the application of rules also needs to be automatized. However, unlike Maths, language learning cannot be de-composed into independent KCs that can be acquired one after the other. Within language learning, vocabulary learning is different from grammar practice and more similar to content domains, so that grammar and vocabulary practice need to be treated differently.

Proceduralization of knowledge is reflected through increased accuracy and fluency in language production (DeKeyser, 2001). Yet, it does not consider the first component of the triad of Complexity, Accuracy, and Fluency (CAF) that forms the basis

of performance, proficiency and development in SLA (Housen et al., 2012). Variability does, however, constitute a parameter of linguistic complexity (Révész et al., 2017) and, as such, links to proficiency (Chen and Meurers, 2019). Learners are more likely to use linguistic variants in spontaneous speech that have been automatized (Ellis, 1985). Therefore, variants that have been practiced more frequently are more likely to be used in production. Variability in practice is thus relevant not only for proceduralization, but also in order to increase the linguistic complexity of a learner's language through higher linguistic variability (Ellis, 2003).

In order to investigate whether practice that intentionally introduces linguistic variability is beneficial for proceduralization, we analyzed learner answers to Gap-filling and elicited exercises in terms of their accuracy and fluency. In addition, we evaluated the variedness of a learner's productions depending on variedness of practice.

The analyses are based on the data described in Section 3.2.2. We manually annotated all learner submissions to the elicited exercises in the pre-, interim and post-tests. The annotations were two-fold. (1) The same two annotators who annotated the exercises now labeled each question produced by a learner with the linguistic annotations also available for the Gap-filling exercises: The specific wh-word, object vs. subject question, the verb category ('be', specific modal, specific modal substitute, or other do-support verb), and the modality (deontic, epistemic, or circumstantial) of modals and modal substitutes. (2) In addition, they annotated the type of error with a focus on errors concerning the formation of questions. Learner productions that contained only errors that are unrelated to question formation, such as spelling mistakes, incorrect formation of possessive forms, or incorrect formation of a verb form, were evaluated as correct. IAA for the correctness of a learner's answer was calculated as Cohen's Kappa at $\kappa = .7266$. For the evaluations, only learner answers that were labeled as correct by both annotators were considered correct answers. IAA for specific error labels was calculated as Krippendorff's Alpha at $\alpha = .4232$. The evaluations used the union set of both annotation sets.

Since no comparable pre-test exercises were available for Gap-filling exercises, we calculated improvement values based on the difference between a learner's performance on practice exercises and on post-test exercises.

The difference between the study conditions in fluency increase from interim to post-test, approximated by the number of produced questions within the given time, was

not significant neither for Gap-filling (t = .3063, p = .7599; g = .0570) nor for elicited exercises (t = 1.5378, p = .1269; g = .2862).

Focusing first on Gap-filling exercises, we assessed improvement in terms of the difference in accuracy on the post-test exercises and accuracy on the practice exercises. To this purpose, we determined accuracy based on the first attempt on each submitted exercise. The intervention group performed overall sightly, but not significantly (t=-1.088, p=.279), lower on the Gap-filling practice exercises with a low effect size of Hedge's g=.2197. Contrarily, the intervention group had non-significantly higher (t=.235, p=.815; g=.0620) improvement between practice and post-test for Gap-filling exercises. However, this interpretation neglects the fact that the control group was expected to perform better on the practice exercises if we assume less variability to result in lower complexity. We would consequently expect improvement to be lower for the control group, which, indeed, it was. In view of these considerations and the non-significance of the effects, the intervention group did not display higher learning gains on Gap-filling exercises than the control group.

An interesting picture emerges, however, when looking at the learners of different class levels and with different English grades separately, as displayed in Table 3.10. While for class 7, the control group did indeed have greater improvement than the intervention group (t=.779, p=.864; g=.2441), this effect is reversed for class 8 (t=-2.162, p=.147; g=.9307). Similarly, mean improvement was higher in the control group for learners with English grades 3 (g=.1901) or 2 (g=.3865), but higher in the intervention group for learners with English grade 1 (g=.8371)[5]. None of the effects

English grade	**Class level**	**Improvement C**		**Improvement I**	
		μ	σ	μ	σ
1		.2342	.2722	.4540	.2101
2		.2341	.2090	.1610	.1541
3		.1491	.1147	.1255	.0111
	7	.2422	.2239	.1928	.1678
	8	.1990	.2489	.4493	.2522

Table 3.10: Improvement on Fill-in-the-Blanks exercises.
The control condition (C) had higher improvement than the intervention condition (I) in class 7 and for learners with English grades 2 and 3, but lower improvement in class 8 and for learners with English grade 1.

[5]Grades in the German school system range from 1 (*very good*) to 6 (*failed*).

are significant, but effect sizes are high for class level 8 and grade 1.

A possible explanation assumes that proceduralization was low during the practice phase for 7th grade learners since the varied practice material did not provide enough repetitions of each `linguistic construct` to support automatization. Procedural knowledge might already have been in place for 8th grade learners before the study, so that varied practice strengthened and broadened this knowledge.

Turning the focus to elicited exercises, accuracy in the pre-test as well as in the interim test was higher for the control group receiving linguistically focused practice than for the intervention group receiving linguistically varied practice, although not significantly and with low effect sizes. In the post-test, however, the intervention group performed better, although only significantly in the sub-group of grade 7 participants (t=2.25, p=.027; g=.4871). This entails higher improvement for the intervention condition, which is significant with a moderate effect size (t=2.6986, p=.0080; g=.5022) between the pre-test and the post-test, but not between the interim and the post-test (t=.6489, p=.5177; g=.2587). Improvement between the pre-test and the interim test was also non-significantly higher for the intervention group (t=1.3900, p=.1673; g=.1208). As can be seen in the box plots presented in

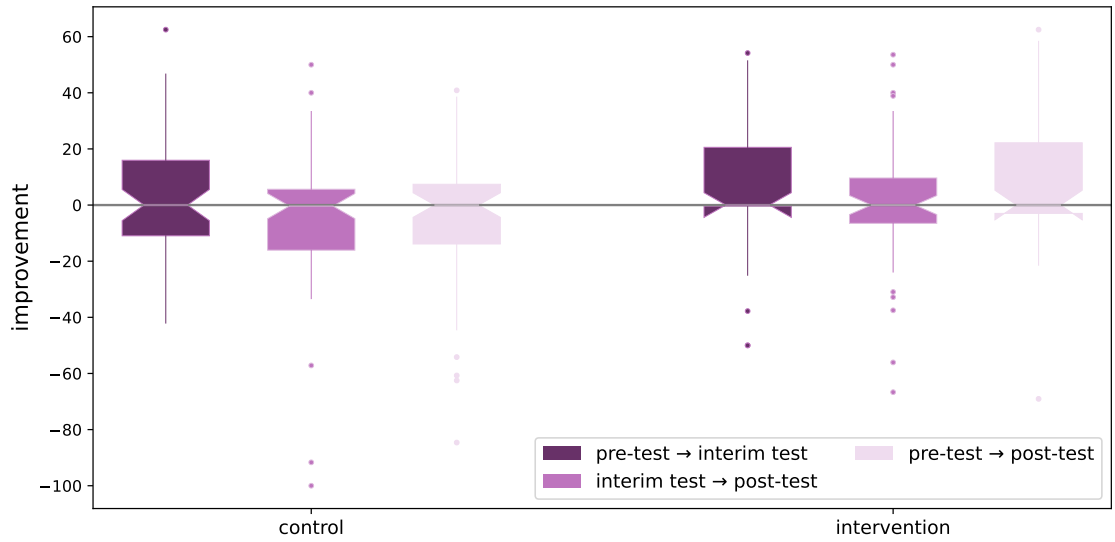
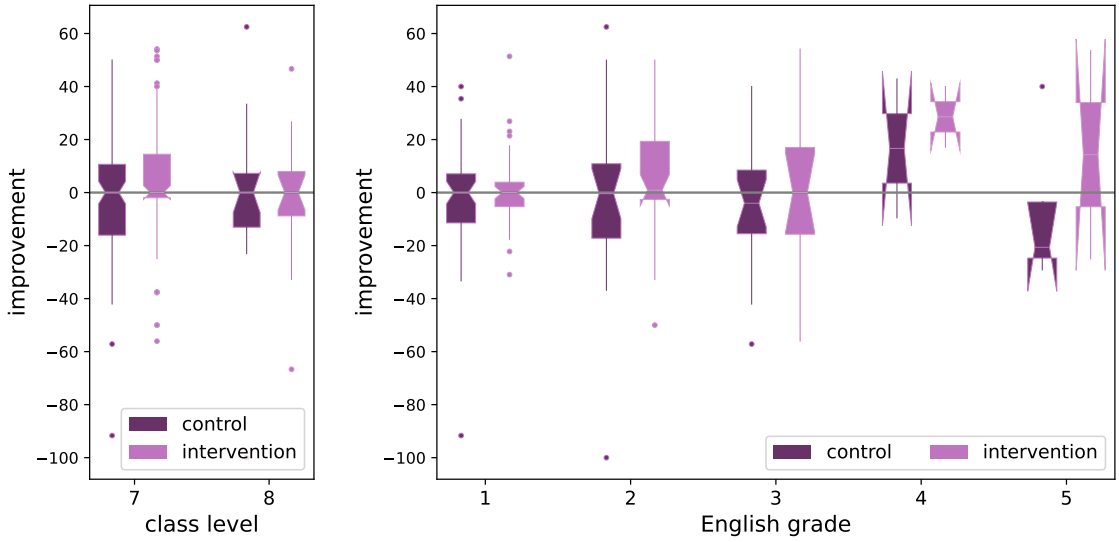


Figure 3.15: Improvement between pre-, interim and post-test.
Improvement in accuracy was lower and predominantly negative in the control group, yet mostly positive in the intervention group.

Figure 3.15, improvement was even negative for many learners in the control group, while it was mostly positive in the intervention group.

This effect is particularly noticeable with lower English grades as well as lower class levels, as can be seen in Figure 3.16. While the difference between control and intervention group is marginal in class 8 (t=0.799, p=.427; g=.0044), the box plots in Figure 3.16a clearly show higher improvement for class 7 participants of the intervention group than of the control group. The effect at this class level is marginally significant at p=.065 (t=1.87; g=.3780). In terms of English grade, the beneficial effect of varied practice relative to the control group is most noticeable in learners of very low grades (4 and 5), as displayed in Figure 3.16b.



(a) By class level. (b) By English grade.

Figure 3.16: Improvement between interim and post-test per subgroup. Higher improvement in accuracy for the intervention group is particularly noticeable in the lower class level of 7th grade and with learners of low English grades.

The bar plot in Figure 3.17 provides the frequency of different errors observed in the productions of the post-test relative to their frequency observed in the interim test. Although the error annotations differ across study conditions, there is no discernible pattern as to what kind of errors varied practice might prevent and what kind of errors it might encourage.

For 7th grade learners, the higher improvement on elicited exercises when practice was varied might be explained by the assumption that the difference in exercise types

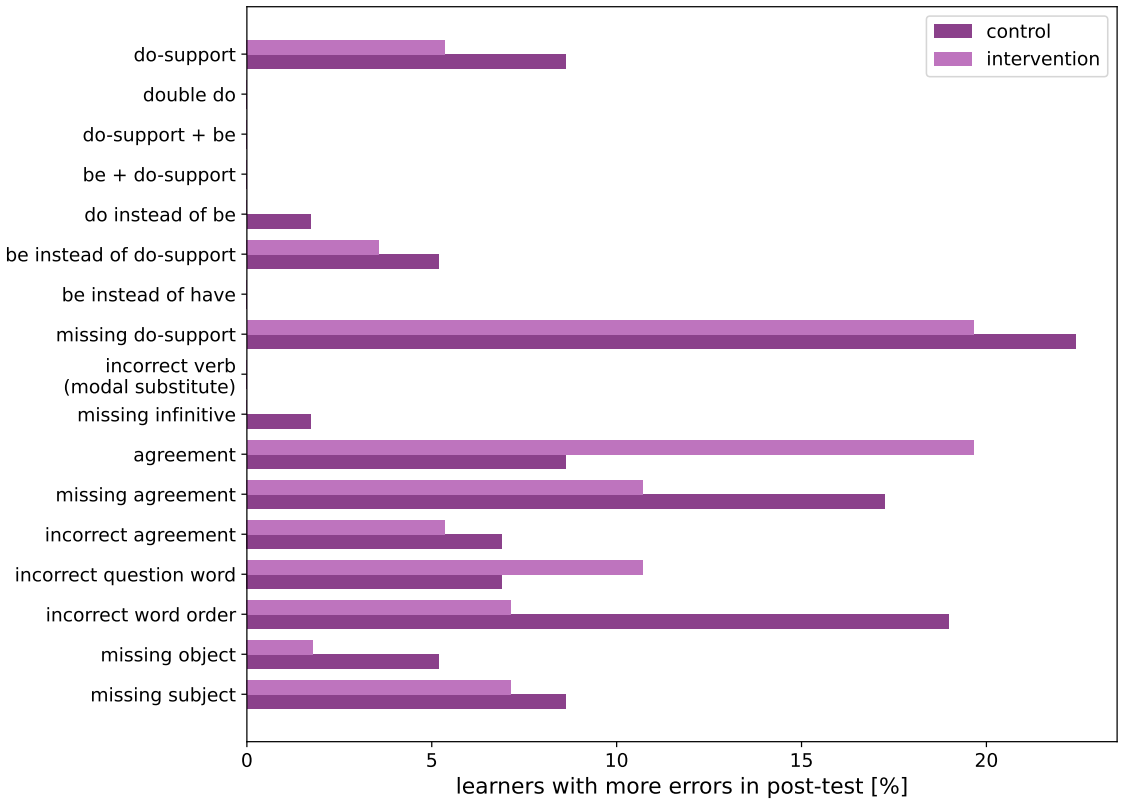


Figure 3.17: Development of the frequency of error annotations.
Increase and decrease in errors per learner and type differed between the study conditions, yet at no discernible pattern.

between the practice and test exercises required them to rely mainly on declarative knowledge instead of on procedural knowledge. Since varied practice exposed the learners to more different `linguistic constructs`, their declarative knowledge was strengthened more than without varied practice. For 8^{th} grade learners, data might have been too scarce to yield reliable results.

The observations regarding improvement on elicited exercises are in stark contrast to the results on Gap-filling exercises. While improvement in accuracy was higher for the control group on the closed Gap-filling exercises, it was higher for the intervention group on the elicited exercises. In addition, variability was shown to be more beneficial for higher class levels and better English grades with respect to the Gap-filling exercises, whereas the current analyses identified a beneficial effect of varied practice for lower class levels and English grades. We suspected the cause of this observation to be rooted in the nature of the `linguistic constructs` that the

learners produced in the two exercise types. An evaluation of the practice states of the produced `linguistic constructs` in terms of having practiced none, some, or all annotations targeted by the exercise indeed showed some differences between Gap-filling and elicited exercises (see Table 3.11).

Annotation practice state	**Percentage among practice states (Gap-filling)[%]**				**Percentage among practice states (Elicitation)[%]**			
	μ_C	σ_C	μ_I	σ_I	μ_C	σ_C	μ_I	σ_I
Together	3.14	5.20	2.83	5.11	11.36	20.79	11.10	12.36
All	4.35	8.01	1.23	6.72	.0	.0	.0	.0
Some	39.37	42.14	37.64	43.30	71.19	36.98	58.33	36.45
None	11.76	27.81	17.23	34.46	5.38	18.81	14.50	28.05

Table 3.11: Percentages of construct practice states.
In elicited questions, participants of the control group used more `linguistic constructs` some but not all of whose annotations they had previously practiced than participants intervention condition. This was not the case in closed Gap-filling exercises.

In the closed Gap-filling exercises study participants in both conditions were forced to use roughly equal numbers of `linguistic constructs` some but not all of whose annotations they had previously practiced. In the more open, elicited exercises, learners produced such `linguistic constructs` more frequently than in Gap-filling exercises relative to other practice states. However, participants of the control group used more of these `linguistic constructs` some but not all of whose annotations they had previously practiced than participants in the intervention condition. Since, for forms in this practice state, learners do not have mappings for all available cues in the form of linguistic annotations, such forms are most likely to result in incorrect language productions. The used `linguistic constructs` can thus, at least to some degree, explain the differences in performance between Gap-filling and elicited exercises across study conditions.

An alternative explanation assumes that the Gap-filling post-test exercises elicited near transfer of the knowledge targeted by the practice exercises, whereas the elicited questions elicited far transfer (DeKeyser, 2018). It further assumes that for 8^{th} grade participants, procedural knowledge was already in place before the study, but not for 7^{th} grade participants. Consequently, if procedural knowledge was already in place, variability was beneficial for near transfer (via procedural knowledge) but negatively impacted far transfer (via declarative knowledge). Inversely, if procedural

knowledge was not yet established, variability negatively impacted near transfer and was beneficial for far transfer. According to this hypothesis, varied practice strengthens knowledge of the type that is used during practice but slows down proceduralization. There is thus a trade-off between acquiring linguistic knowledge in breadth and in depth.

Disentangling the syntactic from the lexical dimension, we further examined learner answers to post-test Fill-in-the-Blanks exercises whose constructs the learner had seen in a similar constellation before, but differing in one annotation. The box plots in Figure 3.18 hint that the control group performed better if the practiced constructs differed only in the wh-word (lexical dimension) or in the verb (both lexical and syntactic dimensions), while the intervention group performed better if the constructs differed in subject vs. object questions (syntactic dimension) or the modality (semantic-pragmatic dimension). None of the effects are, however, significant according to a mixed-effects analysis, and all effect sizes are low.

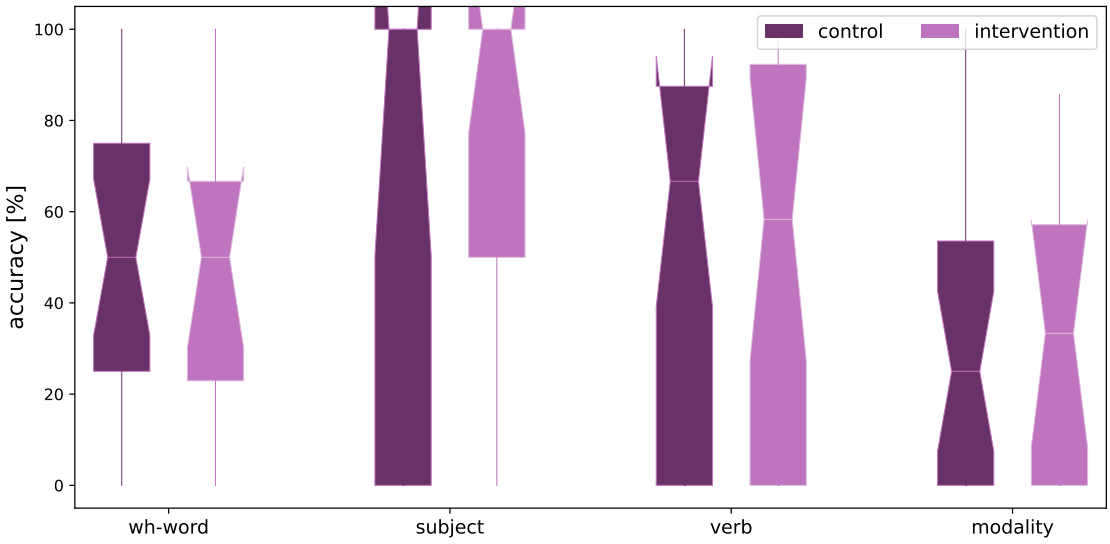


Figure 3.18: Post-test performance on Gap-filling exercises per required transfer. The intervention group performed better than the control group if the constructs differed only in subject vs. object questions or the modality, and worse if they differed only in the wh-word or in the verb.

In order to determine whether variability has an effect on the variedness of learner productions, we further analyzed the `linguistic constructs` produced in elicited questions of the post-test. There were no significant differences between the study

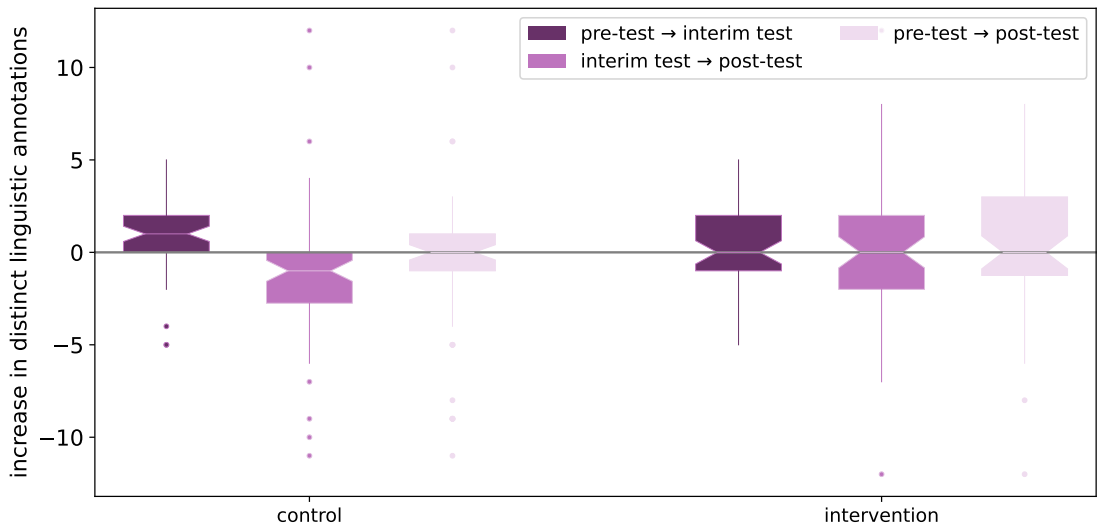


Figure 3.19: Development of variability in produced constructs.
Variability increased between pre- and interim test and decreased again in the post-test for the control group. It remained constant for the intervention group.

conditions. The box plots in Figure 3.19 visualize the development of variability in learner productions for both study conditions on the level of linguistic annotations.

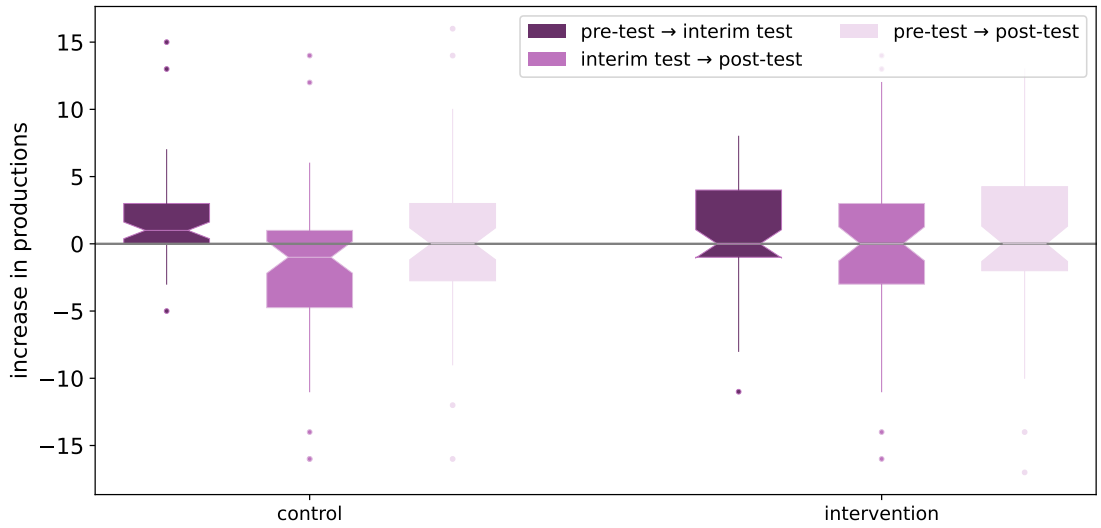


Figure 3.20: Development of the number of learner productions.
The number of produced questions increased between pre- and interim test and decreased again in the post-test for the control group. It remained constant for the intervention group.

In the control group, the number of different constructs decreased between the interim test and the post-test although there had been an increase between the pre- and the interim test. In the intervention group, the number stayed more or less constant across all tests. These differences between the study conditions concerning the number of different annotations were not significant either.

Considering that we observed a difference between the study conditions while the study material was identical, one must consider the decreased variability in the control group's productions with caution. It is rather likely that the differences in variability between study conditions and tests are related to the differences in number of produced questions. As can be seen in Figure 3.20, the picture across test timepoints for the number of productions is rather similar to that for the number of different `linguistic constructs`. Indeed, Pearson's correlation between these two measured values is strong at $\rho = .7064$. The box plots in Figure 3.21 illustrate that, while the intervention condition produced slightly more different constructs, the difference between the two conditions is much less pronounced when normalizing the number of different constructs by the learner's overall number of productions.

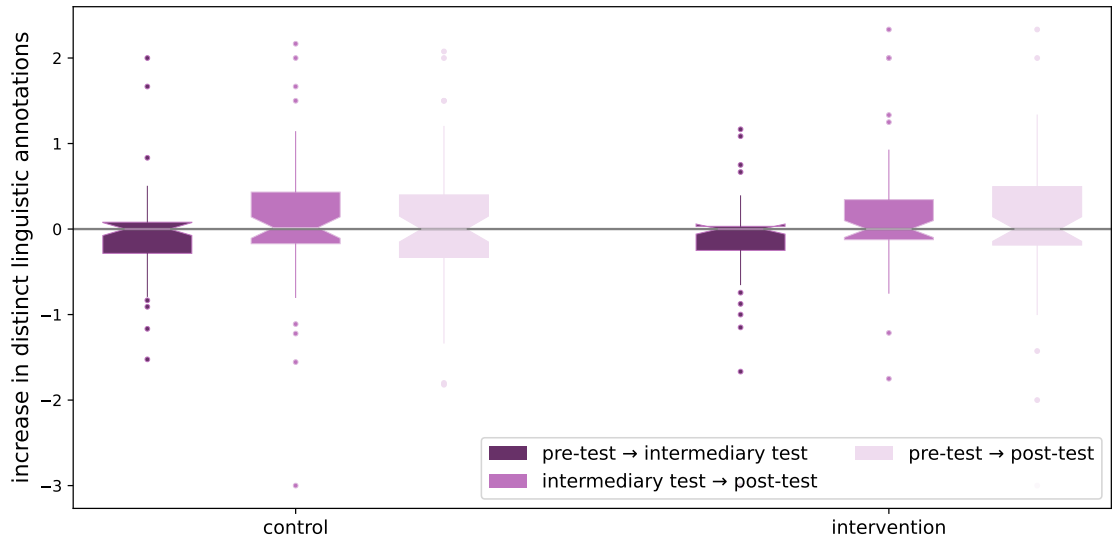


Figure 3.21: Development of the normalized number of learner productions. The normalized number of produced questions increased from test to test for both study conditions. The increase between interim and post-test was slightly higher for the intervention group.

We have, however, already seen in Table 3.11 that the intervention group more

frequently used `linguistic constructs` none of whose annotations they had practiced before, even though the likelihood of having practiced an annotation encountered in the post-test was higher for this study condition than for the control group as practice was more varied for those learners. Varied practice thus seems to encourage the use of other linguistic constructions, possibly because practice does not reduce a learner's focus to the subset of constructions seen frequently in the practice exercises.

In addition to overall numbers of distinct `linguistic constructs` and annotations, we analyzed variedness on two levels: Different verb forms, modals and modal substitutes in particular, allow for output of equivalent meaning but requiring different grammar rules. Different question words, on the other hand, result in productions with different meaning, some of which rely on the same declarative knowledge while others require different rules. We therefore also determined the numbers of different produced forms related to these properties. While we considered different modals and modal substitutes as individual forms, we assigned only a single label to other do-support verbs. As can be seen in Figure 3.22, the control condition produced less distinct forms in the post-test than in the interim test in both categories. Most learners in the intervention group, on the other hand, produced more distinct forms

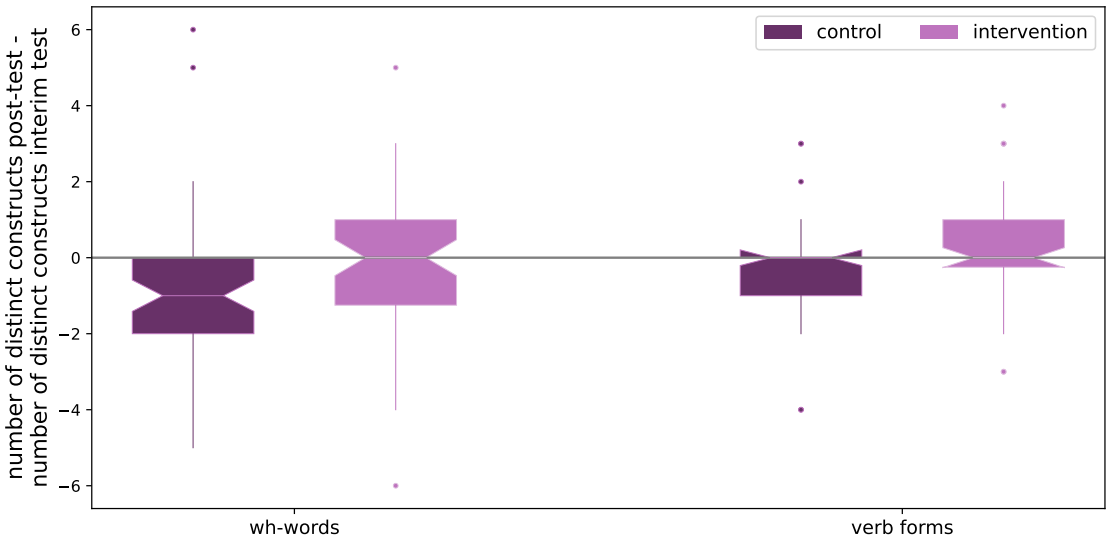
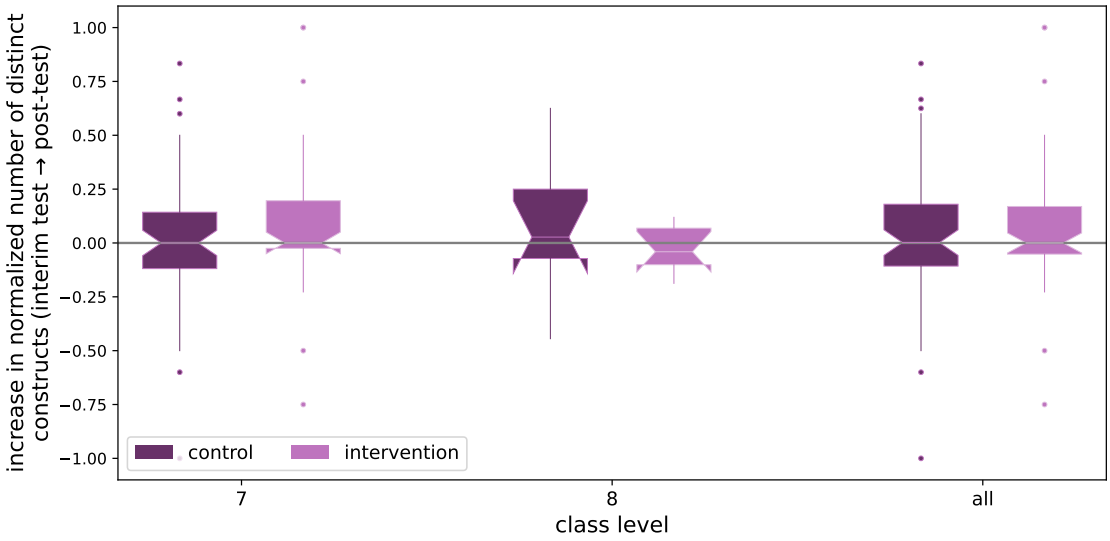
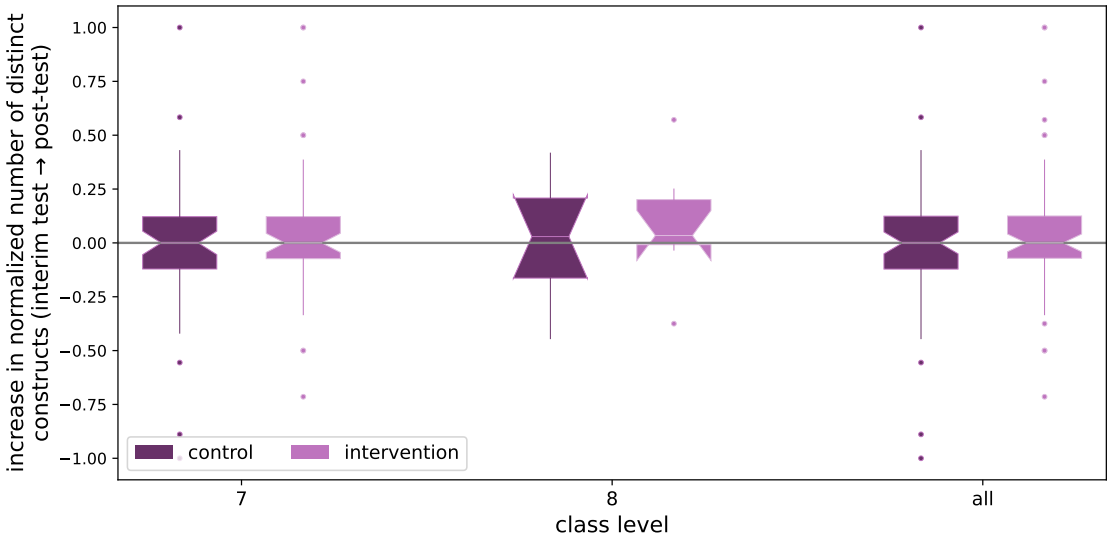


Figure 3.22: Variedness in form variations and meaning variations.
The control group produced less form and meaning variations in the post-test than in the interim test. The intervention group produced more variations of both types.

after the practice phase. The effect of the study condition is significant with a medium effect size for verb forms (t=1.9962, p=.0483; g=.3714) and marginally significant for wh-words (t=1.8639, p=.0650; g=.3468).



(a) Verb forms.



(b) Wh-words.

Figure 3.23: Normalized form and meaning variedness.
7^{th} grade students of the intervention group produced more form and meaning variations after the treatment than the control group. There is no treatment effect for 8^{th} grade students.

When normalizing the values by the number of produced sentences per learner and test, the effect is, however, not significant for the full sample of participants (t=.9736, p=.3324; g=.1812). Yet Figure 3.23 illustrates that there is indeed an effect of the intervention for a sub-sample of the participants. For the sample of 7^{th} grade students we again see a significant positive effect of the intervention for verb variability (t=2.312, p=.023) with a medium effect size (g=.4410) and a marginally significant positive effect for wh-word variability (t=1.862, p=.065; g=.3574). For 8^{th} grade students, the effect is not significant for verbs (t=.094, p=.925; g=.0109) nor wh-words (t=.944, p=.347; g=.2510), and the control group seems to produce slightly more varied verbs. Varied practice thus had a positive effect on variedness both at the level of forms differing in meaning and at the level of synonymous forms for less proficient learners.

In conclusion, varied practice covers linguistic forms in more breadth but less depth. Our evaluations allow us to develop informed hypotheses that may be further pursued in future research. These hypotheses suggest that variability on the one hand increases variedness in learner productions, but on the other hand might slow down proceduralization of knowledge. If this hypothesis is confirmed, macro-adaptive exercise selection needs to provide varied practice material in a balanced manner. It could reduce within-exercise variability in practice exercises as much as possible until the learner has reached mastery, and then increase within-exercise variability while all the time maximizing across-exercise variability. This way, learners' communicative abilities are increased in the initial stages of learning as they may retrieve the practiced variants strengthened through multiple repetitions, while their communicative abilities are extended in later stages through generalized representations that allow them to use the practiced forms more freely and independently (Schulz, 2024).

3.3 Linking Exercises to Variability-Based Knowledge Components

In order to consider linguistic and `skill` variability when selecting an exercise, these two exercise characteristics need to form the basis of KCs. Each KC thus represents a collection of linguistic forms bundled as a `property` and is practiced in combination

with a certain `skill`.

However, this alone does not enable a macro-adaptive algorithm to select suitable exercises. Exercises need to be linked to KCs to enable the algorithm to identify exercises that practice each KC.

3.3.1 Considering the Linguistic Dimension of Knowledge Components

The space of relevant KCs is usually represented in a domain model. Both `skills` and linguistic forms can be modeled in such a structure (Parkinson and Dinsmore, 2019). Considering both curricular constraints and linguistic variability for exercise selection requires a structure for the learning material that links higher-level organizational units to low-level linguistic constructions. If KCs are defined in-between and linked to these two extremes, the macro-adaptivity algorithm can consider the requirements of both when selecting an exercise.

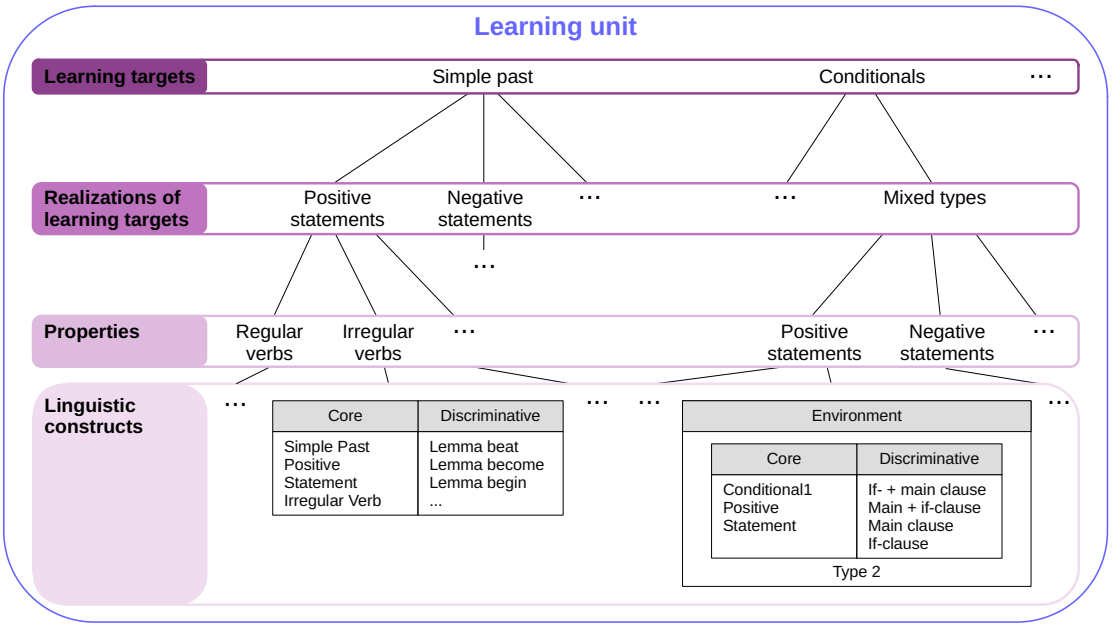


Figure 3.24: Hierarchy of the domain model.
The domain model has a tree structure so that elements are linked to a single element of the next higher hierarchy level.

As Ruiz et al. (2023) point out, a domain model for language learning should be created manually on the basis of a reference framework. For English grammar, the English Grammar Profile (EGP)[6] constitutes a suitable resource. On this basis, we propose a domain model that integrates the pedagogical perspective resulting from curricular constraints and linguistic considerations resulting from linguistic variability demands. It thus encompasses four levels that are organized in a tree structure as shown in Figure 3.24: The highest level comprises `learning targets` such as *Simple past*, or *Conditionals*. The second hierarchy level, `realizations of learning targets`, comprises more concrete instantiations of a `learning target` that indicate palpable learning goals for students, such as *Positive statements*, *Negative statements* and *Mixed statements*. All `realizations` are specific to a `learning target`. While these two levels are directly visible for learners in a student dashboard that supports the two entry points to adaptivity proposed in Section 2.1.3 (practice of the entire `learning target` or of a specific `realization`), learners are exposed to the third level only indirectly through progress visualizations, as illustrated in Figure 3.25.

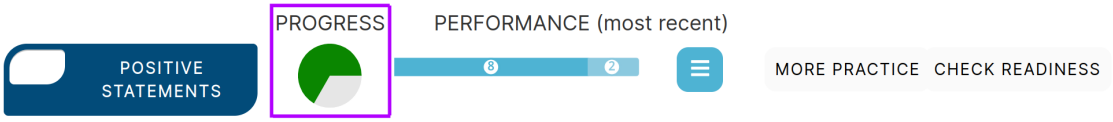


Figure 3.25: Progress visualization in the student dashboard.
Progress is represented in a pie chart that attributes a slice to each `property` of the `realization`.

This level defines `properties` of the `realizations`, which, in combination with `skills`, make up KCs. They thus mark milestones in the acquisition process that a learner has to master. `Properties` are also specific to a `realization`, such as *Irregular verbs* and *Regular verbs* for the `realization` *Positive statements* of the `learning target` *Simple past*.

The top three levels of the domain model constitute pedagogical layers informed by common curricular goals and developmental sequences of SLA. Developmental sequences represent the natural order in which learners acquire syntactic and morphological features (Ellis, 2015). The fourth layer, comprising `linguistic constructs`, is a linguistic layer to which learners are not exposed at all. These `constructs` are

[6]The EGP is available at https://www.englishprofile.org/english-grammar-profile/egp-online.

defined on a very fine-grained level based on combinations of linguistic features such as *simple past tense*, *regular verb*, and *question*. A naive approach simply introduces a new `linguistic construct` with a new name for each possible feature combination. However, we should then consider that, for instance, each irregular verb constitutes a linguistic feature that results in a `linguistic construct` in combination with other features such as polarity and tense of a sentence. This quickly escalates into a very large and very unmanageable domain model, especially if further insights later reveal that an additional feature should also be considered for some `constructs`. These considerations call for a more flexible approach. We therefore define `linguistic constructs` that each consist of three types of linguistic annotations: (1) core annotations, (2) environment annotations, and (3) discriminative annotations. All annotations can be obtained through Natural Language Processing (NLP) as described by Quixal et al. (2021). Core annotations comprise annotations that must be present in an actionable element of an exercise in order for the element to practice the construct, for example *statement*, *simple past* and *irregular verb* for exercises practicing the formation of irregular simple past verbs. Environment annotations must be present in any of the actionable elements of the exercise. This allows to, for example, enforce that positive and negative statements are both present in an exercise, which is a requirement for the *Mixed statements* `realization`. Defining *simple past* as a core annotation and *present perfect* as an environment annotation, for instance, results in the selection of exercises that contain actionable elements practicing simple past as well es elements practicing present perfect. Discriminative annotations are those annotations that make a `construct` unique. An example constitutes the specific verb when practicing irregular verbs in the simple past. They can be used by the adaptivity algorithm to select exercises of varied `constructs` by prioritizing those exercises that practice `constructs` where no other `construct` with the same core, environment, and discriminative annotations has yet been practiced. This approach keeps the domain model slim, scalable, and manageable.

Since the domain model hierarchy constitutes a tree structure, each element is connected only to a single element in the hierarchy level above its own level. There are no connections between elements of the same hierarchy level. Only linguistic annotations may be linked to multiple `linguistic constructs`.

The domain model organizes siblings at each hierarchy level according to devel-

opmental sequences of SLA whenever available. This is, however, predominantly the case for questions (Ruiz et al., 2023). Other language means rely on criterial features, which make it possible to associate language means with different proficiency levels according to the Common European Framework of Reference (CEFR) (Hawkins and Filipović, 2012). If none of these two options are applicable, the order is best determined by consulting practicing teachers.

In addition to these four levels of the domain model, learning units group multiple `learning targets` outside of the domain model. Learning units can in principle be individually assembled for any language class. It is possible to include only a subset of `realizations of learning targets` and `properties` specified in the domain model into a learning unit.

In order to link exercises to the domain model, they need to be mapped to domain model components at one of the four levels. This link is most specific when established on the lowest level of the domain model, as links to higher levels result indirectly from links within the domain model hierarchy.

Following the procedure suggested by Quixal et al. (2021), our approach annotates actionable elements with linguistic constructions such as *conditional type 1*, *if-clause*, or *irregular verb 'go'* based on automatic analyses of actionable elements and non-actionable co-text. The annotations are determined based on labels corresponding to EGP constructs, which are processed further to map them to `linguistic constructs` of the domain model. The EGP labels are obtained from a micro-service that was developed in the course of the POLKE[7] project. The implementation of the micro-service relies on NLP processing tools including a sentence segmenter, tokenizer, POS tagger, and constituency and dependency parsers. Combinations of the labels obtained from this micro-service form requirements for the presence of a `linguistic construct` of the domain model in an actionable element. Depending on the specific construction, it needs to be contained entirely or only partially by the actionable element. Apart from annotations contained partly in an actionable element and partly in the co-text, annotations contained in the co-text are not further considered. The specific combinations of core, environment, and discrimina-

[7]Pedagogically oriented language knowledge extraction and readability-controllable natural language generation (POLKE) aims to provide means to model a learner's foreign language knowledge and to automatically generate language material adapted to the learner based on that model (LEAD Graduate School & Research Network and Eberhard Karls Universität Tübingen).

tive annotations that make up a `linguistic construct` are defined in the domain model. All constructions that are annotated for the actionable element itself can represent both core and discriminative annotations, all constructions annotated in other actionable elements of the exercise are considered environment annotations. If all annotations of a `linguistic construct` specified in the domain model are present, the actionable element practices that `construct`. `Linguistic constructs` are, however, not stored with the exercise in order to enable dynamic assembly of exercise items into exercises. Environment annotations thus may or may not be present depending on the other items included in the exercise. Practiced `linguistic constructs` are therefore determined for each exercise dynamically based on the linguistic annotations when selecting the next exercise.

Thanks to these annotations, exercises can be selected that practice a specific `property`. An evaluation of the annotated exercises revealed that there was considerable room for improvement with respect to the exercise pool of the AI2Teach study. Suitable exercises were found for only 29 of the 91 `properties` (31.87%), with a mean of 7.5824 (sd=15.1343) exercises per `property`. First improvements to the linguistic annotation process, however, yielded very promising results. Exercises were now found for 80.36% of the `properties`. An overview of the number of exercises linked to each of the `properties` of the domain model with the improved annotations is provided in Appendix A.3.1.

It is, in principle, possible to maintain multiple `linguistic constructs` in the domain model that encompass the same linguistic annotations. This makes it possible – even though the domain model does not support many-to-many connections between elements of different hierarchical levels – to update mastery for multiple `properties` at the same time, even across `realizations` or `learning targets`.

Inversely, an exercise may receive annotations for multiple linguistic constructions. Since different combinations of these constructions may be associated with different `properties`, the linguistic annotations do not clearly place an exercise within the domain model at a single node of the tree. Instead, they might link an exercise to multiple `properties` and even multiple `realizations` or `learning targets`. This is particularly likely for exercises designed for practice contrasting two linguistic features in `learning targets` like *Simple past vs. present perfect*. Such an exercise is very likely to receive annotations of `linguistic constructs` linking to

higher-level domain model elements of the two distinct `learning targets` *Simple past* and *Present perfect*. Yet, `realizations` constitute a pedagogical level in the domain model. This implies that the exercises assigned to a `realization` should be designed in line with a specific pedagogical goal. We therefore aimed to determine whether, for linking exercises to the domain model, the linguistic level alone is sufficient to select pedagogically suitable exercises, or whether an additional mapping at a higher, pedagogical level of the domain model is required. Assuming that each exercise is pedagogically suitable for a single `realization`, an automatic approach needs to identify this `realization` based on `properties` of the exercise. If the automatic approach to link exercises to `realizations` identifies only a single suitable `realization` candidate for an exercise, making the link is straightforward. If, however, the approach identifies multiple potential `realizations`, these candidates need to be prioritized in order to determine the best fitting `realization` for the exercise. With the goal of identifying those exercises for which prioritization of potential `realizations` is relevant, we first determined the number of automatically assigned `realizations` based on annotated `properties` for each exercise. The evaluations are based on the exercises generated for the AI2Teach study, which were manually linked to `realizations`. They thus provide gold standard labels for `realizations` that can be compared with the automatically determined `realizations` in order to identify the best approach.

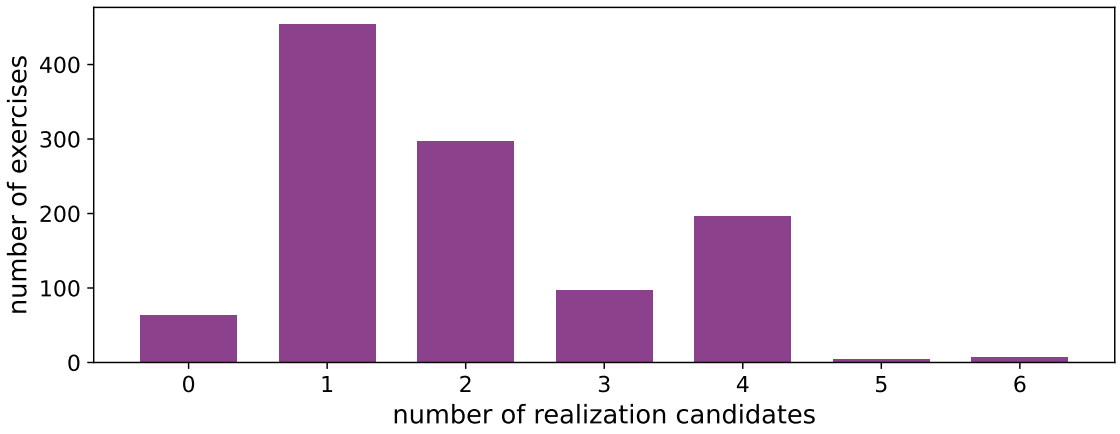


Figure 3.26: Distribution of the number of `realization` candidates for exercises. Many exercises constitute good practice material for only one `realization` according to the annotated `linguistic constructs`. A considerable number of exercises might be suitable for various `realizations` according to their annotations.

The results are shown in Figure 3.26. For 64 exercises, the linguistic annotation process was apparently not very successful so that no `properties` and consequently no `realizations` could be determined that the exercise intends to practice. For 454 exercises, only one `realization` candidate was identified. This corresponds to the gold-label `realization` in 89.21% of the exercises (405 correct, 49 incorrect). For the remaining 603 exercises, multiple `realizations` were found as possible candidates.

For the purpose of assessing whether the link between exercises and `realizations` can be determined automatically, we evaluated a range of approaches that aim to identify the best-fitting `realization` for an exercise in terms of its suitability for the pedagogical goal.

Our first approach (1a) assumes that the more `properties` of a `realization` an exercise practices, the more likely it is to be suitable for that `realization`. This approach therefore selects the `realization` with the largest number of associated `properties` annotated in the exercise. The results are summarized, along with the results of our other approaches to link exercises to `realizations`, in Table 3.12

Approach	n correct	n incorrect	Accuracy
(1a) `Property`-based (binary)	157	446	.2504
(1b) `Property`-based (element count)	312	291	.5174
(1c) `Property`-based (construct count)	312	291	.5174
(1d) `Property`-based (combined)	337	266	.5589
(2a) Prerequisite-based (exercise count)	328	275	.5439
(2b) Prerequisite-based (manual)	489	114	.8109

Table 3.12: Accuracy of automatic `realization` determination.
Approaches to automatically determine the most suitable `realization` for an exercise may use different weighting mechanisms for the linguistic annotations. The approach relying on manually defined prerequisites achieved highest accuracy.

Since this approach achieved very poor accuracy ($acc = .2504$), we experimented with some variations. The first variation (1b) in addition takes into account the number of actionable elements that practice each `property`. While it yielded considerably better results ($acc = .5174$), a substantial number of `realizations` still differed from the manually determined `realizations`. Considering the number of

distinct `linguistic constructs` instead of `properties` (1c) yielded identical results ($acc = .5174$). Further analyses revealed that this does not result from `properties` always being represented by a single `linguistic construct` in an exercise, which is not the case. The third variation of the `property`-based approach (1d) is based on the assumption that exercise creators focus on a specific `property` within an exercise. This variation therefore assumes that the most suitable `realization` is represented by only a single `property` in an exercise, and that the `property` in turn is practiced by most of the exercise's actionable elements. This variation therefore tries to minimize the number of different practiced `properties` while at the same time maximizing the number of actionable elements practicing the `property`. Results were even better for this variation ($acc = .5589$).

The second approach (2a) assumes that `realizations` that are prerequisites for another `realization` have more practice opportunities. For example, exercises with the pedagogical goal to practice conditionals type 2 will inevitably require learners to provide forms in the simple past. Positive and negative forms in the simple past, however, constitute themselves `realizations` of a different `learning target`. The approach therefore selects the `realization` that least often is a `realization` candidate in the entire exercise pool. The results were almost as good as those of the best variation of the first approach ($acc = .5439$). Since prerequisites determined by means of a frequency-based approach may differ from pedagogically determined prerequisites, we compared our automatically determined order of `properties` to a manually defined `property` hierarchy (2b). In the domain model of the Ai2Teach system, `realizations` and `properties` are by default ordered in the pedagogically most reasonable order within a `learning target`, so that prerequisite `properties` are always practiced before higher-level `properties`. We therefore determined the order of `learning targets` manually, and assumed the domain model structure to provide a reasonable `property` hierarchy in which prerequisite `properties` are listed as left-hand siblings of dependent `properties`. This variation of the approach yielded the best results ($acc = .8109$). It does, however, require at least a minimum of manual effort to reliably define prerequisites in the domain model. Although some approaches to macro-adaptivity rely on prerequisites for other purposes as well (e.g., Schwarz et al., 2012), this is not the case in our approach. The invested manual labor would therefore only serve the purpose of linking exercises to the domain model

more or less reliably on a pedagogical level.

Note

The prerequisite-based approaches to link exercises to `realizations` outlined that establishing a prerequisite hierarchy requires little effort in our approach. If desired, macro-adaptive exercise selection could therefore take prerequisites into account as additional curricular constraints.

In light of these results, we consider manually linking exercises to `realizations` to be the best approach, especially since it requires little additional effort. Exercise generation differs from `learning target` to `learning target`, sometimes from `realization` to `realization`. The `realization` for which the exercise is designed therefore needs to be specified in any case in order to create suitable exercises. In this process annotating the exercise with the `realization` for which it is pedagogically intended constitutes a much simpler, more efficient, and much more reliable option than handing the responsibility to the macro-adaptivity algorithm to identify the pedagogically best suited `realization` at runtime. Exercises should thus be linked to the domain model on two levels: linguistic annotations and `realizations of learning targets`.

3.3.2 Considering the `Skill` Dimension of Knowledge Components

Kwasnicka et al. (2008) highlight the importance of practicing different `skills` such as logical thinking, spacial imagination, and perceptiveness. While their instantiations of `skills` are tailored towards Maths, grammar exercises can also target different `skills`. Research has not yet established the effectiveness of certain exercise types for specific `skills` (Torre et al., 2011). However, most exercise types are naturally suited for practice of a specific `skill` or multiple `skills`, although there may be differences in the coding of the exercises in terms of exercise-to-`skill` matching across `learning targets`. Due to these potential variances, it makes sense to explicitly annotate the practiced `skill` in each exercise.

SLA researchers generally distinguish between the four `skills` *reading*, *writing*, *listening*, and *speaking* (Burns and Siegel, 2018). In the context of written grammar

exercises without oral input or output, this classification is less suitable. Munby (1981, p. 117) provides a much more fine-grained classification into overall 260 `skills`. However, differentiating between such a large number of `skills` when defining KCs is unpractical for multiple reasons. IAA for annotations of `skills` in exercises can be expected to be rather poor with so many classes. In addition, requiring learners to master all 260 `skills` is unreasonable, even if about half of them do not apply in our system due to their application exclusively in oral contexts. Since we aimed for a one-to-one or at least one-to-few mapping of exercises to `skills`, we therefore instead applied a bottom-up approach to define `skills` that should be practiced and mastered in our system. We thus took the exercises from the AI2Teach study as a starting point and defined `skills` that differentiate one exercise type from another and at the same time group exercises of the same type.

We illustrate `skill` mappings for the six exercise types presented in Section 2.2.2. Table 3.13 provides an overview of this mapping. **Mark-the-Words** exercises always target *metalinguistic knowledge*. In some cases, *semantic parsing* is required in addition. **Categorization** exercises are always used to practice *metalinguistic knowledge*. **Mapping** exercises can target different `skills` depending on the learning goal they encode. In our system, they are currently used for *semantic parsing* practice. **Jumbled Sentence** exercises always map to *word order* `skills`. If they include additional distractor chunks that are not part of the target answer, they also require *semantic parsing*. This is, however, not the case in our implementation. **Gap-filling** exercises target multiple `skills`. *Construct formation* is almost always

Exercise type	`Skills`
Mark-the-Words	metalinguistic knowledge, semantic parsing
Categorization	metalinguistic knowledge
Mapping	semantic parsing
Jumbled Sentence	word order
Gap-filling	construct formation, word order, semantic parsing
Short Answer	construct formation, word order, semantic parsing, transposition, sentence formation

Table 3.13: Mapping of exercise types to `skills`.
Each exercise type can be used to practice one or multiple `skills`. Although the actually practiced `skills` depend on the concrete exercise, the exercise type provides guidance about most likely `skills`.

relevant for this exercise type. *Word order* and *semantic parsing* can be relevant depending on the gap size and the hints provided by the exercises for the blanks. **Short Answer** exercises also practice multiple `skills` at once, at an even broader range than Gap-filling exercises. *Word order*, *construct formation*, and *semantic parsing* are always relevant for this exercise type. Depending on the learning goal the exercise encodes, it can in addition target *transposition* and *sentence formation*.

These considerations result in six `skills` that are practiced in our system: *Metalinguistic knowledge* targets a learner's ability to understand technical language and correctly identify meta-linguistic properties of language productions. *Word order* `skills` are relevant to form syntactically correct sentences. *Construct formation* targets morphology. It relates to a learner's ability to form correct inflections of a lemma while at the same time paying attention to correct agreement between dependent constituents. *Semantic parsing* is a meaning-focused `skill` that is required in order to correctly process the semantic content of an exercise's co-text. *Transposition* `skills` might be considered a sub-type of *construct formation*. However, transposing the verb form of a sentence into another tense or aspect requires to derive the correct lemma from an inflected form in addition to forming a different inflected form. We therefore keep transposition as a distinct `skill`. *Sentence formation* is a higher-level `skill` that requires the ability to use *construct formation* and *word order* `skills` in tandem. In order to emphasize that combining these two `skills` results in higher complexity than the sum of the individual `skills`, we attribute a separate `skill` to *sentence formation*.

In order to map exercises to `skills`, the exercise type constitutes a good indicator for potential `skills` that an exercise might encode. However, while `skill` annotations may initially be set automatically in exercises based on the exercise type, they should be verified manually in order to make sure that the annotations are correct. If exercises are created systematically for a specific learning goal, this verification is not necessary for each individual exercise. It is sufficient to make sure that exercises created for a certain `realization of a learning target` encode the annotated `skills`.

3.4 Deliberate Variability Impacting Exercise Difficulty

In order to support transfer of knowledge, deliberately introduced variability in exercises comprises varied exercise types as well as linguistic variability. Varied exercise types are included on the one hand to boost motivation through a less monotonous learner experience (Mettadewi et al., 2023), and on the other hand are necessary for transfer-appropriate practice (Suzuki, 2022). Linguistic variability has been shown in the previous sections to be relevant to strengthen varied declarative knowledge and transfer to other task types. It helps to prevent proceduralization from being limited to a narrow range of linguistic forms.

If, however, this variability impacts exercise difficulty, thus constituting an additional exercise difficulty parameter, it needs to be considered in exercise difficulty calibration. If RQ 5 is answered in the affirmative and exercise difficulty varies across varied linguistic forms, this adds to the already considerable challenge of calibrating exercise difficulty. Similarly, if exercise difficulty varies across different `skills`, the `skill` also needs to be considered when calibrating exercise difficulty. We therefore look into potential effects of `skills` and linguistic variants on exercise difficulty.

3.4.1 Effect of practiced `Skill` on Exercise Difficulty

Transfer-appropriate practice requires practice of knowledge for different `skills` (Lightbrown, 2007; DeKeyser, 2018). Although the definition of `skills` varies from domain to domain and from study to study, we illustrated in Section 3.3.2 how different `skills` map to different exercise types. The evaluations presented in Section 3.1.1 revealed the exercise type to impact exercise difficulty. Macro-adaptive systems should therefore support different exercise types, but be sensitive of their effect on exercise difficulty. In order to determine which exercise types are more challenging for learners, we analyzed learner submissions collected in the I4S and Didi studies (see Section 3.1.1).

In this analysis, we ranked the exercise types according to their average difficulty for each learner. The rankings were created based on the percentage of incorrect submissions for an exercise item. A submission was considered correct if the learner

answered the exercise item correctly on the first attempt. If two exercise types obtained identical percentages, they were both assigned the same rank. Exercise types that the learner did not attempt altogether were excluded from the ranking. The ranking values were scaled to the range [0, 1] for each exercise type. While aggregates over all learners reveal overall rankings, they might not be suitable for each individual learner. We therefore specifically focused on the distributions of the rankings across learners.

The results are visualized in the heatmaps in Figure 3.27, where darker colors of a matrix cell indicate higher numbers of learners for whom the exercise type is placed at the corresponding rank relative to the other exercise types. A single dark cell and white color for all remaining cells of the row would indicate perfect agreement in ranking for that exercise type among all learners. Uniform coloring of a row would indicate highest possible diversity across learners. The heatmaps show considerable differences between learners although, for most exercise types, the most frequent rank positions correspond to adjoining cells of the matrix, indicating that the difficulty placements are roughly similar for most learners. The results are clearest for Memory exercises, where the majority of learners gave the most incorrect responses among all exercise types. Jumbled Sentence and Categorization are the easiest exercise types according to these results. Short Answer and Mark-the-Words exercises constitute an interesting case as they feature a main peak as very easy, and additional less pronounced peaks as rather difficult exercise types, indicating that they are among the least critical types for some learners, and among the most critical ones for others. Single Choice and Fill-in-the-Blanks exercises are placed in middle to difficult ranks, although Single Choice exercises are solved correctly slightly more often than Fill-in-the-Blanks exercises. In addition, the heatmaps differ considerably from one `learning target` to another, sometimes even reversing ranking positions. Assuming that learners perform better on exercises if they are more proficient in the `skill` that the exercise practices, this seems to indicate that the exercises of the dataset do not encode the same `skill` for an exercise type across all `learning targets`.

In order for the exercise type to reliably predict exercise difficulty, it is therefore imperative to systematically create exercises so that all exercises of a certain exercise type target the same `skill`. Alternatively, exercises may be annotated with the

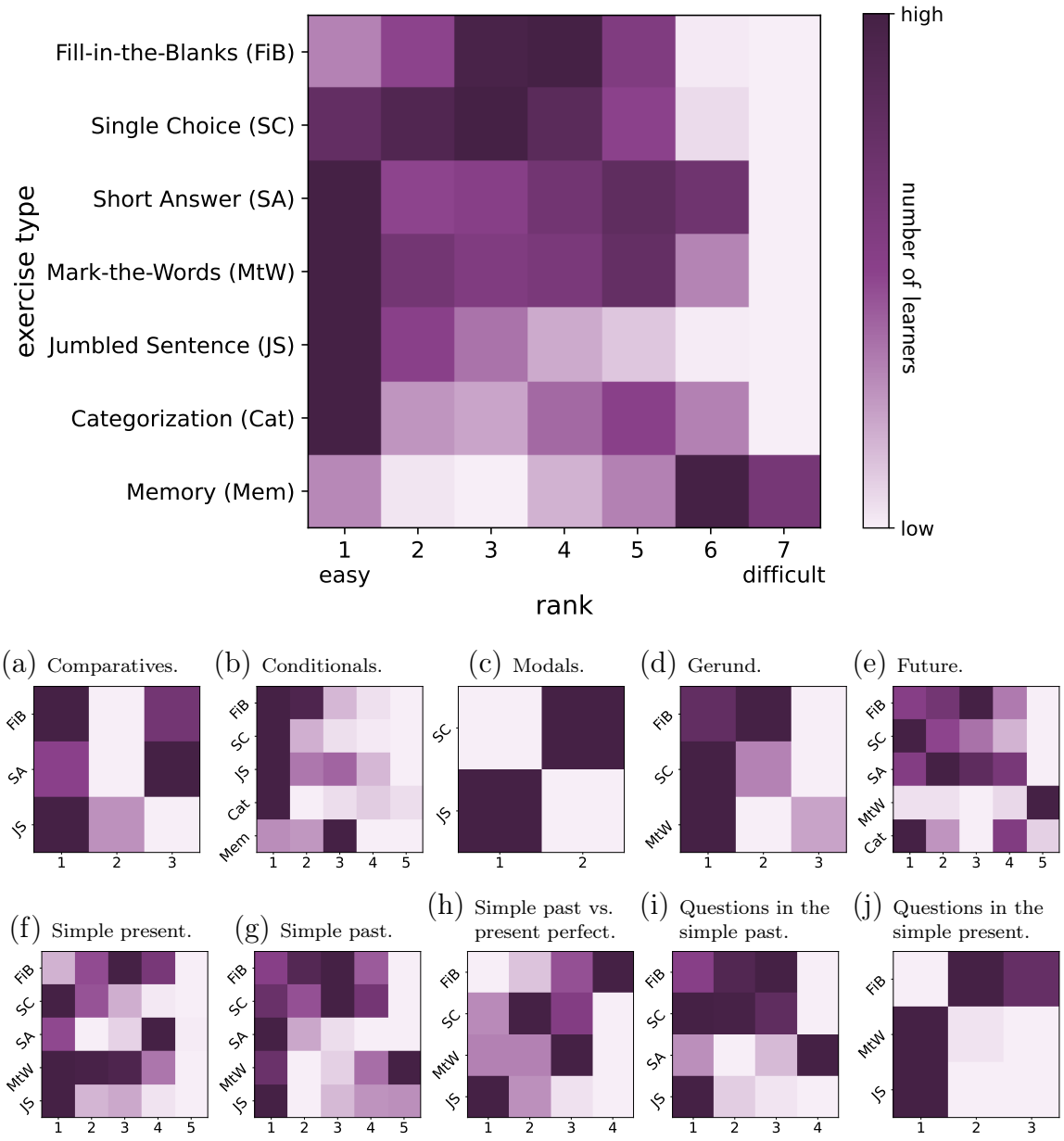


Figure 3.27: Distributions of exercise difficulty rankings.
Darker colors indicate greater numbers of learners for whom the exercise type of the respective row obtained the difficulty ranking of the respective column. Higher difficulty rankings indicate that the exercise type was more challenging for a learner.

`skill` they practice so that the `skill` can be used instead of the exercise type as exercise difficulty parameter.

3.4.2 Effect of Linguistic Variations on Exercise Difficulty

Didi's exercises of the `learning targets` *Conditionals* and *Relative clauses* used automatically generated, systematic variations of `properties`. The study therefore collected learner data for exercises with identical textual material and varying only in a selection of controlled, syntactic features. These variations target the *clause order*, *targeted clause* and *negation of clauses* for conditionals, and *clause order* for relative clauses.

Note

The exercises did not systematically introduce lexical variations. The analyses are therefore limited to the effect of syntactic variations.

In order to determine the effect of the syntactic variations on exercise difficulty, we compared the distributions of difficulty scores across the different realizations of the variations. We tested for statistical significance with a two-tailed t-test.

For the overall dataset, all effects were statistically significant, indicating that syntactic variations indeed impact exercise difficulty. In order to verify whether this is the case for each exercise type, we performed robustness checks for the individual types. The violin plots given in Figure 3.28 illustrate some variance across different exercise types. Almost all effects are statistically significant. For the clause order of conditionals, exercises of almost all types are more difficult when putting the if-clause before the main clause. The effect, although not significant (t=1.7796, p=.0760; g=.1884), is inverted for Categorization exercises. With respect to the targeted clause, exercise items are slightly more difficult if the actionable element is in the if-clause rather than in the main clause with significant effects for all exercise types. Although we hypothesize that items simultaneously targeting both clauses are more difficult, the dataset does not contain according exercises. The question whether this variation of the exercise parameter makes a difference thus remains an open research question. Concerning negation, there is a statistically significant effect indicating that exercises are easiest if only the if-clause is negated, and most difficult when both clauses are negated. The only contradictory – and non-significant – effect appears with Jumbled Sentence exercises between negation of the if-clause and of

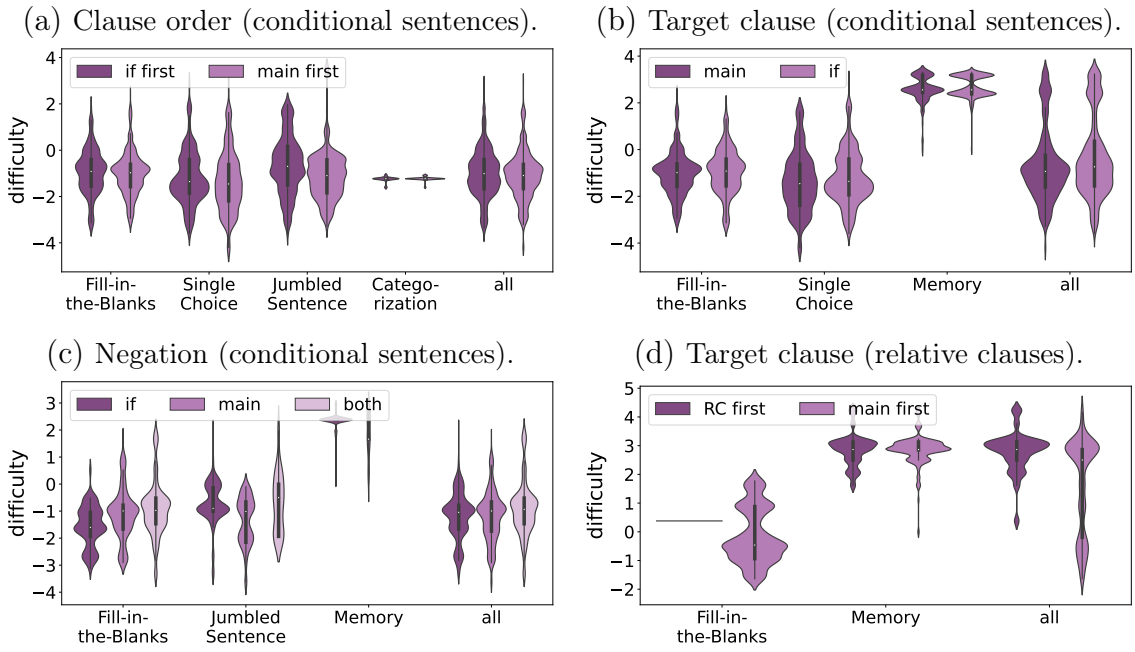


Figure 3.28: Difficulty distributions for exercises of syntactic variations. The violin plots show exercise difficulty distributions across different syntactic variations of conditional sentences and relative clauses. Difficulty varies primarily between exercise types, but also between different syntactic variations.

both clauses (t=.1993, p=.8421; g=.0264). For relative clauses, exercise items appear to be more difficult if the prompt gives the clause corresponding to the relative clause before that corresponding to the main clause. The effect is significant for all exercise types but Memory (t=-.4771, p=.6333; g=.0125).

Deliberately introduced syntactic variations thus do have an impact on exercise difficulty and generally affect different exercise types in the same way.

3.4.3 Effect of Linguistically Varied Practice on Practice Difficulty

While RQ 5 focused on differences in exercise difficulty depending on different variations of a target construct, RQ 6 aims to determine whether integrating different such variations into a practice session introduces desirable, context-related practice difficulty.

As Likourezos et al. (2019) point out, linguistically varied practice is more difficult

than less varied practice since it requires more parallel processing in working memory in order for learners to establish patterns from analyzing similarities and differences. On the other hand, Madlener (2015) found learners with prior knowledge to be less sensitive to different degrees of variability in the input than learners without prior knowledge. We therefore analyze whether practice variability impacts learners who received explanations of the varied target constructions before practice.

The analyses are based on the data described in Section 3.2.2 that was collected in the variability RCT. We again approximated exercise difficulty with accuracy on the first attempt. The box plots in Figure 3.29 indeed show higher accuracy for the control group who received less varied practice exercises, yet the effect is not significant (t=1.0860, p=.2802; g=.2197). When considering class level and English grade, there is no significant effect of the study condition either.

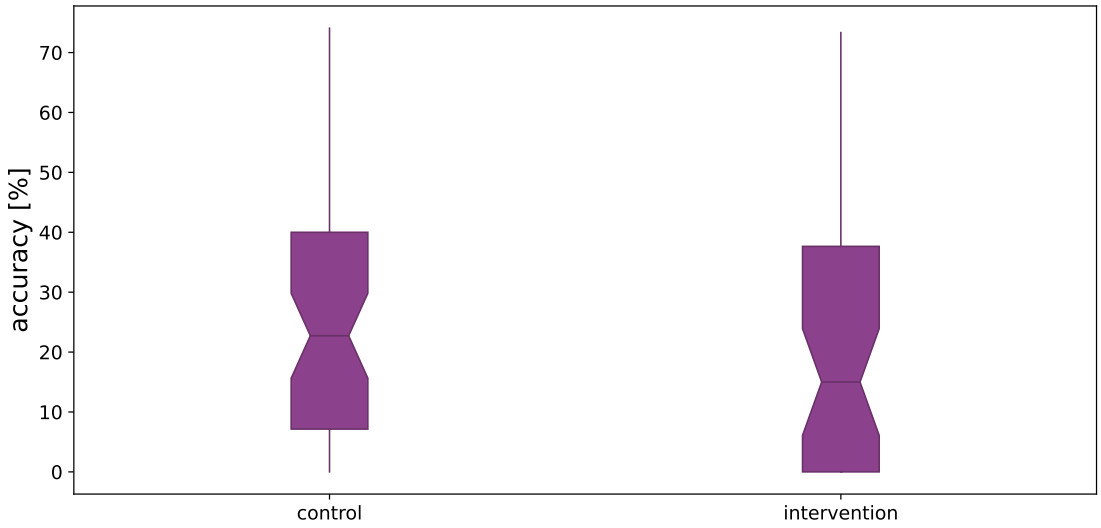


Figure 3.29: Exercise difficulty across variability conditions.
Exercise difficulty was operationalized as accuracy on the first attempt. The intervention condition achieved slightly lower accuracy than the control condition.

A very insightful observation can be obtained from the results of the working memory test. We calculated scores as F_1-scores for all items of the test. The histogram in Figure 3.30 outlines that scores were distributed across the entire range between 0 and 1 for the intervention group that were concentrated around higher F_1-scores, yet the control group had low scores throughout.

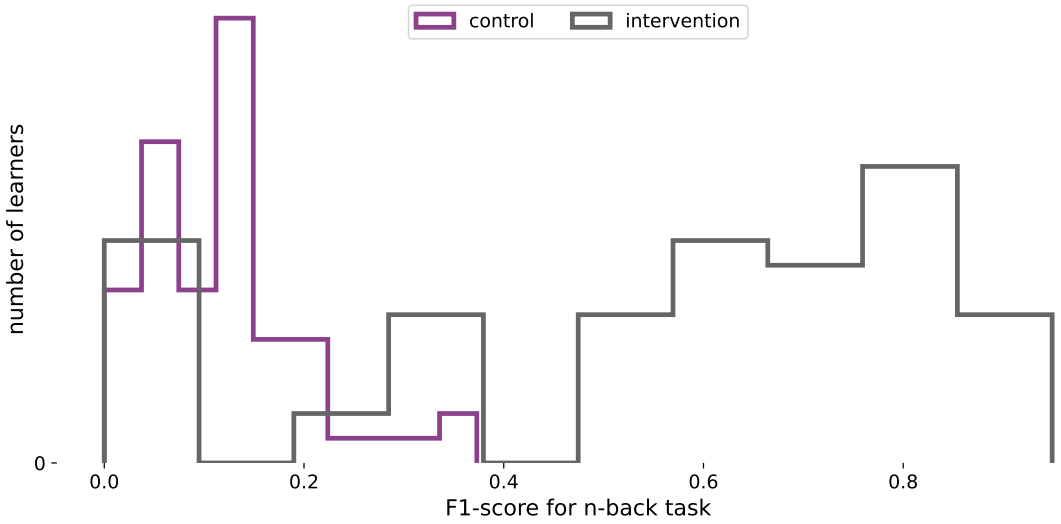


Figure 3.30: Working memory capacity after practice.
Scores of the working memory test are distributed across the entire spectrum for the intervention group and concentrated around higher accuracy scores. The control group achieved only low to moderately high scores.

The scatter plots in Figure 3.31 make it clear that this is not only the case for the aggregate sample of participants, but also within each of the participating classes for which working memory data was available.

Although we did not administer a working memory test at the beginning of the study, it is quite unlikely that learners with high working memory capacity were all randomly assigned to the intervention condition in all classes. Since the working memory test was administered after the practice phase, it is much more likely that the observed differences in scores result from the study intervention. Less varied practice thus seems to result in significantly (t=-10.7554, p=.0000; g=2.3559) lower working memory capacity than more varied practice. This might indicate that more varied practice increases brain activation more than less varied practice (Plante et al., 2015). This increased activation would then result in increased performance on the subsequent working memory capacity test. However, Mashal and Metzuyanim-Gorelick (2019) found no effect of increased brain activation on test takers' accuracy in a working memory test. An alternative interpretation suggests that in less varied practice conditions, learners recognize the possibility to apply a correct answer to the succeeding exercises (Martin and Ellis, 2012). Their increased effort to store

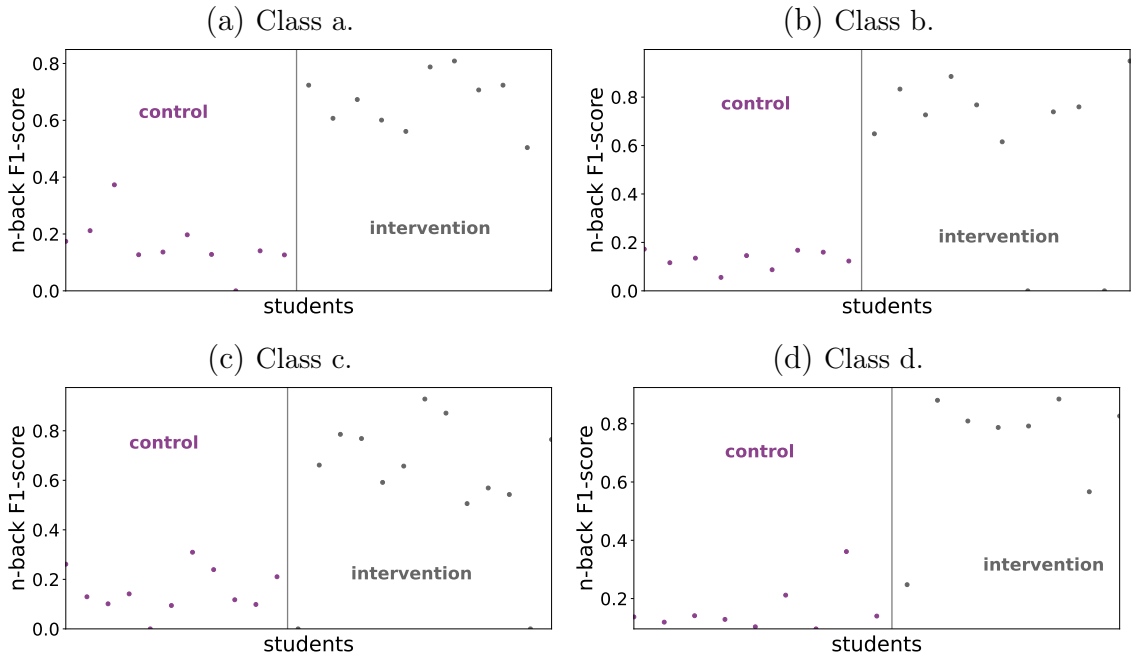


Figure 3.31: Working memory capacity after practice per class.
Each dot represents the working memory test score (y-axis) of a student (x-axis). The difference in working memory test scores between the study conditions is similar across all classes.

and process the current answer thereby exhausts the working memory and thus decreases performance in the subsequent test. Less varied practice therefore seems to put considerable strain on learners' working memory so that their working memory capacity is reduced after the practice session compared to more varied practice.

3.5 Inherent Linguistic Variability Impacting Exercise Difficulty

Some parameters of an exercise can be manipulated without changing the core of the exercise itself, such as including images and other context elements, or the number, order and kind of distractors. Linguistic features, on the other hand, cannot be manipulated without changing the exercise core. They may either be deliberately introduced to control linguistic variability, or are inherent to the source co-text. While it is possible to control exercise difficulty influenced by deliberately introduced manipulations, the co-text of an exercise typically remains unchanged, especially in

task-supported teaching. In this instructional practice, the co-text should constitute authentic language material that is not generated for the explicit purpose of the pedagogical learning goal (Thomas and Reinders, 2013). The source text can in principle constitute an arbitrary text document, yet it is subject to certain restrictions imposed in particular by a curriculum. These restrictions target primarily the semantic topic, as well as certain vocabulary and grammatical learning targets that the text must comply with.

Linguistic co-text material of exercises always contains linguistic constructions other than the learning targets. In order to process the semantic context of the exercises, learners need to have at least passive knowledge of these constructions. Approaches trying to determine difficulty based on exercise parameters have indeed found that general language parameters influence exercise difficulty (Pandarova et al., 2019). However, these approaches focus on a specific exercise type each. Since different exercise types elicit different processing of the linguistic co-material and target different `skills` (Grellet, 1981, p. 5), the relevance of individual linguistic parameters can be expected to vary from one exercise type to the other. They might not be relevant in mechanical drill exercises that practice the linguistic forms of the learning target in isolation (Wong and Van Patten, 2003). However, contextualized exercises, which practice linguistic constructions in the broader context of a coherent text, require learners to understand the clues provided by the co-text in order to give the correct answer (Walz, 1989).

In order to address RQ 7 assessing the effect of co-text complexity on exercise difficulty, we conducted a range of experiments to determine the relevance and learner dependence of co-text complexity for exercise difficulty. For these analyses, we used the dataset described in Section 3.1.3. The subset containing co-text sensitive exercises is of particular interest to these analyses as co-text complexity might be less relevant in exercises where correct processing of the text is not essential (Hoshino, 2013).

3.5.1 Relationship between Co-text Complexity and Exercise Difficulty

If exercise difficulty increases linearly with increasing co-text complexity, there should be a positive correlation between these two variables. We therefore determined Pearson's correlation between the text complexity scores and the IRT difficulty scores. Since there might not be a global relationship for all exercise types and `learning targets`, we calculated correlations for the various subsets in addition to the correlation for the entire dataset.

The results, summarized in Table 3.14, show that there is no linear relationship between co-text complexity and exercise difficulty, neither for all exercises ($|\rho| = .10$) nor for those of the individual I4S ($|\rho| = .01$) and Didi ($\rho = .14$) studies. The values vary considerably between `learning targets` ($|\rho| = .01$ for *Future Tenses* to $|\rho| = .73$ for *Modals*) and exercise types ($|\rho| = .00$ for Fill-in-the-Blanks to $|\rho| = .33$ for Jumbled Sentence). For the subsets comprising combinations of `learning targets` and exercise types, the differences across combinations are equally pronounced ($|\rho| = .02$ for Fill-in-the-Blanks exercises on *Simple past vs. Present perfect* to $|\rho| = .83$ for Single Choice exercises on *Conditionals*[8]). There is no linear re-

Exercise set	ρ	**Sample size**
All	.0991	3,104
I4S	.0076	1,101
Didi	.1381	2,003
Future Tenses	-.0094	127
Modals	.7270	34
Fill-in-the-Blanks (FiB)	.0022	1,849
Jumbled Sentence	.3337	444
FiB – *Simple past vs. Present perfect*	-.0231	241
Single Choice – *Conditionals*	.8291	8
Core	.0024	131
Co-text sensitive	.0804	208

Table 3.14: Correlation of co-text complexity with exercise difficulty.
Pearson's correlation ρ of text complexity with exercise difficulty varies considerably across different exercise subsets.

[8]We excluded combinations with sample sizes $|n| \leq 2$, although sample sizes may be too small in other cases ($4 \leq |n| \leq 385$) as well to yield reliable results.

lationship ($|\rho| = .00$) for the subsets containing only key exercises, which most learners submitted, or only co-text sensitive exercises ($|\rho| = .08$). Interestingly, some correlations are negative, suggesting that exercises are more difficult when co-text complexity is lower. While this might be due to insufficiently large sample sizes, it could also indicate that exercise creators try – consciously or subconsciously – to compensate some difficulty features with others in order to create exercises of overall approximately similar difficulties. The results, while not entirely conclusive due to data sparseness considering the multitude of parameters influencing exercise difficulty, suggest that co-text complexity does not have the same effect on exercise difficulty for all `learning targets` and exercise types. There is no overall linear relationship between these two parameters.

For the subsets controlling for exercise difficulty, the difficulty values by definition differ only marginally. We therefore determined the mean as well as the minimum and maximum text complexity scores within these subsets and compared them between the sets. Following the logic that higher text complexity scores result in higher exercise difficulty, these metrics should then be lowest for the subset of low-difficulty

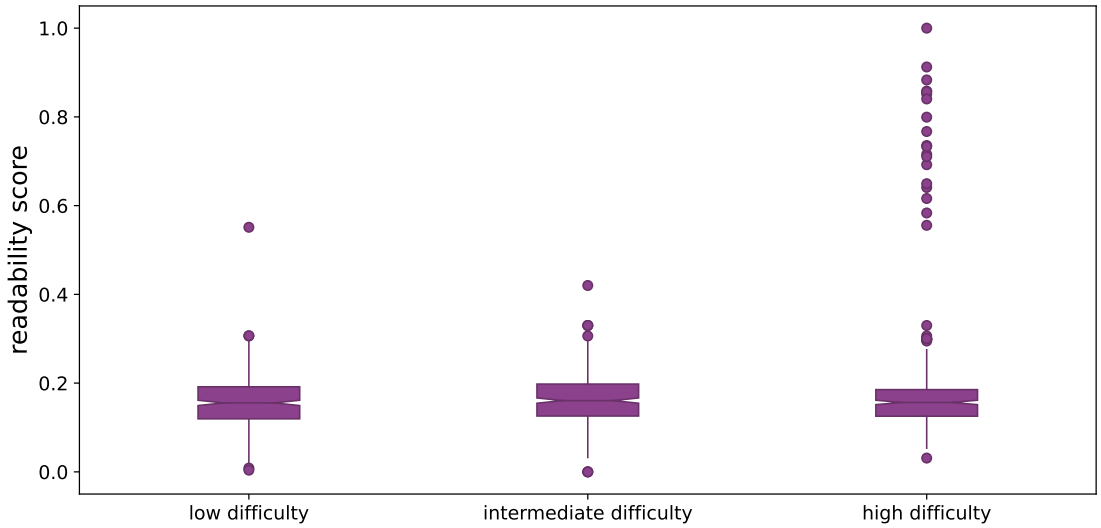


Figure 3.32: Text complexity score distributions for subsets of controlled difficulty.

Text complexity score distributions are rather similar for exercises of low, intermediate and high difficulty. Very high text complexity scores appear only with exercises of high difficulty.

exercises and highest for the subset of high-difficulty exercises. However, the box plots in Figure 3.32 illustrate that text complexity scores are very similar for all three subsets, with values ranging from .0000 to .4632 (μ = .1390), from .0172 to .3841 (μ = .1503), and from .0074 to 1.0 (μ = .1776) for low-, intermediate-, and high-difficulty items respectively. It should be noted, though, that very high text complexity scores appear only with high-difficulty exercises, which could indicate that such high text complexities might indeed have an influence on overall exercise difficulty.

In order to examine non-linear relationships between co-text complexity and exercise difficulty, we trained AdaBoost regressors from the Python `scikit-learn` library to predict exercise difficulty based on simple exercise features as outlined in Section 3.1.3 and co-text complexity. We focused on the sub-sets of exercises based on the exercise type and the subset of co-text sensitive exercises. The results, outlined in Table 3.15, indicate that integrating co-text complexity even reduces the amount of variance explained by the model. Co-text complexity does therefore not successfully predict exercise difficulty.

	All features	**Without co-text complexity**
All exercises	.2135	.2142
Categorization	.0000	.0000
Memory	-.9393	-.0489
Mark-the-Words	-1.0609	.1872
Jumbled Sentence	.1017	-.0135
Single Choice	.0099	-.0063
Fill-in-the-Blanks	.2322	.2540
Short Answer	.2587	.2811
Co-text sensitive	.1563	.2148

Table 3.15: Explained variance in regression models predicting exercise difficulty. The full feature set encompassed simple exercise features and co-text complexity. The models without co-text complexity as predictor throughout outperformed the models with co-text complexity in terms of explained variance.

Overall, we found neither any linear relationship between co-text complexity and a learner's performance on the exercise, nor could machine learning models capture a significance of this parameter in combination with other exercise specific features.

3.5.2 Learner-dependent Relevance of Co-text Complexity

The analyses presented in Section 3.5.1 could not identify an impact of co-text complexity on exercise difficulty. Considering that it tends to be aligned with general language proficiency, operationalized for instance as C-test scores, in readability assessment applications (Chen and Meurers, 2019), it is possible that the relevance of co-text complexity depends on the learner's general language proficiency. A first naive comparison of mean differences between a learner's general language proficiency in terms of C-test scores and an exercise's co-text complexity score indeed reveals that this difference is on average higher for correctly answered exercises ($\mu = .1646, \sigma = .1053$) than for incorrectly solved ones ($\mu = .1436, \sigma = .1030$).

We next determined whether the relevance of co-text complexity as exercise difficulty parameter is relevant only within a specific range of general language proficiency. To this purpose, we compared the performance of classifiers for the subsets of controlled general language proficiency including only submissions of learners with low, intermediate, or high language proficiency. In this analysis, we used co-text complexity as a single predictor. If the predictive power of co-text complexity varies with the learners' proficiency levels, we expected performance to differ between the subsets.

The results indeed show differences in model performance, which is best for high learner proficiency ($F_1 = .7755$) and lowest for low proficiency ($F_1 = .6627$). However, a majority baseline model shows a similar effect of higher F_1-scores for the subset of high-proficient learners, indicating that the dataset is not balanced. When resolving this by means of over-sampling, the effect is quite shifted: The model now performs best for learners of intermediate proficiency ($F_1 = .6251$), which is considerably better than the majority baseline model ($F_1 = .2704$). Co-text complexity thus has the greatest impact at intermediate proficiency levels. A possible explanation suggests that lower-performing learners use more local clues of the actionable element itself to solve grammar exercises while high-performing learners can process co-texts of arbitrary complexity. Learners of intermediate proficiency, on the other hand, seem to make use of top-down skills to solve grammar exercises. Adaptivity algorithms taking co-text complexity into consideration need to identify means to determine the thresholds defining intermediate general language proficiency for their

target group of learners so that co-text complexity can be considered exclusively for those learners for whom it does make a difference.

Note

Other learner characteristics might also have an influence on the predictiveness of co-text complexity for exercise difficulty. Further learning analytics in the line of work by Deininger et al. (2025) can shed light on this question.

3.6 Summary

In this chapter, we investigated the relevance of variability in grammar practice. In answer to RQ 3, our analyses did not yield conclusive results, yet they allow us to form hypotheses directing future research. We hypothesize that (1) knowledge transfers across variations of a linguistic form, although in contrast to the lexical and semantic levels, syntax is often not sufficiently productive in that it would offer enough variations of a characteristic for learners to benefit from varied practice. (2) Varied practice is beneficial for far transfer relying on declarative knowledge, but negatively impacts near transfer relying on procedural knowledge. (3) Linguistic variability is necessary in order to foster co-occurrence of variedness with proceduralization in order to prevent under-use of some linguistic forms (Schmidt, 1990). If future research attests these hypotheses to be true, they answer RQ 4 partially in the affirmative, negating that varied practice is beneficial for proceduralization but confirming that it results in more varied learner productions. In order to strike a balance between practice of linguistic forms in breadth and in depth, exercise selection that aims to globally maximize or minimize variability in practice material is not suitable. Instead, we propose to reduce within-exercise variability as much as possible while learners have not mastered the learning target, and at the same time increase between-exercise variability. Once the learner fulfills the mastery criterion, within-exercise variability should also be maximized. Revision learning units, for which learners are presumed to have already mastered the learning targets at an earlier point in time, can aim for high variability as well.

Having identified the need to systematically introduce variability into practice exercises, we introduced a complex domain model that, although labor-intensive in creation, considerably simplifies automatic annotation of exercises in preparation

for macro-adaptive sequencing. Exercises need to be linked to KCs so that a macro-adaptivity algorithm can select suitable exercise candidates. In order to be able to focus on linguistic and `skill` variability, these two parameters of exercises define KCs. For the purpose of linking exercises to KCs, the linguistic annotations – the linguistic dimension of KCs – as well as the exercise type – the `skill` dimension of KCs – can be automatically determined and exercises can be linked to KCs thanks to the hierarchical structure of the domain model. In order to also take into account the pedagogical level, which is particularly relevant in school settings, exercises also need to be linked to a node on a pedagogical level of the domain model. Only then can the macro-adaptivity algorithm determine whether the exercises are within the scope granted to it by the curricular constraints. It is the responsibility of exercise creators to establish this link.

In order to assess the effect of controlling variability in practice exercises, we looked at the impact of both inherent and deliberately introduced linguistic variability on exercise difficulty. The evaluations showed that both sources of variability influence exercise difficulty. This implies differences across otherwise similar exercises depending on the linguistic forms practiced by the exercise, as referred to in RQ 5, as well as, for learners of intermediate proficiency, forms contained in the co-text, as referred to in RQ 7. The question whether exercise difficulty varies across linguistic forms of similar properties is thus answered in the affirmative, as is the question whether co-text complexity affects the difficulty of form-focused grammar exercises. The impact of linguistic variability on exercise difficulty also implies differences in context-dependent exercise difficulty of the practice session depending on the overall variedness of linguistic forms targeted in that session, as referred to in RQ 6. While simply introducing high variability into practice material does not contribute to desirable difficulty, controlling the linguistic variability does. Answering this research question therefore requires some more distinguished considerations. While inherent variability of co-text material is primarily relevant to learners of intermediate proficiency, this source of variability cannot be prevented. As we have seen in Chapter 2, exercise-inherent linguistic variability cannot be entirely controlled in school settings where the selection of text material often underlies some restrictions. This is, in fact, a challenge with other context-related difficulty factors as well. The timing of instruction (Suzuki et al., 2019), for instance, is often subject to a course schedule and cannot be freely manipulated. Being prey to one's circumstances and

surroundings, exercise difficulty may thus at times be lower or higher than desirable, without the ILTS being able to monitor and adjust to those factors.
In addition, variability in the second factor of KCs, namely `skills`, affects exercise difficulty as well. Although each `skill` is often realized as a particular exercise type, inconsistent coding in terms of exercise-to-`skill` matching may result in unreliable difficulty estimates based on exercise types.

Selecting exercises matching a learner's proficiency level while at the same time acknowledging the existence of as well as the need for variability is therefore extremely challenging. Macro-adaptive approaches will most likely need to compromise and are able to provide merely approximations of good matches even with the most sophisticated implementation. Reducing the focus on exercise difficulty in macro-adaptive exercise selection would, however, make it possible to instead focus on linguistic variability and curricular restrictions. We will discuss this possibility in the next chapter.

Chapter 4

Adaptive Exercise Selection in the Zone of Proximal Development with Reduced Focus on Exercise Difficulty

Parts of this chapter are based on Publication 4.

The previous two chapters have outlined the challenges of considering curricular structures and linguistic variability in a macro-adaptive ILTS. While considering either one of the two in isolation poses substantial challenges, accounting for both at the same time while trying to select exercises matching a learner's proficiency has rather limited prospects of success.

Although we agree that all learners should be able to solve the exercises presented to them, rather small differences in exercise difficulty need not be considered in macro-adaptive exercise sequencing. Moeyaert et al. (2016) did not observe any significant effects in their evaluation of the impact of item difficulty of French verb inflection exercises on learning outcomes. As a possible explanation, they acknowledge that the differences in difficulty scores of their items might not have been high enough. In addition, we must not forget that adaptivity does not only comprise macro-adaptive exercise sequencing. Micro-adaptive scaffolding aspires to provide sufficient assistance so that learners can perform beyond what they would be capable of doing on their own (Vygotsky, 1978). A new form of adaptivity operating in-between macro- and micro-adaptivity, which we call *meso-adaptivity*, might therefore be able

to mitigate exercise difficulty in the learner's ZPD to a certain degree. If this is the case, then incorporating meso-adaptive adjustments in addition to macro-adaptive exercise sequencing makes it possible to more dynamically handle the responsibility of making sure that a learner can successfully complete an exercise.

The resulting approach incorporates two adaptivity strategies: The macro-adaptive strategy focuses on selecting linguistically varied exercises within the scope of curricular constraints, whereas the meso-adaptive strategy focuses on modulating an exercise's difficulty within the ZPD. Such an approach has the additional benefit of addressing both the issue of performance of ITSs raised by Pelánek and Řihák (2017) and the cold-start problem of data-enabled macro-adaptivity algorithms (Abdi et al., 2019).

In this chapter, we address RQ 8 by presenting and evaluating macro-adaptivity with a reduced focus on exercise difficulty in combination with meso-adaptive adjustments.

Research Questions

RQ8 Can meso-adaptivity mitigate exercise difficulty sufficiently to enable learners to solve (almost) all exercises?

4.1 Reducing the Focus on Exercise Difficulty in Macro-adaptivity

In most adaptivity implementations, exercise selection for adaptive grammar exercises consists of two steps: Filtering and ranking based on the exercises' fit according to the learner model (Effenberger, 2018). Since filters reduce the amount of exercises that need to be considered in the subsequent exercise selection steps, the approach that we present here uses these instead of rankers whenever possible in order to reduce processing time.

The implementation of the macro-adaptive exercise selection is kept as simple as possible for four reasons: (1) A simple, rule-based approach makes it possible to tweak certain parameters of the implementation that are subject to a research question. The simpler the implementation, the easier it is to introduce various study

conditions. (2) Complex implementations result in trade-offs between breadth of features and implementation depth of all features (Pelánek, 2017). (3) From a practical perspective, complex, machine learning based models have high maintenance costs (Pelánek, 2017). (4) Most importantly, for the goals of our adaptivity approach, namely minimizing or maximizing variability of `linguistic constructs`, there is no benefit of machine learning-based approaches over simple statistical calculations.

4.1.1 Exercise Filters for Coarse-grained Exercise Selection

Our exercise filters ensure that exercises are selected that provide opportunities for the learner to progress towards mastery of some underlying `property` or `skill`. Depending on the entry point (see Figure 4.1), the algorithm considers only exercises of the selected `realization` (entry points 2) or exercises of all `realizations` of the `learning target` (entry point 1).

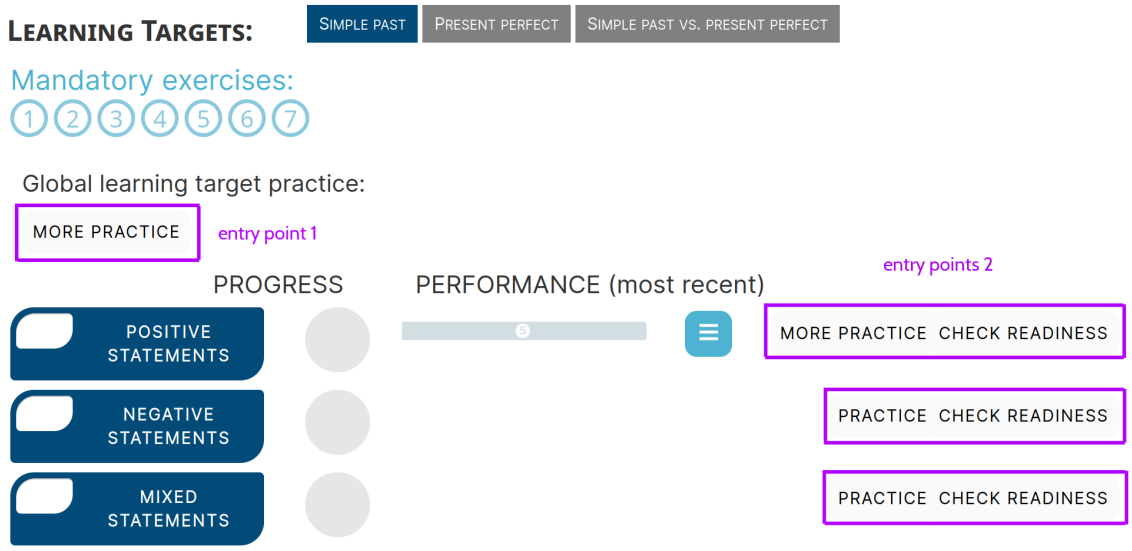


Figure 4.1: Entry points for practice in AI2Teach.
Learners can request an adaptively selected exercises from one of two entry points: Entry point 1 considers all exercises of the `learning target`. Entry point 2 considers only the exercises of the corresponding `realization`.

If all `realizations` of the `learning target` are selected, the first step consists in prioritizing the `realizations`. `Realizations` for which the learner has already

successfully completed the diagnostic exercise, which indicates readiness for the TSLT-inspired target task, are only considered if there are no exercises for any other `realization` without successfully completed diagnostic exercise.

Within each `realization`, mandatory exercises receive special treatment; they are given highest priority. Therefore, if a `realization` has already been mastered but there are not yet completed mandatory exercises, the algorithm considers only those. Once all mandatory exercises have been submitted, additional filters are applied to the remaining exercise candidates. Inside a `realization`, `properties` constitute the central milestones that the exercises practice. Effenberger (2018) distinguishes between (1) mastery decision, (2) curriculum sequencing, and (3) task selection for macro-adaptivity. Applied to our implementation, mastery decision corresponds to determining whether the last practiced `property` has been mastered; curriculum sequencing corresponds to determining the next `property` to practice; and task selection corresponds to selecting the specific exercise. Mastery Learning presents a straightforward answer as to whether the last practiced `property` has been mastered. In line with Effenberger's (2018) three steps, the algorithm thus checks for each `realization` whether all its `properties` but not all `skills` have been mastered (*mastery decision*). If this is the case, the algorithm tries to find an exercise for this `realization` while only exercises that practice a missing `skill` are considered. Mastery is tracked for each `skill` on two levels. The *interactive* level requires learners to practice the `skill` in an interactive manner in closed exercises. The *productive* level requires learners to practice the `skill` in a productive manner in half-open exercises. If there are any not yet mastered interactive `skills`, only these are considered; otherwise, productive `skills` are considered. This way, each `skill` has to be mastered in closed exercises first before the learner is exposed to more open exercises. If multiple `skills` have not yet been mastered, they are considered in rotating order so that the last practiced `skill` is given lowest priority (*curriculum sequencing*). The initial priority assignment follows the pedagogically motivated default order of `skills` specified in the domain model. Similarly, if there are multiple `realizations` with all `properties` but not all `skills` mastered, the algorithm considers them in the default order specified in the domain model.

If there is no `realization` for which the learner has mastered all `properties` but not all `skills`, the algorithm applies a filter based on the `properties` the exercises

practice. If the last practiced `property` has not yet been mastered, the algorithm focuses on this `property` (*mastery decision*). Otherwise, if the learner has fulfilled the mastery criterion, the algorithm moves on to the next `property`. The `properties` are practiced in the default order specified in the domain model (*curriculum sequencing*). Depending on the learner's proficiency level, however, certain shortcuts are possible. The proficiency level is maintained in the learner model. Learners of advanced proficiency can skip practice of one `property` if it can also be mastered by mastering a succeeding `property`. This is the case if at least one of its `linguistic constructs` is identical to a `linguistic construct` of a succeeding `property`. If the algorithm finds exercises for a `property` that it considers, these exercises constitute candidates that are passed on to the succeeding filters, and no further `properties` are considered.

If all `properties` and all `skills` have been mastered already, the algorithm considers all exercises of all `realizations` of the `learning target`, or of the `realization` selected by the learner.

In the next step, the concrete exercise is selected (*task selection*). Regardless of whether the pool of exercises resulting from the previous filtering steps is based on the `skill`, `property`, or fallback filter, the algorithm now checks whether the learner has already practiced the exercises. All exercises that share the same co-text are grouped in an exercise set. This enables the algorithm to prevent selecting exercises whose co-text the learner is already familiar with. The exercises within each set differ in exercise type and linguistic variations. If any of the exercise candidates belong to an exercise set that the learner has not yet practiced, only these exercises are considered further. Otherwise, exercises from an already practiced exercise set where that concrete exercise instantiation has not yet been practiced are considered. If no exercise passes this last filter, already practiced exercises are considered. The last practiced exercise is then excluded from the candidate list. This prevents learners from getting the same exercise over and over in case there are not enough exercises to practice a specific KC. In addition, closed exercises are prioritized if the learner has not yet mastered closed exercises for that `property`. Otherwise, the algorithm prioritizes half-open exercises. This filter integrates a minimum of difficulty control since closed exercises are generally easier than more open exercises (Lee and Muncie, 2006). If none of the remaining exercise candidates match, candidates of

non-optimal openness are considered. If there are any mandatory exercises in the filtered exercises, these are given priority and only they pass the filter.

These filters in most cases considerably reduce the pool of exercise candidates. The remaining, manageable pool of candidates is passed on to the next step, which ranks the filtered exercises.

4.1.2 Exercise Rankers for Fine-grained Exercise Selection

An exercise ranker in our implementation ensures that the learner receives varied linguistic contexts and that co-text is optimally adjusted to the learner's general language proficiency.

If only exercises could be found that have already been practiced, the ranking algorithm considers the order in which they were practiced before. However, it does not take the exact same order but instead divides the exercises into two groups. Exercises in the first group were practiced by the learner before the median submission timestamp of all submissions by the learner. Exercises in the second group were practiced after that timepoint. The algorithm then rejects all exercises of the second group so that only those 50% of the exercises can be issued to the learner again that were practiced rather long ago. Out of these exercises, the ranker selects a random one for the learner.

If the exercises that passed the filters have not been practiced before, linguistic variability and co-text complexity are considered.

The co-text complexity score borrows from the field of readability assessment and determines the fit of an exercise's co-text to the learner's general language proficiency (Chen and Meurers, 2019). The learner's proficiency level is maintained in the learner model. The item's co-text complexity can be determined when the exercise is created through automated calculations based on the co-text and must be stored with the exercise in the system. We based complexity scores on aggregates of a range of readability formulas. This approach is analogous to the implementation described in Section 3.1.3. Since the evaluation in that section was based on data representing the target group of our implementation, we adopted the thresholds that

resulted from those evaluations for converting the continuous scores into discrete complexity classes of three levels. The difference between learner proficiency and co-text complexity is averaged over all exercise items in order to calculate a co-text complexity score for an entire exercise.

Note

Alignment of learner proficiency with co-text complexity merely serves as a proof of concept to demonstrate that aspects of the co-text, in this case its linguistic complexity, can be considered in a macro-adaptive approach focusing on varied practice. Applications for real-world use are more likely to, for instance, focus on alignment of the topic with the curriculum the learner follows.

The linguistic variability score is calculated based on the `linguistic constructs` practiced by the exercise. To this purpose, the exercises must be enriched with linguistic annotations relevant to the `learning target`. Based on these annotations and Formula 4.1, the algorithm calculates the linguistic variability score s_v. a_e denotes the annotations of the actionable element. a_l denotes the annotations previously practiced by the learner. p_l denotes the learner's proficiency level. p_e denotes the proficiency level targeted by the actionable element. Low proficiency corresponds to the Integer *1*, intermediate proficiency to *2*, and advanced proficiency to *3*.

$$s_v = \frac{\mid a_e \setminus a_l \mid}{p_l - p_e}, \qquad \text{where } p_e = \begin{cases} 1, & \text{for } \mid a_e \mid \leq 2 \\ 2, & \text{for } 2 < \mid a_e \mid < 8 \\ 3, & \text{for } \mid a_e \mid \geq 8 \end{cases} \tag{4.1}$$

Looking at this formula in more detail, the initial score is the sum of all never practiced linguistic annotations of the actionable element. From this initial score, the linguistic variability score is determined based on the learner's proficiency. While high-proficient learners are exposed to as many linguistic contexts within an exercise as possible, the number is kept as low as possible for less proficient learners. The higher the learner's proficiency level, the higher the number of distinct linguistic annotations that is considered ideal for the learner.

Note

The results from the variability study described in Section 3.2.2 were not available at the time of implementing the algorithm. The implementation is therefore based on research on working memory that found learners with low working memory capacity to require less varied input (Lv and Liang, 2019). Our productive system did not collect overhead data in the form of working memory capacity values. We therefore relied on general language proficiency instead in order to approximate this learner characteristic.

The ideal number of annotations per proficiency level was determined based on the distribution of the number of annotations across exercises. An actionable element in our system has a mean of $\mu = 4.5882$ linguistic annotations with a standard deviation of $\sigma = 2.1225$. Although the distribution is not normal (with $\alpha = 0$ for Shapiro-Wilk and Kolmogorov-Smirnov tests; $\alpha > .05$ for normal distribution), Figure 4.2 shows that the distribution approximates a bell-shaped curve. We therefore applied thresholds of $\mu \pm \sigma$ to determine the optimal number of linguistic annotations per proficiency group. For low-proficient learners, the optimal number of annotations was set to $n < \mu - \sigma$. For learners of intermediate proficiency, the optimal number of

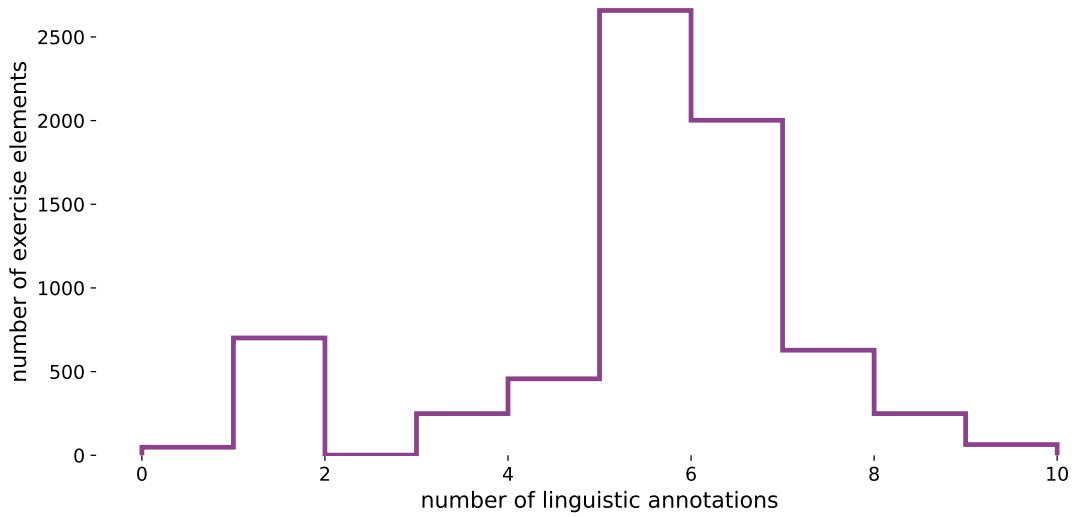


Figure 4.2: Numbers of linguistic annotations per actionable element.
The numbers of linguistic annotations per actionable element range from 0 to 10. The distribution is not normal, but it approximates a bell-shaped curve.

annotations was set to $\mu - \sigma \leq n \leq \mu + \sigma$. For high-proficient learners, the optimal number of annotations was set to $n > \mu + \sigma$. Rounding up to whole numbers, this resulted in an optimum of 0 to 2 annotations for beginner learners, 3 to 7 annotations for intermediate learners, and 8 or more annotations for advanced learners. The algorithm determines the distance between the proficiency group corresponding to the actionable element's number of annotations and the learner's actual proficiency group according to the learner model. Dividing the initial score by this distance results in the final linguistic variability score.

These two scores, co-text complexity and linguistic variability scores, are calculated for all exercise candidates. In order to determine an aggregate score, they are each normalized into numbers between 0 and 1 over all exercises. When summing them up per exercise, weights are set so that the co-text complexity has half the weight of linguistic variability. The exercise score s is thus calculated with Formula 4.2 from the normalized co-text complexity score s_c and the normalized linguistic variability score s_v.

$$s = 0.5s_c + s_v \tag{4.2}$$

The highest-ranked exercise is then provided to the learner.

4.2 Mitigating Exercise Difficulty through Meso-adaptivity

In the presented macro-adaptive implementation, exercise difficulty is only considered indirectly through manually determined sequences of `realizations`, as well as at the level of exercise openness. The approach relies on the assumption that meso-adaptive adjustments of exercise elements provide enough support to enable learners to solve any exercise selected by an algorithm with this minimal focus on difficulty. In order to test this assumption, we implemented a number of meso-adaptive adjustments for six exercise types and evaluated their mitigating effect on exercise difficulty using the data from the large-scale field study of the AI2Teach project.

4.2.1 Solvability of Exercises with a Restricted Answer Space

In socio-cultural contexts also referred to as scaffolding (Elbers et al., 2013), micro-adaptivity aims to support learners in finding the correct solution in response to learner errors, for instance through step-by-step feedback and hints (Monteira et al., 2022).

This focus is slightly shifted in the context of Self-Regulated Learning (SRL), where micro-adaptive scaffolds have received considerable attention as well. Implemented as prompts, tools, pedagogical agents, strategies, or feedback, **SRL scaffolds** help learners to identify relevant learning goals and learning objects based on their learner model (Lim et al., 2023; Zheng, 2016). Research in this field has made considerable advances in the categorization of scaffolds. Classifications may target different functionalities comprising conceptual, metacognitive, procedural, and strategic functionalities. When focusing on the delivery mode, they contrast scaffolds embedded into the system to those initiated by the learner outside the system, fixed scaffolds to learner-adaptive ones, hard scaffolds to soft ones in terms of customizability of when to provide them, and direct scaffolds to indirect ones in terms of explicit training vs. prompts.

In adaptive ITSs, a range of scaffolding strategies has been explored as well. Somewhat related to macro-adaptivity, **next-step hints** can be adjusted to the learner model for problems that can be structured into sub-problems (Koedinger et al., 2013). Where multiple paths towards solving the exercise are possible, the next suggested step should take the learner's answer up to that point into account (Aleven et al., 2016). Most implementations rely on data-driven approaches to decide which step to scaffold next. *Proofs Tutorial*, for instance, uses a Markov Decision Process (Barnes and Stamper, 2010). *DT Tutor* relies on a Bayesian Network for this decision (Murray et al., 2004).

Tutorial Dialogue Systems tailor the **feedback message** to the learner's answer. For example, the ITS *Rimac* for Physics implements micro-adaptive scaffolding in the form of questions in different variations that differ in the level of support they offer, feedback for learner answers, hints, and explanations (Katz et al., 2021). The conversational tutoring system for natural language practice *AutoTutor* provides binary

feedback and hints formulated as Socratic questions (Nye et al., 2014). McKendree (1990) found that, for problems that can be broken down into multiple steps, beginners benefited more from feedback that pinpointed the goal of the current step in solving a geometry problem than from feedback specifying which condition was being violated in the currently applied rule. In the domain of Maths, the *Assistment System* breaks down exercises into smaller parts if learners struggle with the original question (Razzaq and Heffernan, 2006).

In macro-adaptive ILTSs, where the learning objects are automatically and adaptively assigned to learners, micro-adaptivity is usually limited to adaptive feedback (Bimba et al., 2017). While some implementations also provide static support such as hints, links to information pages, or additional information displayed as tooltips, they do not provide them micro-adaptively as the learner struggles to complete an exercise item (Antoniadis et al., 2004). Feedback may be adjusted to the learner's profile in terms of language variation and type. Meta-linguistic feedback can use either the learner's native or the target language (Choi, 2016; Heift and Nicholson, 2001), and simple or technical language (Heift, 2001). In terms of feedback type, Shute's (2008) overview presents different types ranging from binary feedback to giving partial answers. Combinations of multiple feedback types are also possible. Merely giving the correct response does not constitute scaffolding towards the correct answer, even if elaborate explanations are provided of why the learner's answer is incorrect and the target answer is correct. The more sophisticated feedback types constitute scaffolds of increasing elaborateness. A minimal scaffold referred to as *error flagging* points the learner to the location of the error. Textual elaborations can target specific attributes of the `learning target` (*attribute isolation*) or the target topic, often by repeating the general rule or piece of information (*topic contingent*). Explanations regarding the learner's misconception rely on formal error diagnosis. Hints, cues, and prompts comprise guidance for the next step to take and examples illustrating the rule that needs to be applied. According to Shute (2008), more specific feedback, in particular feedback focusing on the learner's error, is most effective. She also suggests that, the more proficient a learner becomes, the less specific feedback should be. While the *German Tutor* applies this principle by selecting one of multiple equivalent feedback messages according to their specificity (Heift and McFetridge, 1999), examples from non-linguistic domains present alternative implementations. They vary the feedback type with the number of at-

tempts on an exercise, so that more proficient learners receive only the first, more general, feedback messages whereas less proficient learners receive more and more specific feedback after each additional attempt. For instance, *Coach* applies multiple levels of hints to each error, starting with a rule required for the solution and that constitutes a misconception of the learner, next outlining all possible next steps, then selecting the best next step, before explaining the best next move (Burton and Brown, 1979). Similarly, the *SQL-Tutor* increases the level of detail of its feedback messages with each attempt of the learner on the same exercise, ranging from only pointing to the location of the error to giving the solution (Mitrovic et al., 2002). A system can thus provide multi-step scaffolding support through feedback by gradually adjusting the language and/or the type of the feedback.

Systems incorporating only micro-adaptivity have reported high learning gains, although their application in real-world contexts remains to be evaluated (Rus et al., 2015). Meta-linguistic feedback in particular has been shown to increase learning gains in language learning (Meurers et al., 2019). However, providing meta-linguistic feedback is only possible if the nature of the learner's error can be reliably and correctly identified. Despite significant advances in error diagnosis, this is not always possible for all learner answers. In order to still provide scaffolding support when learners struggle, alternative adaptive adjustments need to be implemented. Different strategies may be more or less suitable for different learners. Adjusting the instructions, for instance, may help learners who struggle to correctly understand the learning goal. Instructions can be made available in multiple languages and provide more or less detail. For example, Mark-the-Words exercises become more specific if the instructions indicate the number of words to find in the text. If the learner's error does not relate to a lack of understanding of the task at hand, exercise elements provide considerable potential for scaffolding similar to multi-step feedback. As illustrated in Figure 4.3, this is especially noticeable with Gap-filling exercises. The initial representation uses text fields as input type and the lemmas of the target words given as a bag of words with additional semantic distractors that do not fit into any gap. The first scaffolding step may consist in eliminating the semantic distractors ①. The next step might place the lemmas in parentheses behind the gaps so that the learner needs only produce the correct form ②. Still failing to do so, the learner would then receive options of potential forms, thus changing the input type to a drop-down ③. This input type still supports additional scaffolding

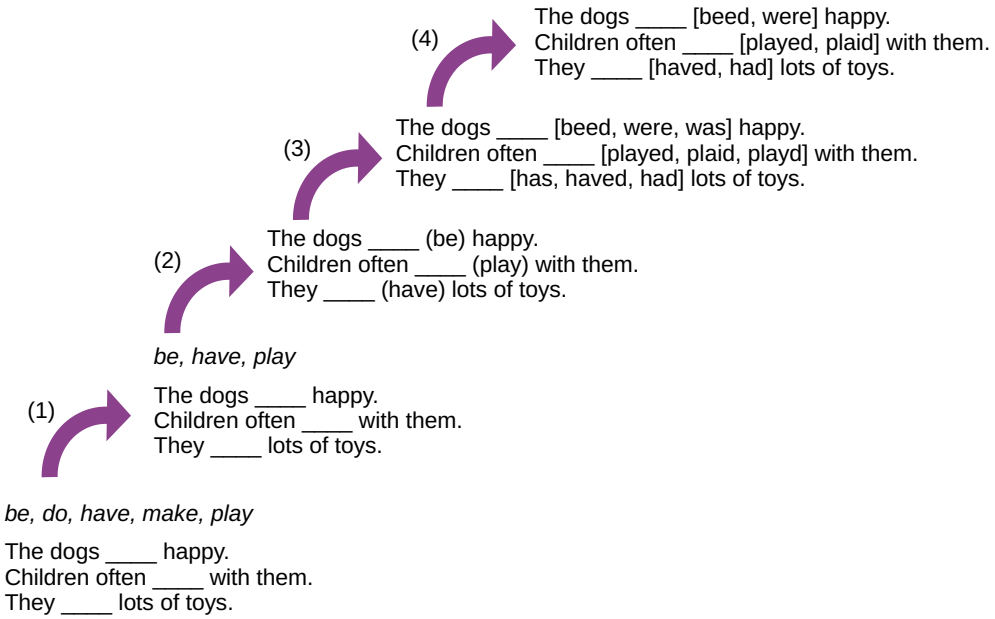


Figure 4.3: Example of meso-adaptive adjustments of exercise elements. Meso-adaptive scaffolds can step by step make Gap-filling exercises easier by making the hints more specific. The steps consist in removing the distractor lemmas from the bag of words of lemma candidates, specifying the required lemma in parentheses behind the corresponding blank, replacing the blank with distractors, and reducing the number of distractors.

by successively reducing the number of distractors ④.

Our approach focuses on this type of adaptivity targeting exercise elements. Although it can be seen as a type of feedback and hint provision, we introduce the term *meso-adaptivity* to distinguish adaptive adjustments to exercise elements from micro-adaptive scaffolds described in the literature. The basic idea builds on the assumption that, by restricting the space of possible answers, any exercise can be made easy enough for any learner to eventually find the correct solution. Meso-adaptivity thus basically transforms open or half-open exercises into closed exercises when necessary.

In order to validate this assumption, we examined the submissions to Gap-filling exercises from the I4S and Didi studies (see Section 3.1.3). They were sub-divided into Fill-in-the-Blanks an Single Choice exercises. Single Choice exercises basically constitute Fill-in-the-Blanks exercises where the blank has been replaced with drop-downs that offer answer candidates and are therefore considered less complex

(Medawela et al., 2018). If our assumption is correct, we would expect that a number of students did not eventually arrive at the correct solution for all Fill-in-the-Blanks exercises. For Single Choice exercises, on the other hand, we would expect almost all exercises to be answered correctly in the end.

For our evaluations, we considered only the final submission of each exercise instead of all attempts. The evaluations show that the percentage of Fill-in-the-Blanks exercises submitted incorrectly was indeed higher (6.06%) than that of Single Choice exercises (5.53%). Interestingly, when considering only those submissions where the learners did not provide the correct answer at the first attempt, the picture is reversed. For Fill-in-the-Blanks exercises, 2.62% of the submissions with at least one incorrect attempt were submitted without providing the correct solution. For Single Choice exercises, this amounts to 5.08%. Considering that 30.66% of SC exercises had only two options, we assume that many learners did not make the effort of selecting the correct option after concluding that the remaining option must be the correct answer. Indeed, when only considering those submissions for Single Choice exercises with at least three drop-down options, only 1.81% were submitted incorrectly. For 37.72% of the incorrect submissions, the learner had tried all incorrect options before submitting the exercise. We thus conclude that learners are more likely to try and find the correct solution with closed exercises than with more open exercises, even if they do not always submit correct answers to closed exercises. We therefore adopt an approach to meso-adaptivity that aims to reduce the space of possible answers in order to gradually transform half-open exercises into closed exercises.

4.2.2 Enriching Exercises with Meso-adaptivity

We present strategies for meso-adaptivity that dynamically reduce the space of possible answers for six exercise types, two of which are half-open and four are closed types. Each exercise generally comprises one to five items, although more items may be included, for example in diagnostic exercises. An item is defined as the smallest practice unit that can be used independently of other such units. This makes it possible to re-assemble items from different exercises into new exercises, and to extract single items for more focused practice of a `linguistic construct`. At the same time, statically assembling multiple items into an exercise allows to

make use of global elements that need to be unambiguous within an exercise, such as lemmas of target answers in Gap-filling exercises for which learners need to identify the correct item themselves. We will elaborate on this example when we look at Gap-filling exercises. Each item, in turn, contains actionable and non-actionable elements. Non-actionable elements are co-text elements for which the learner does not need to provide an answer. Actionable elements are target elements for which learners can obtain one scoring point if they provide the correct answer. The specific definition of an actionable element varies from one exercise type to another. Each exercise should have roughly the same number of actionable elements so that performance aggregation metrics in open learner models can use a constant number of the n most recent actionable elements in order to represent the learner's performance on m exercises. We define n as 10, which should then represent approximately $m = 2$ exercises.

As meso-adaptive adjustments change an exercise, the issue of scoring should always be considered when discussing meso-adaptivity. Scoring of a learner's attempt on an exercise depends on the exercise type. There are generally three possible timepoints for triggering evaluation: (1) when a learner finishes interaction with the element, (2) when a learner submits the entire exercise, or (3) on demand when a learner clicks an evaluation button. Since feedback is most effective when provided immediately after an error has been made (Opitz et al., 2011), our approach implements the first option. In reaction to this feedback, learners may adjust their answer as often as they desire, or until they provide the correct answer. Each adjustment results in an additional attempt. Correct elements cannot be modified anymore so that no further attempts are triggered once the element has been evaluated as correct. When the learner does not wish to adjust the answer to any actionable element anymore, they can submit the exercise. The ILTS then determines a criterion-referenced performance of *correct at first try*, *correct after feedback*, *incorrect*, or *missing* for each element (Colling et al., 2024). *Missing* is assigned if no attempts have been registered when the exercise is submitted. If a correct attempt is registered but there was an incorrect attempt before for that actionable element, the criterial category *correct after feedback* is assigned. If there was no incorrect attempt before for that actionable element, the criterial category *correct at first try* is assigned. If at least one incorrect attempt but no correct attempt has been registered when the exercise is submitted, the element is evaluated as *incorrect*.

In **Mark-the-Words** exercises (see Figure 4.4), each item constitutes a text split into clickable chunks ①. Learners have to find and click on chunks with specific properties ② that are indicated in the instructions. Incorrectly selected chunks can be de-selected by clicking on them again. There are two options to define an actionable element: (a) each chunk or (b) each correct chunk. The first option would violate our requirement that all exercises have roughly the same number of actionable elements. In addition, incorrect chunks are subject to meso-adaptive changes so that the number of clickable chunks is not constant while the learner works on the exercise. Therefore, we define actionable elements as correct chunks.

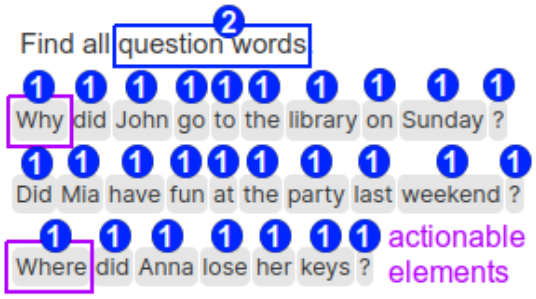
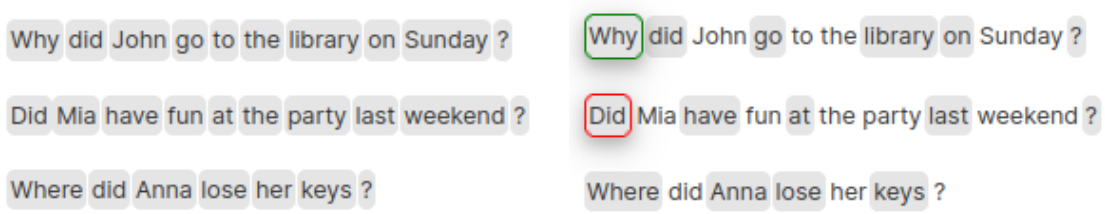


Figure 4.4: Mark-the-Words exercise.
Each clickable chunk ① that has the meta-linguistic properties ② specified in the instructions constitutes an actionable element.

For Mark-the-Words exercises, both actionable and non-actionable elements can be clicked. The number of interactive elements thus deviates from the number of actionable elements. In order to provide a score of how many of the actionable elements were identified at first try, it would therefore be best to attribute an error in terms of an incorrectly clicked chunk to an actionable element. This is, however, challenging since the system cannot possibly determine the reason for which the learner clicked the chunk. Attributing the error to the wrong actionable element could result in incorrect scoring. For example, a presumed third attempt on one actionable element might instead have been the first incorrect attempt on another actionable element for which the system did not record any incorrect attempts at all. This would result in 1 *correct after feedback* and 1 *correct at first try* instead of 2 *correct after feedback*. It therefore makes more sense to only consider as many clicks for *correct at first try* as there are actionable elements in the exercise. Any succeeding correct clicks will then be evaluated as *correct after feedback*. The number of actionable elements limits the chunks that can be selected at the same time. If the

learner tries to select yet another chunk, a popup appears alerting them that they first need to de-select another chunk. This makes it more transparent to learners that the first n clicks (n = number of actionable elements) determine the *correct at first try* evaluations. In addition, it ensures that the sum of correct and incorrect selections at any time is no higher than the number of actionable elements.

In the exercise's initial state, each chunk is clickable (see Figure 4.5a). When a chunk is correctly clicked, it is de-activated so that it cannot be clicked anymore. In order to illustrate scaffolding, Figure 4.5b shows how the number of clickable elements is reduced by half when a learner clicks on an incorrect chunk. Learners have to actively de-select incorrect chunks themselves. This ensures that they recognize that it is not a target chunk. The currently (incorrectly) selected chunk remains clickable until after the learner has de-selected it. Then it becomes inactive instantly. For the remaining chunks, the first or second chunk of each adjoining pair is randomly de-activated. If the learner clicks on another incorrect chunk, half of the remaining active chunks are de-activated. This is repeated until only actionable elements remain activated.



(a) Initial state. (b) After first scaffolding step.

Figure 4.5: Meso-adaptive adjustments in Mark-the-Words exercises. Meso-adaptivity disables half of the currently clickable chunks.

Categorization exercises (see Figure 4.6) comprise textual, draggable elements ① that learners need to put into one of two or more drop areas representing meta-linguistic categories ② via drag and drop interaction based on meta-linguistic properties. Incorrectly dropped elements can be dragged to another category, or else placed back with the unsorted elements via drag and drop or, for all incorrect elements of a category, via a reset-button ③. The two options to define an actionable element are (a) each draggable element or (b) each category. An exercise usually has two categories and five draggable elements, and all draggable elements are independent of each other. We therefore define an actionable element as a draggable

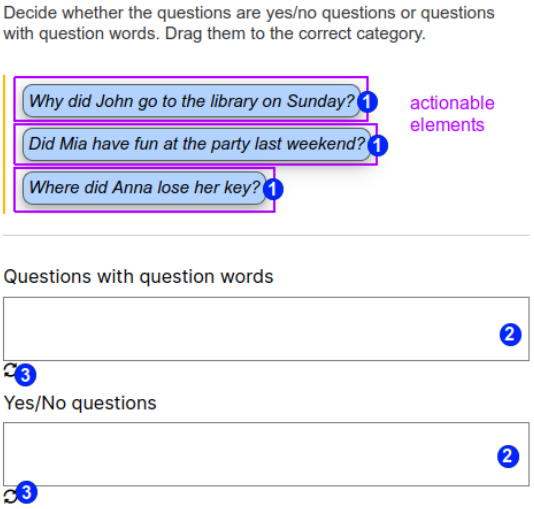


Figure 4.6: Categorization exercise.
Each draggable element ① that has to be sorted into a category ② constitutes an actionable element. The reset-button ③ clears all elements from a category.

element.

Actionable elements are evaluated as correct if they are placed in the correct category.

This exercise type does not offer any meso-adaptivity. Considering that Categorization exercises usually only have two categories, we consider it within reason to assume that learners conclude that the second category is correct if the first one results in negative feedback.

Mapping exercises (see Figure 4.7) consist of pairs of text or image cards ① belonging together. The order of the cards is random and may change with each re-load of the exercise. A card can be selected by clicking on it. No more than two cards can be selected at any time. Learners can de-select a card by clicking on it again. A learner can match two cards by selecting them both. Feedback is provided in the form of color highlighting (red/green). After three seconds, correctly matched cards are put on a stack of solved items. Incorrectly matched cards automatically lose their color highlighting and are de-selected after three seconds. As usual, there are two options to define an actionable element: (a) each card or (b) each pair of cards. The first option would result in a higher number of actionable elements, and a correct match would inevitably result in a score increase by 2. We therefore define

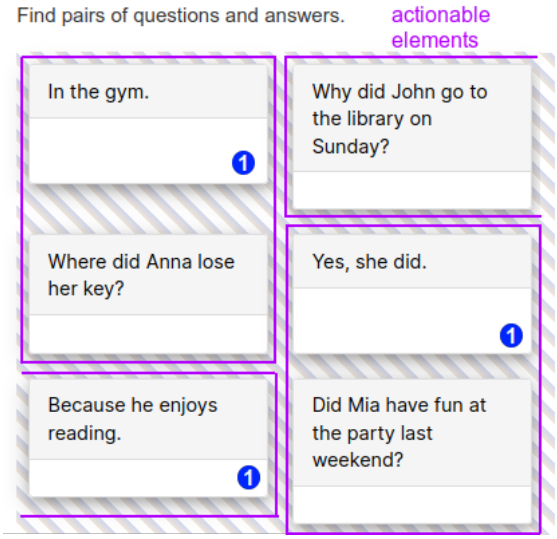
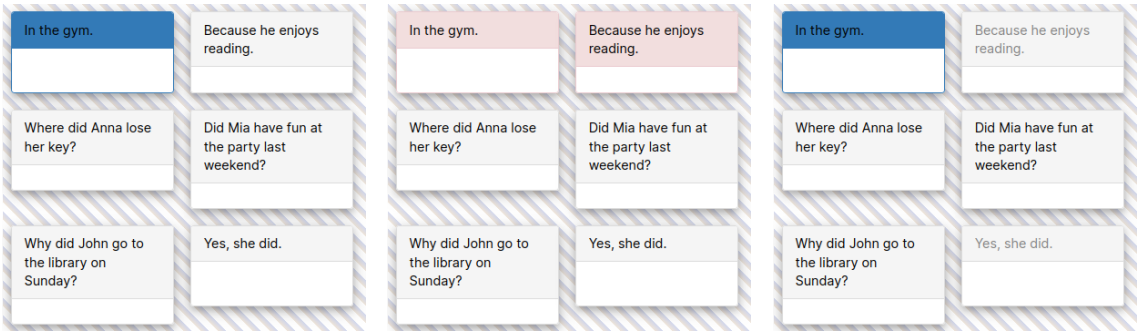


Figure 4.7: Mapping exercise.
Each actionable element is represented as two cards ① that need to be mapped.

an actionable element as a pair of cards.

In Mapping exercises, an incorrect answer consequently involves two actionable elements. While it would be possible to only attribute the error to the actionable element of the card that was clicked first or last, the learner apparently had some misconception about both cards. We therefore evaluate the actionable elements of both involved cards as incorrect. *Correct at first try* is only awarded if both cards of the actionable element were not mapped with any other card before.

For meso-adaptivity, the challenge lies not in the involvement of two actionable elements, but in that only one half of each element is concerned. Therefore, meso-adaptive adjustments do not operate on the entire actionable element but on individual cards. The general idea is that pairs in Mapping exercises tend to consist of cards from two meta-linguistic categories and that the categories are similar across all items of the exercise. For Mapping exercises on conditionals, for instance, cards in the first category might give the if-clause whereas cards in the second category would give the main clause of a conditional sentence. As can be seen in Figure 4.8, scaffolds become active when a learner selects a card again that they have previously selected in combination with another card that is not its correct match. Therefore, when a card is selected again that was previously involved in an incorrect mapping, meso-adaptivity de-activates all cards of the same category as that card so that only



(a) Initial state. (b) Incorrect match. (c) With scaffolding.

Figure 4.8: Meso-adaptive adjustments in Mapping exercises.
Meso-adaptivity disables all cards of the same category as the first clicked card if the first clicked card has already been involved in an incorrect match.

cards of the opposing category remain.

In **Jumbled Sentence** exercises (see Figure 4.9), each item consists of chunks ① that must be put in a drop area ② by clicking on them. If the drop area already contains chunks, the next chunk is placed to the right of the right-most chunk. Incorrectly placed chunks can be removed from the drop area by either clicking on a specific chunk, or by clicking on a reset button ③, which places all incorrect chunks back outside the drop area. There are again two options for defining actionable elements: (a) each chunk or (b) each drop area. The first option faces the same issue as the discarded options for Mark-the-Words and Mapping exercises of resulting in too many actionable elements. We therefore define each drop area as an actionable element.

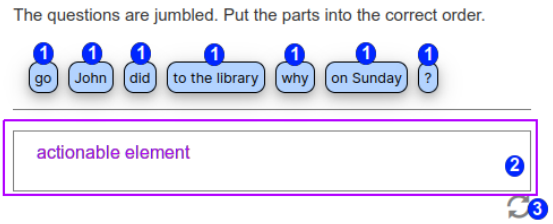
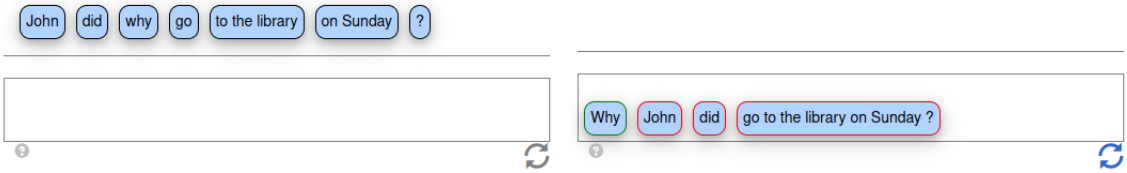


Figure 4.9: Jumbled Sentence exercise.
Each sentence constitutes an actionable element whose chunks ① have to be assembled in correct order in the drop area ②. The reset button ③ clears all incorrectly placed chunks from the drop area.

For Jumbled Sentence exercises, each actionable element thus consists of multiple

chunks. If the chunks of the element are in an order corresponding to that of an accepted target answer, the element is evaluated as correct. If not all chunks have been placed when the exercise is submitted, the element is evaluated as *missing*. Since an actionable element is represented by multiple chunks in this exercise type, binary feedback is slightly different from the other exercise types. Color highlighting indicates up to which chunk the sentence is correct and these chunks cannot be moved anymore. All chunks to the right of the last correct chunk are evaluated as incorrect. Since Jumbled Sentence exercises target word order, only the entire sentence can be evaluated as a unit. Similar to the other exercise types, an attempt is therefore triggered when a learner finishes interacting with the actionable element. This is the case when all chunks have been dropped.

As shown in Figure 4.10, meso-adaptivity concatenates all chunks of a correctly ordered sequence within the chunks evaluated as incorrect.



(a) Initial state. (b) After scaffolding.

Figure 4.10: Meso-adaptive adjustments in Jumbled Sentence exercises. Meso-adaptivity concatenates all incorrect chunks with their adjoining chunks that are in correct position relative to each other.

This may sometimes result in chunking that does not allow any correct solution. For example, the chunk order given in Example 4.1 has the correct target answer given in Example 4.1a. It would, however, result in the concatenated chunks shown in Example 4.1b, which cannot be put into the correct order. Valid chunking would instead constitute the options given in Examples 4.1c and 4.1d.

4.1 Will | freeze to death | if | they | they | don't | wear jackets

a. They will freeze to death if they don't wear jackets

b. * Will freeze to death if they | they don't wear jackets

c. Will freeze to death if they | they | don't wear jackets

d. Will freeze to death | if | they | they don't wear jackets

In order to prevent invalid chunking, the algorithm takes the correct chunk order as starting point and searches for matches in the learner's answer by walking through the correct answer. This way, the same part of the correct answer cannot be contained in two concatenated sequences. Some sentences license multiple correct chunk orders, for instance if the subject and object constitute individual chunks that each represent a person and can thus be interchanged. In this case, the algorithm iterates over all correct answer alternatives and uses the one as target alternative whose string-initial overlap with the learner's answer is longest. If multiple answer alternatives match the beginning of the learner's answer at equal lengths, the first one specified in the exercise definition is taken. Since the items of Jumbled Sentence exercises are disjunct, each item is scaffolded individually.

Gap-filling exercises (see Figure 4.11) combine two commonly used exercise types, Fill-in-the-Blanks and Single Choice exercises. Each item consists of one or more blanks ① embedded in some co-text ②. Learners need to type their answer into the blanks. Although each item typically contains a single blank, this does not necessarily have to be the case. We define an actionable element as a blank. Hints for each blank may be given in parentheses behind the blank, as drop-down selections, or as exercise-global bags of words. These bags of words need to be designed in a way that no contained item constitutes a candidate for multiple actionable elements. This implementation is therefore only valid when the exercise is used in its static assembly of items.

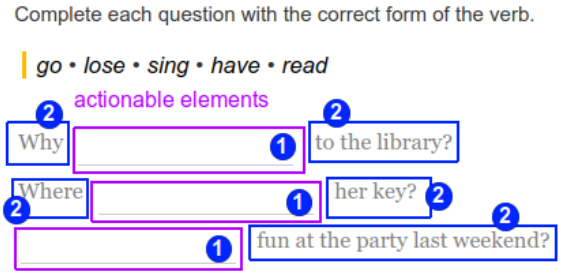
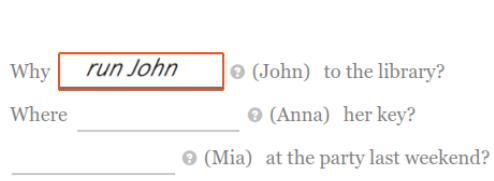


Figure 4.11: Gap-filling exercise.
Each blank ① embedded in some co-text ② represents an actionable element. The lemmas of the target answers are here given in an exercise-global bag of words, which represents the first scaffolding step.

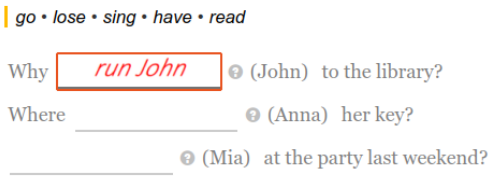
Each actionable element has one or more accepted target answers. If the learner's answer matches one of the accepted target answers, the element is evaluated as

correct. If the learner does not correct a previous incorrect attempt, the element is scored as incorrect.

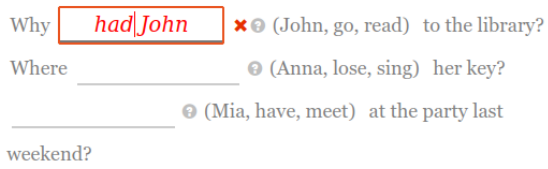
This exercise type supports micro-adaptivity in the form of meta-linguistic feedback. Therefore, if the learner gets the answer wrong, the system always tries to provide this kind of feedback. Only if no specific meta-linguistic feedback can be provided is meso-adaptivity applied as a fallback strategy. Scaffolding is applied to each item individually for this exercise type as well. If an exercise parameter required for a scaffolding step, such as distractors, is missing, it is skipped. The initial state in principal provides only blanks into which to insert a word. This shifts the focus of Gap-filling exercises from being purely on form to also encompassing meaning. The first meso-adaptive adjustment then gives a bag of words at the top of the exercise. This bag of words contains the lemmas of the target answers of all actionable elements, as in Figure 4.12b, if the exercise definition specifies any. Lemmas are typically provided in more form-focused exercises that practice formation of an inflected form. Examples include exercises where learners need to provide the verb in the correct form. For more meaning-focused exercises, on the other hand, the target answer is often identical to the lemma. Examples include exercises asking learners to provide the correct question word. In this case, the bag of words contains the target answers of all actionable elements, plus an additional two semantic distractors. The bag of words remains at the top of the exercise until the exercise is either entirely correct, being submitted, or higher meso-adaptive steps have been applied to all items. The next step, shown in Figure 4.12c consists in putting the lemma and a semantic distractor into parentheses behind the blank that was answered incorrectly. The subsequent scaffolding step, illustrated in Figure 4.12d, removes the distractor lemma. These two steps only make sense for form-focused exercises where the target answer is not identical to its lemma. Figure 4.12e shows the last scaffolding step, which transforms the Fill-in-the-Blanks exercise into a Single Choice exercise by replacing the blank with a drop-down menu that gives the correct form along with a limited number of distractors. The number of distractors can vary from exercise to exercise and is up to the exercise creator. It is, however, also limited by the adaptivity algorithm to a maximum of five. If more distractors are available in the exercise, this enables the system to adapt the displayed distractors to the learner's abilities. In this case, the algorithm selects those distractors that correspond to the most frequent misconceptions in the learner's bug model. To make this possible, the



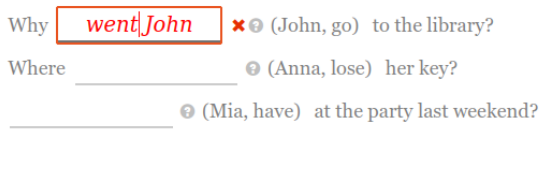
(a) Initial state.



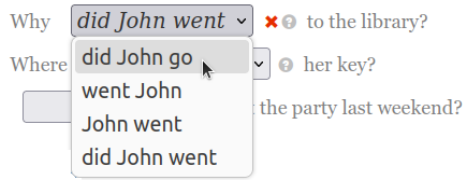
(b) Lemmas as bag of words.



(c) Lemma + distractor in parentheses.



(d) Lemma in parentheses.



(e) Drop-down.

Figure 4.12: Meso-adaptive adjustments in Gap-Filling exercises. Meso-adaptivity specifies the lemmas as a bag of words, next puts the correct lemma along with a distractor lemma behind the blank, then removes the distractor lemma, and finally replaces the blank with a drop-down selection of the correct answer and distractors when a learner types an incorrect answer into the blank.

distractors must be annotated with misconceptions in the exercise definition.

In **Short Answer** exercises (see Figure 4.13), each item comprises a prompt ① eliciting a complete sentence. Learners have to type the answer into an input field ② following the prompt. We define each input field as an actionable element.

Evaluation is similar to that of Gap-filling exercises. The answer is only correct if it matches one of the target answer alternatives.

This exercise type also supports meta-linguistic feedback. However, since the answers are more open than those of Gap-filling exercises, it here makes sense to combine meta-linguistic feedback with additional meso-adaptivity. Meta-linguistic

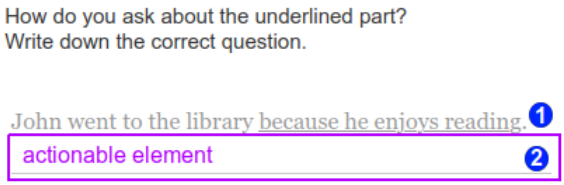
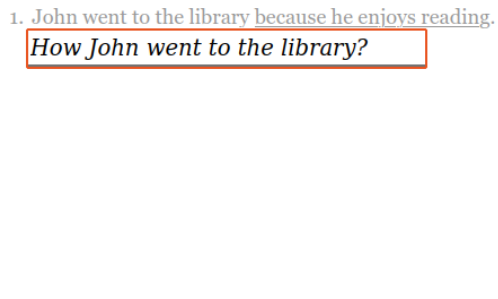
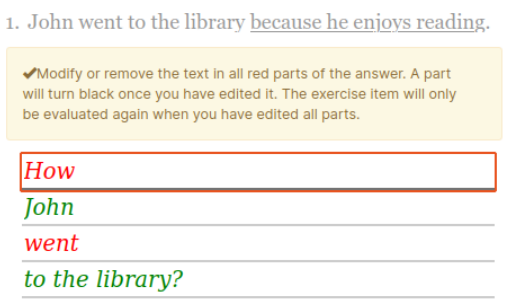


Figure 4.13: Short Answer exercise.
Each answer ② to a prompt ① constitutes an actionable element.

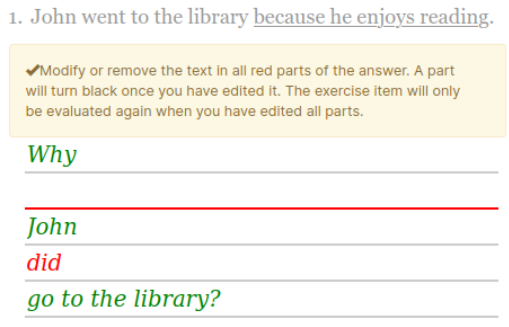
feedback is therefore always shown when available in addition to meso-adaptive adjustments. For these scaffolds, we applied the approach by Yin et al. (2012) with minor adjustments. Focusing on free text production, they convert a learner's input text into a Gap-filling exercise where incorrect parts of the learner's answer are transformed into blanks and the remaining text into non-interactive co-text. Similar to binary feedback in our Jumbled Sentence exercises, the scaffolds thereby indicate the specific parts of the learner's answer that are incorrect and encourage them to



(a) Initial state.



(b) After first scaffolding step.



(c) After second scaffolding step.

Figure 4.14: Meso-adaptive adjustments in Short Answer exercises.
Meso-adaptivity highlights all incorrect parts of the answer and de-activates correct parts. This procedure is identical for all scaffolding steps.

edit these incorrect parts. In our adaptation of this approach, the answer is divided into correct and incorrect parts as illustrated in Figure 4.14b. Part boundaries are always set at the start or the end of a token delimited from other tokens by a blank or the start or end of the answer. All correct parts are highlighted in green and deactivated. Incorrect parts are highlighted in red and remain active. Figure 4.14c shows how, for missing parts, an empty red line is inserted. If multiple possible target answers are given, the one with the shortest edit distance to the learner's answer is applied. This implies that after each re-evaluation, the applied target answer can potentially change if the learner's modifications to their answer suggest that another alternative constitutes a better target hypothesis of the learner's intended production. If the learner edits an incorrect part, that part loses its color highlighting. Only when all incorrect parts have been modified is the exercise evaluated again. This evaluation behavior is made explicit to the learner in an alert that only disappears once all incorrect parts have been edited.

Our meso-adaptive adjustments do not make use of sophisticated NLP methods, but rely on simple string operations. This has the advantage that they can be implemented as client-side operations, which ensures timely reactions to learner interactions and robustness of meso-adaptivity independent of available internet connection.

Since some of the meso-adaptive adjustments are not obvious as helping steps to all learners, such as the chunks of Mark-the-Words exercises being deactivated, the popup shown in Figure 4.15 is displayed after an answer has been evaluated as incorrect, alerting the learner of the adjustments.

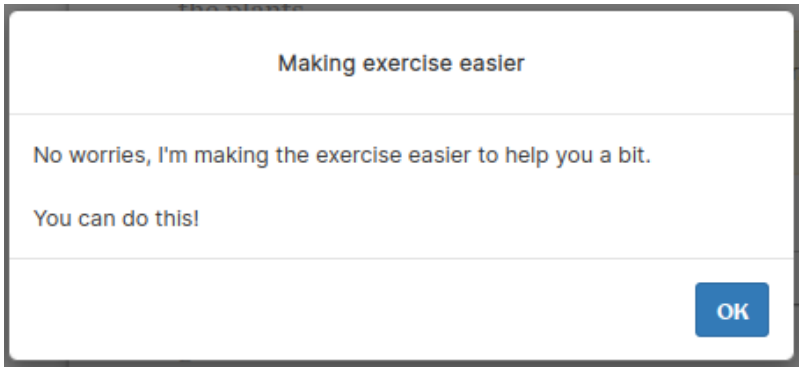


Figure 4.15: Popup to inform about meso-adaptive adjustments.
Learners are informed about the adjustments and their meso-adaptive intention in a popup.

In order to support test scenarios in instructional settings or field studies, such as class tests or study pre- and post-tests, scaffolding can be de-activated. The concrete presentation of the exercise can then be defined by setting the desired meso-adaptive step in the exercise definition. The fixed meso-adaptive step constitutes a static exercise parameter that must be provided when the exercise is created.

4.2.3 Effect of Meso-adaptivity

As learning is most effective when cognitive effort is high (Shea, 2002), our approach presents each exercise initially at its most difficult. Scaffolding is provided only if necessary. We hypothesized that higher exercise complexity might then even result in better uptake and longer-term retention (Robinson, 2021).

In order to evaluate the suggested approach and thus answer RQ 8, we implemented the macro- and meso-adaptive strategies described in Sections 4.1 and 4.2, respectively. The implementation of determining mastery of KCs corresponds to that described in Section 2.2.1. The implementation was used and put to test in the AI2Teach study.

In order to test whether meso-adaptivity can successfully replace a stronger focus on exercise difficulty in macro-adaptive exercise selection, we introduced some modifications to the macro-adaptivity algorithm that only became effective for the control group. This group did not receive meso-adaptive adjustments. Instead, in addition to linguistic variability and co-text complexity, the exercise ranking also considered exercise type complexity for these learners, which emerged as the most influential exercise parameter with respect to exercise difficulty in Section 3.1.1. The type score was determined for each exercise type and `learning target` based on the exercise's type complexity level as specified by a teacher involved in exercise creation. Although there are slight differences across `learning targets`, the type complexity level increases from Mark-the-Words over Mapping, Categorization, Jumbled Sentence, and Gap-filling to Short Answer exercises. The optimal type complexity level was initially set to *1* for all learners in all `realizations`. If the learner answered the last five elements (most of the time one exercise) practicing the same `property` incorrectly, the optimal type complexity remained unchanged. If the learner answered the last five elements correctly, the optimal type complexity level was incremented by *1*. If the learner answered not only the last 5 but the last 10 elements correctly,

the optimal type complexity level was incremented by *2*. Exercises were then scored according to the deviation of their type complexity level from the optimal complexity level. Exercises with higher type complexity than the optimal complexity were thereby penalized twice as heavily as exercises with lower type complexity than the optimal complexity level. Similar to the scores for linguistic variability and co-text complexity, the resulting type score was also normalized over all exercises into the range [0, 1]. As indicated by Formula 4.3, when calculating the overall exercise score s, the type score s_t was accorded double the weight of linguistic variability s_v and four times the weight of the co-text complexity score s_c.

$$s = 2s_t + 0.5s_c + s_v \tag{4.3}$$

The implementation thus supported two study conditions: For the **intervention group**, the macro-adaptive algorithm did not take exercise difficulty into account other than half-open vs. closed (interactive) exercises. Learners needed to first master a construct in closed exercises before receiving half-open exercises for that construct. The remaining exercise candidates were ranked according to their dynamically determined linguistic variability score. Scaffolding was provided in the form of meta-linguistic feedback, if available, and meso-adaptive adjustments. For the **control group**, the macro-adaptive algorithm considered the specific exercise type in the final ranking. While the algorithm did not strictly enforce a progression from easier to more difficult exercises, both linguistic variability and exercise type difficulty were considered for exercise ranking, with double weights for type difficulty. Scaffolding comprised only meta-linguistic feedback. If the exercise type in principle supported meta-linguistic feedback but the system could not diagnose the exact nature of the error, default meta-linguistic feedback ('This is not what I was expecting') was displayed.

All analyses are based on the dataset described in Section 2.2.1. Since the study encompassed within-class assignment to two different study conditions, one concerning the Fitness Center and one concerning meso-adaptivity, we applied a 2x2 factorial design. 567 Gymnasium students were assigned to the intervention condition, 558 Gymnasium students and all 66 Gemeinschaftsschule students to the control condition.

The evaluation encompasses several analyses ranging from simple Pearson's correlations and two-tailed t-tests to mixed effects analyses. They consider pre- and post-test data, in particular the improvement between the two assessments, as well as practice exercises, diagnostic exercises, and small polls issued after each exercise. When a learner submitted an exercise with an incorrect or missing answer, the poll inquired after the reason for not attempting to find the correct answer. When a learner submitted a fully correct exercise, the poll asked for a subjective assessment of the exercise's difficulty on a 5-point scale.

Note

The implementation of the adaptivity algorithm could not be tested extensively before being released for the study. Some bugs were therefore discovered during the running study. We were able to resolve them, but this intervention into the running study might have introduced some noise in the collected data. Despite our best efforts, not all such noise could be removed from the data. Other bugs were intentionally not corrected during the study as they affected both study conditions and fixing them would have introduced further noise in the evaluation data.

We first look at the effects of the reduced focus on exercise difficulty in macro-adaptivity with the aim of determining whether the selected exercises were still appropriate for the learners. Since macro-adaptive exercise selection considered exercise difficulty to a higher degree in the control condition, we evaluated the learners' performance on practice exercises for a more objective analysis. Performance was here defined as the percentage of attempted actionable elements that were answered correctly at the first attempt. As discussed in Section 2.2.2, we defined appropriate accuracy as ranging between 60 and 80% based on mastery thresholds typically applied in Mastery Learning. Mean values of practice accuracy within a `learning target` ranged from 83.04% (Conditionals) to 91.70% (Simple past) for the intervention group, and from 78.89% (Conditionals) to 91.55% (Simple past) for the control group. A full overview is provided in Table 4.1. According to a two-tailed t-test, the control condition had significantly higher pre-test scores when considering all test exercises (t=2.4938, p=.0128; g=.1523) as well as for the `learning target` *Conditionals* (t=2.5097, p=.0122; g=.1537), so that previous knowledge must be taken into account in the evaluations. In a mixed effects analysis controlling for previous

`Learning target`	**Accuracy C [%]**				**Accuracy I [%]**				**C vs. I**	
	μ	σ	min	max	μ	σ	min	max	t	p
Simple past	91.55	11.53	28.42	100	91.70	11.08	33.33	100	-0.216	.829
Present perfect	89.59	13.08	0.00	100	89.48	13.33	0.00	100	-.079	.937
Simple past vs. present perfect	85.03	19.40	0.00	100	84.70	19.02	0.00	100	-.069	.945
Conditionals	78.89	23.51	0.00	100	83.04	18.55	3.64	100	-2.588	.010
Questions	86.56	14.92	15.00	100	84.29	16.65	19.29	100	.821	.412

Table 4.1: Practice performance per `learning target`.
Negative t-values indicate higher performance in the intervention group according to a mixed-effects analysis controlling for previous knowledge. Accuracy was higher for the control group (C) in some `learning targets`, and for the intervention group (I) in others. None of the differences are significant.

knowledge in terms of pre-test scores, we did not find any significant difference between the two study conditions in getting the practice exercises correct at the first attempt.

Although mean values are overall within or above the aimed for accuracy of 60 to 80% for both study conditions and in all `learning targets`, there is considerable variance across exercise types. Closed types have very high accuracies with means between 88.79% (σ = 16.24) for Mark-the-Words exercises and 100% (σ = .00) for Mapping exercises across the control condition[1]. Half-open exercises, on the other hand, had rather low accuracy scores with considerable standard deviations (μ = 60.48%, σ = 16.04 for Gap-filling exercises; μ = 43.43%, σ = 34.99 for Short Answer exercises). The values are comparable for the intervention condition. Closed exercises thus do not need to be practiced excessively as learners are not challenged by them. Since meso-adaptivity can only make exercises easier but cannot increase exercise difficulty, it is important to macro-adaptively select exercises that are below the learner's current abilities. In our system, the prioritization of mandatory exercises frequently required learners to complete closed exercises even after proving their mastery of the required `skill` before selecting half-open exercises. However, since learners can always explicitly click on a specific mandatory exercise in the user

[1]We did not consider Categorization exercises since scoring was erroneous for this exercise type during the study.

interface to practice it, sequencing them all in is actually not strictly necessary. In addition, the dashboard highlights already submitted mandatory exercises so that learners see easily enough which mandatory exercises are still missing. We therefore suggest to reject the idea of this prioritization of mandatory exercises.
Contrary to closed exercise types, most learners were not ready to instantly answer half-open exercises by themselves when the macro-adaptive algorithm selected them. The control group, for which in particular the transition from closed to half-open exercises was made smoother through Single Choice exercises, did not perform better on half-open exercises. The intervention group even outperformed the control group on Gap-filling exercises, although performance was generally lower in this study condition. This suggests that considering the exercise type to a greater degree did not impact the learners' performance on practice exercises and that the more controlled progression from easier to more difficult exercise types in the control condition did not result in selecting easier exercises. Therefore, other means of modulating exercise difficulty, for instance meso-adaptivity, need to be incorporated for half-open exercises.

However, this interpretation is based on the assumption that learning gains from each exercise were similar in both conditions, so that the learner's probability of answering an exercise correctly is not conditioned by the availability of meso-adaptivity or a stronger macro-adaptive focus on exercise difficulty in the previously submitted exercise. If, however, practice in the meso-adaptive condition was more efficient, subsequent exercises of a similar complexity would have been easier for the intervention group. In order to test for this interaction effect, we analyzed the number of submitted exercises before attempting a diagnostic exercise as a measure of efficiency of practice, and the learners' performance on the diagnostic exercise as a measure of effectiveness of practice. Mixed effects analyses controlling for previous knowledge show that while there is no significant difference in the number of submitted practice exercises (t=.837, p=.403; g=.0194), the intervention condition performed worse on the diagnostic exercises. The effect was only significant for the `learning target` *Questions* (t=3.300, p=.001; g=.3428). The box plot in Figure 4.16 suggests a ceiling effect, yet the range of accuracy scores was wider for the intervention group, and scores were generally lower. The number of not attempted actionable elements, for which no learner interaction was registered at all in the system, was also similar for both conditions (t=.2012, p=.8406; g=.0120). Although the number

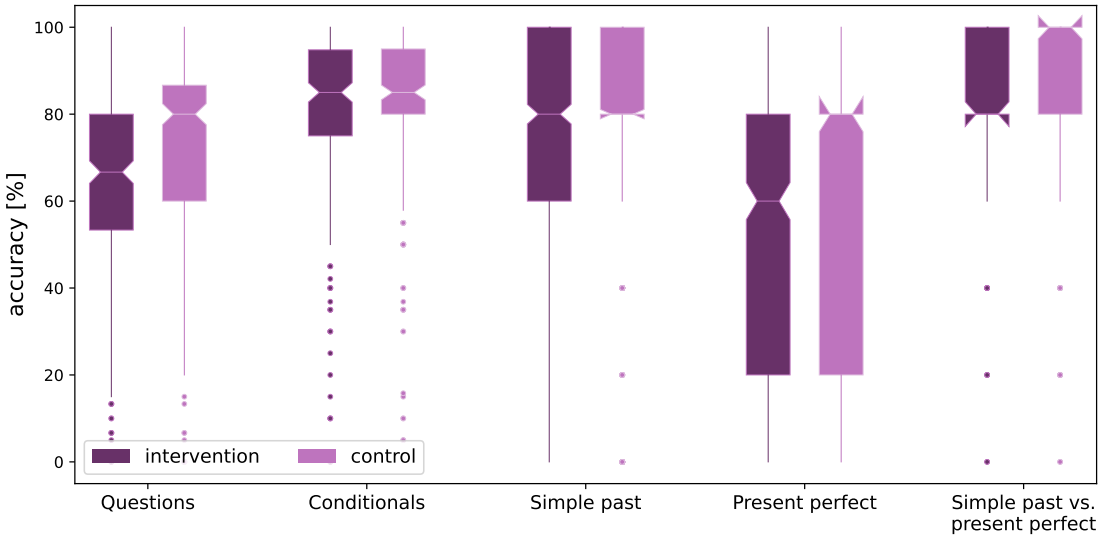


Figure 4.16: Accuracy on diagnostic exercises.
Learners in the control condition had higher accuracy on diagnostic exercises than learners in the intervention condition. There seems to be a ceiling effect for most `learning targets`.

of practice opportunities was thus similar, short-term learning gains were lower in the intervention group.

In addition to being less effective, practice was thus also less efficient when meso-adaptive adjustments were provided. The lower practice efficiency in the intervention condition as compared to the control condition is reflected in the number of practice exercises needed to master a KC. This kind of mastery of a KC is not assessed by a diagnostic exercise, but determined by the adaptivity algorithm. The difference between the study conditions with respect to the number of exercises required to reach this kind of mastery is significant for only three `realizations` (*Mixed questions*: t=2.485, p=.015, g=.5657; *Questions with question words*: t=2.285, p=.025, g=.5668; *Positive statements in the simple past*: t=2.304, p=.036, g=.5588), but for these, the intervention condition needed more practice exercises. Not having to go through the entire sequence of macro-adaptive exercise types thus does not reduce the number of exercises required to reach mastery.

Zooming into the exercises, we can also look at efficiency in terms of the number of attempts on each exercise instead of as the number of practiced exercises. As shown in Figure 4.17, the control condition had more incorrect attempts for all

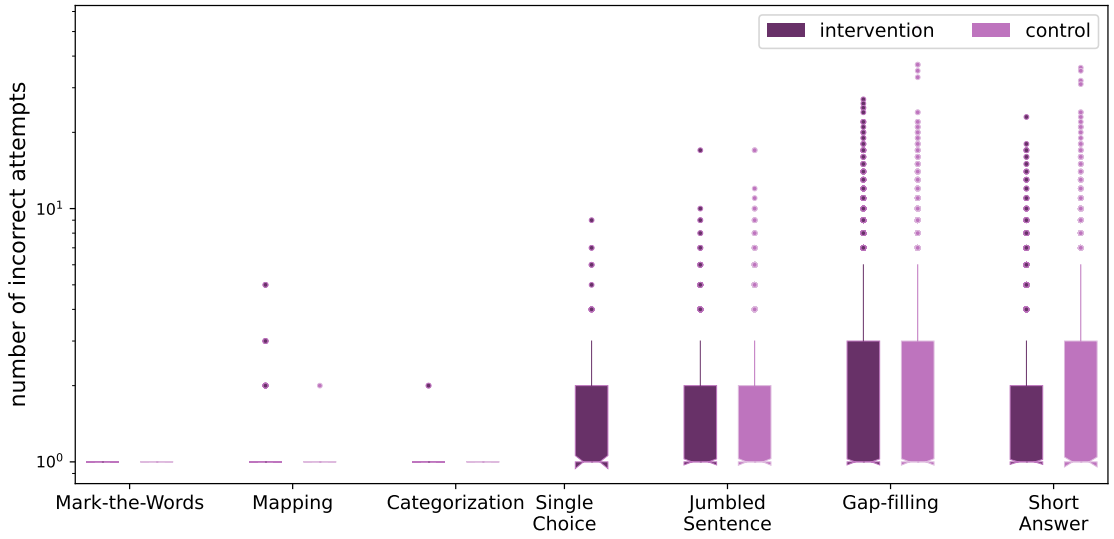


Figure 4.17: Incorrect attempts on actionable elements.
The intervention condition had less incorrect attempts for all exercise types, although the effect is only significant for Jumbled Sentence, Gap-filling, and Short Answer exercises. The number of incorrect attempts is overall very low for the remaining exercise types.

`learning targets` and exercise types. The effect is significant for Jumbled Sentence (t=6.459, p=.000; g=.1944), Gap-filling (t=3.590, p=.000; g=.0449), and Short Answer (t=10.569, p=.000; g=.2266) exercises. Both groups had very few incorrect attempts on closed exercise types, yet the difference is rather pronounced on half-open exercises. Meso-adaptivity thus generally helped learners to find the correct answer faster.

This is reflected in the poll of subjective exercise difficulty, issued after each exercise in the form of a 5-point scale. It showed significant differences (t=4.9339, p=.0000; g=.1607) in the perceived difficulty of the exercises between the study conditions. The intervention condition, who received meso-adaptive adjustments, rated the exercises easier than the control condition, who received less support. The effect is not significant for individual closed exercise types, but it is significant for the two half-open exercise types.

Mastery was reached very rarely, but in those cases where the entire `realization` was mastered, the intervention condition needed more exercises to achieve it. The low mastery values can be explained by the system design that allowed learners to attempt a diagnostic exercise regardless of their mastery of the `realization` as

estimated by the adaptivity algorithm. Since passing the diagnostic exercises was the main goal for most learners[2], they apparently did not see the need to follow the adaptivity module's recommendations to continue practice. The pie chart in Figure 4.18 shows that most learners passed the diagnostic exercises irrespective of mastery of the corresponding `realization`. Failure on a diagnostic exercise correlated with the system's recommendation to first practice more in only 4.99% of the submitted diagnostic exercises. In these cases, however, the learners might have benefited from following the system's recommendation. As outlined in Section 2.2.1, learners overall performed better on diagnostic exercises if they had mastered the `realization` according to the algorithm. In those cases where the learners had proven mastery, they generally reached the accuracy threshold of 60-80% that is

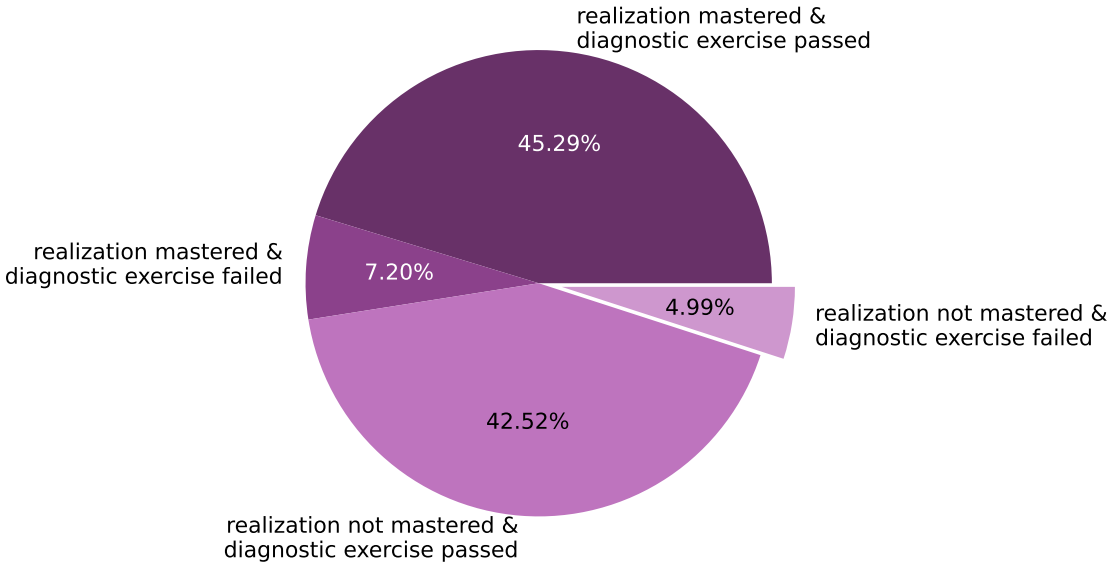


Figure 4.18: Performance on diagnostic exercises by mastery.
Most learners passed the diagnostic exercises regardless of whether the system estimated them to be ready or not. Almost 5% of the failed diagnostic exercise submissions correlated with a lack of mastery according to the adaptive system.

[2]Teachers participating in the AI2Teach project reported that most of their students completed mainly the mandatory and diagnostic exercises in order to receive the trophy symbol in their dashboard that indicates passing of the diagnostic exercise and thus readiness for the functional target task.

required to pass a diagnostic exercise[3]. In view of these findings it is little surprising that the intervention group achieved lower accuracy scores on diagnostic exercises than the control group.

As an inevitable result of practicing the same number of exercises, but requiring more exercises to reach mastery, the intervention condition reached mastery significantly less often (t=2.471, p=.014; g=.1497) than the control condition when controlling for previous knowledge. Figure 4.19 highlights that this is the case for almost all `realizations`.

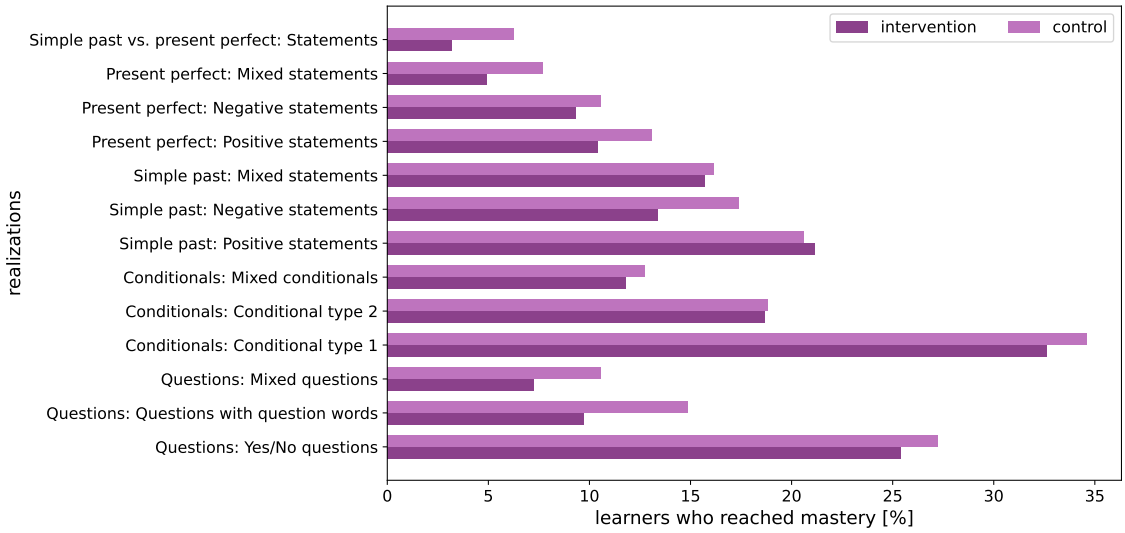


Figure 4.19: Percentage of learners having reached mastery.
In almost all `realizations of learning targets`, more learners of the control condition reached mastery according to the adaptivity algorithm than did learners of the intervention condition.

Pearson's correlation between the number of mastered `properties` and improvement in test accuracy between pre- and post-test was, however, very low, although slightly higher for the intervention group ($\rho = .2076$) than for the control group ($\rho = .0559$).

An interesting observation points out that although learners in both conditions practiced similar numbers of exercises before attempting a diagnostic exercise and showed similar practice performance, the control condition performed better on the diagnostic exercises. As can be seen in Figure 4.20, the number of practiced exercises

[3]The actual passing threshold was set to 80%. However, learners with 60% accuracy were allowed to revise their answers with the hint that there were some errors (although without indicating the location of these errors) and could re-submit the exercise.

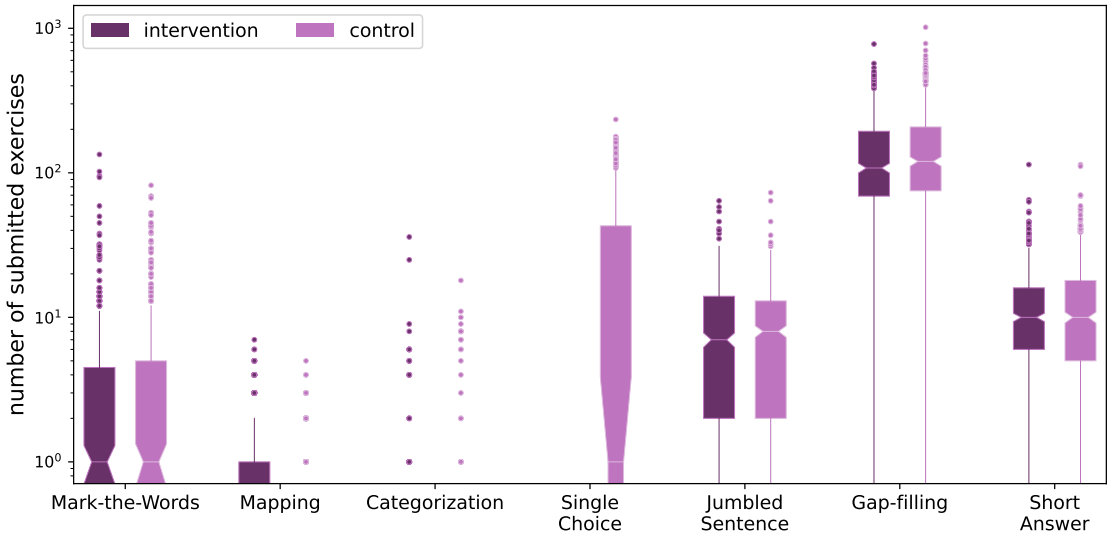


Figure 4.20: Numbers of practiced exercises.
The number of exercises practiced per exercise type is similar for both study conditions across all exercise types.

per exercise type was similar for the groups as well. Table 4.2 outlines that only for Mapping (t=6.1577, p=.0000; g=.3669) and Short Answer (t=-2.3103, p=.0211; g=.1377) exercises, there was a significant difference in the number of practiced exercises between the conditions. Submission numbers for Mapping exercises were, however, so low ($\mu_C = .3315, \sigma_C = .6854; \mu_I = .6984, \sigma_I = 1.2322$) that one cannot attribute any significance to the observed difference. In terms of Short Answer exercises, learners in the control group submitted more exercises. For the remain-

Exercise type	**n exercises C**		**n exercises I**		**C vs. I**	
	μ	σ	μ	σ	t	p
Gap-filling	160.2957	125.4910	147.2646	108.2912	1.8656	.0624
Fill-in-the-Blanks	131.7599	97.2961	.0	.0	32.2465	.0000
Single Choice	28.5358	46.2162	.0	.0	14.7025	.0000
Categorization	.2276	1.3964	.2628	2.0297	-.3382	.7352
Mapping	.3315	.6854	.6984	1.2322	-6.1577	.0000
Mark-the-Words	4.7903	9.3063	4.3898	10.9811	.6595	.5097
Short Answer	13.6344	12.5670	12.0511	10.3271	2.3103	.0211
Jumbled Sentence	8.8620	8.1866	8.8148	8.6733	.0938	.9253

Table 4.2: Number of practiced exercises per exercise type.
The control (C) and intervention (I) groups practiced a similar number of exercises per exercise type.

ing exercise types, learners in both study conditions practiced similar numbers of exercises. The better performance of the control group might therefore indeed be related to having practiced more exercises of a more beneficial type in terms of learning gains. In addition, Short Answer exercises were usually the last exercise type practiced before attempting a diagnostic exercise. If we assume Short Answer exercises to be particularly beneficial for learning gains, this effect obtained in the last practice exercise would not be reflected in any further practice exercises, but only in the directly succeeding diagnostic exercise. Supporting this, diagnostic exercises are encoded as half-open exercise types as well, so that the last practice exercise was most similar to the diagnostic exercise. Since test performance is highest when no transfer is required to different skills that were not targeted in practice (DeKeyser, 1997), this would explain the high relevance of the last practice exercise in terms of performance on the diagnostic exercise. Another possible explanation emerges when not controlling for previous knowledge. A simple two-tailed t-test attributes better practice performance to the control group, although the difference is only significant for the `learning target` *Conditionals* (t=2.2157, p=.0271). Since both groups practiced similar numbers of exercises before attempting a diagnostic exercise, the intervention group might thus not have been as well prepared.

The analyses so far revealed that short-term learning gains were lower in our study when meso-adaptivity was provided. This is in line with findings from related work that too much scaffolding decreases learning gains (Vanderhyde, 2019; Richey and Nokes-Malach, 2013). In order to shed further light on the usefulness of meso-adaptivity, we looked at long-term learning gains that we determined based on the learners' improvement between pre- and post-test. The improvement data reflects the findings from the analysis of diagnostic exercises: The control condition had significantly higher learning gains according to a mixed effects analysis controlling for previous knowledge (t=1.857, p=.064; g=.1128). On individual `learning targets`, however, the effect is only significant for *Conditionals* (t=1.945, p=.052; g=.1069). Interestingly, this `learning target` is the only one where the control group had significantly different (i.e., higher; t=2.5097, p=.0122; g=.1537) pre-test scores representing previous knowledge than the intervention group according to a two-tailed t-test. Considering our earlier findings of higher efficiency in the intervention condition, we surmise that increased time of engagement with the exercise is beneficial for learning outcomes. This is in line with Whitmer et al.'s (2022) findings on efficiency

in terms of the number of trials required to correctly answer all flashcards of the training session, and long-term retention of learning in the military sector. In their study, more efficient learning correlated with decreased long-term retention.

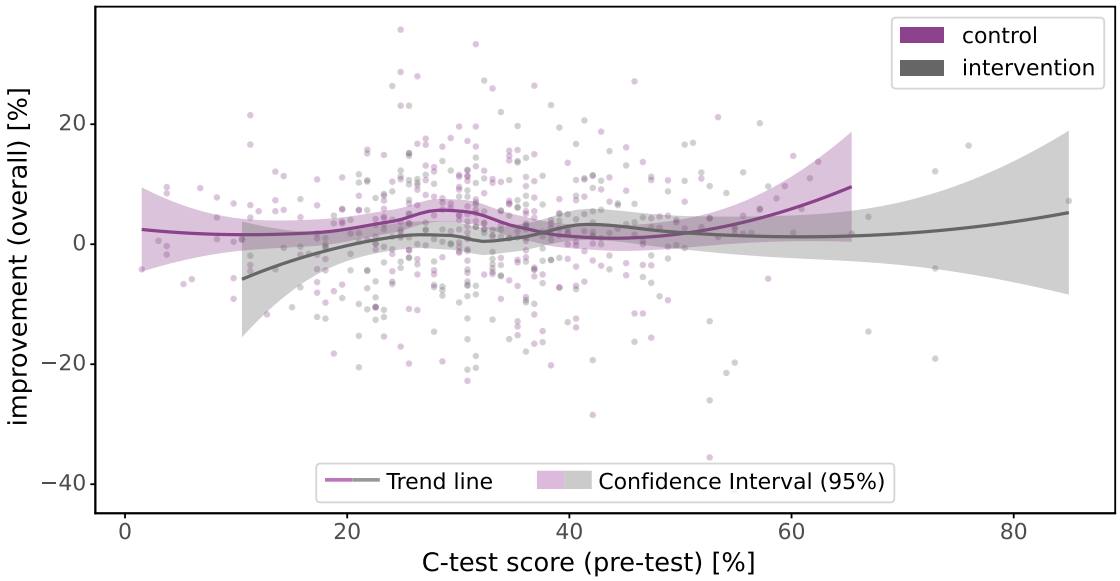
Before rejecting the usefulness of meso-adaptivity altogether, we should therefore consider the possibility that the adjustments were applied too early. If this is the case, it might have caused frustration in learners as they were not allowed another chance at solving the exercise at the higher difficulty level. Such an effect could be mediated by asking learners whether they want meso-adaptive adjustments to be applied after each incorrect answer. An additional issue related to motivation concerns a potential lack of effort when knowing that the exercise will become easier after providing an incorrect answer. Some learners possibly did not make an honest effort to solve the exercise in its most complex state, but instead waited for meso-adaptivity to reduce its complexity. Mastery of a KC was calculated based on the correctness of the learners' first attempts on the exercises so that they could only progress towards mastery if they answered exercises correctly at the first attempt. This was, however, not made transparent to the learners so that they might not have seen the relevance of correctly answering the exercises on the first attempt. In addition, since diagnostic exercises could be attempted irrespective of the learners' mastery of the `learning targets` and their `realizations`, motivation to reach mastery might have been low. Meso-adaptivity might have further decreased learners' motivation to get the exercises correct at the first attempt as it guided them towards the correct solution – for instance in the form of a drop down menu including the correct answer – with little effort from their side. In order to test this hypothesis, we extracted nonsense answers from learners' first submissions to practice exercises. To this purpose, we identified a number of tokens that we often observed in valid learner answers and labeled all answers containing any of these tokens as valid answers. We then automatically labeled all remaining learner answers with less than two characters as nonsense answers, as well as answers where the edit distance to the target answer was greater than the length of the target answer. The resulting list of nonsense answers indicates that, contrary to our hypothesis, learners in the intervention group provided significantly less (t=-2.0469, p=.0418; g=.2668) nonsense answers than the control group according to a two-tailed t-test. In the intervention condition, nonsense answers made up 5.10% of the submitted answers, whereas they amounted to 7.62% in the control condition. If motivational

shortcomings are responsible for the negative effect of meso-adaptivity despite these contradicting findings, they could be addressed by integrating an assessment module to determine when a learner should receive help. Aleven et al. (2006) present such an assessment module for their Cognitive Tutor *Help Tutor* that aims to improve learners' help-seeking behavior.

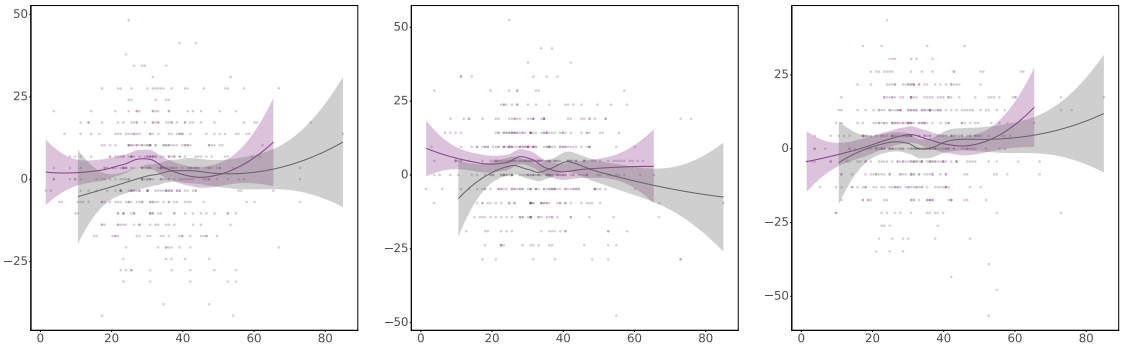
Another hypothesis to explain the negative effect of meso-adaptivity assumes that the adjustments deflected the learners' attention from the exercise content to the changed elements. Such a distraction might lessen as the novelty effect wears off (Wetzel et al., 2009). Further assessments after prolonged use of the ILTS are required to test this hypothesis.

In order to determine whether meso-adaptivity is beneficial for sub-groups of learners, we examined differences between the study conditions when considering a range of learner parameters. They comprise the learner's English grade, scores on the grammar test as well as on the C-test of the pre-test, the number of submitted practice exercises, accuracy on all practice exercises as well as on the subsets of individual exercise types and the learning units. We identified loess trend lines modeling the relationship between each learner parameter and improvement on the practiced `learning targets` between pre- and post-test. We determined trend lines individually for the two study conditions.

The most revealing insights could be found for the learner parameter of C-test scores in the pre-test. Figure 4.21 shows that improvement decreased with intermediate C-test scores for the control condition, but increased for the intervention condition in this range. The effect is similar for all `learning targets` that were practiced in the system, although the range of C-test scores within which the intervention was beneficial is slightly higher for Tenses than for the other `learning targets` (see Figures 4.21b–4.21d). The turning point seems to be at 35%, and the effect is again reversed for C-test scores above 55%. A significance test shows that between 35% and 55% accuracy in the C-test, learners in the intervention condition indeed performed better than the control group, although not significantly (t=1.3507, p=.1783; g=.1885). The control group performed significantly better than the intervention group below (t=2.8157, p=.0052; g=.3218), and not significantly better, although with a medium effect size, above this threshold (t=1.4150, p=.1694; g=.5374). Meso-adaptivity thus generally resulted in greater long-term learning gains for learners



(a) Overall.



(b) Conditionals. (c) Questions. (d) Tenses.

Figure 4.21: Improvement relative to C-test scores.
Improvement between the pre- and the post-test was generally higher for the control group for C-test scores <35%, and higher for the intervention group for C-test scores >35% and <55%.

with intermediate C-test scores, data for high C-test scores being too sparse to yield representative results of large learner populations.

This conclusion that meso-adaptivity is beneficial for learners with higher C-test scores is unfortunate if we assume that learners with lower C-test scores also struggle more with the practice exercises and thus require increased support to complete the exercises. However, Pearson's correlation between C-test scores and performance on practice exercises is very low at $\rho = .1276$. The correlation is also low between C-test

scores and English grades ($\rho = .3774$), although it is considerably higher between C-test scores and the grammar pre-test ($\rho = .5318$). It is therefore possible to support learners through meso-adaptivity who perform poorly on practice exercises, and at the same time achieve higher learning gains, if they performed moderately well on the entry C-test.

While the evaluations so far analyzed the intervention as randomized, we next assess the robustness of the obtained results by analyzing the effect of the intervention as treated. As it has not yet been established what amount of meso-adaptive scaffolding is required for a learner to be categorized as treated, we incrementally increased the threshold between the *treated* and *not treated* categories. Since very few learners received meso-adaptivity in more than 21 exercises, as illustrated by the barplot in Figure 4.22, we considered treatment thresholds up to this number. The amount of meso-adaptive scaffolding received by a learner was operationalized as the number of exercises in which meso-adaptivity was applied. Only those learners were considered in each increment step that met or surpassed the treatment threshold. Learners in both study conditions who submitted less exercises, with or without meso-adaptive scaffolding, than specified as treatment threshold were excluded from the analyses.

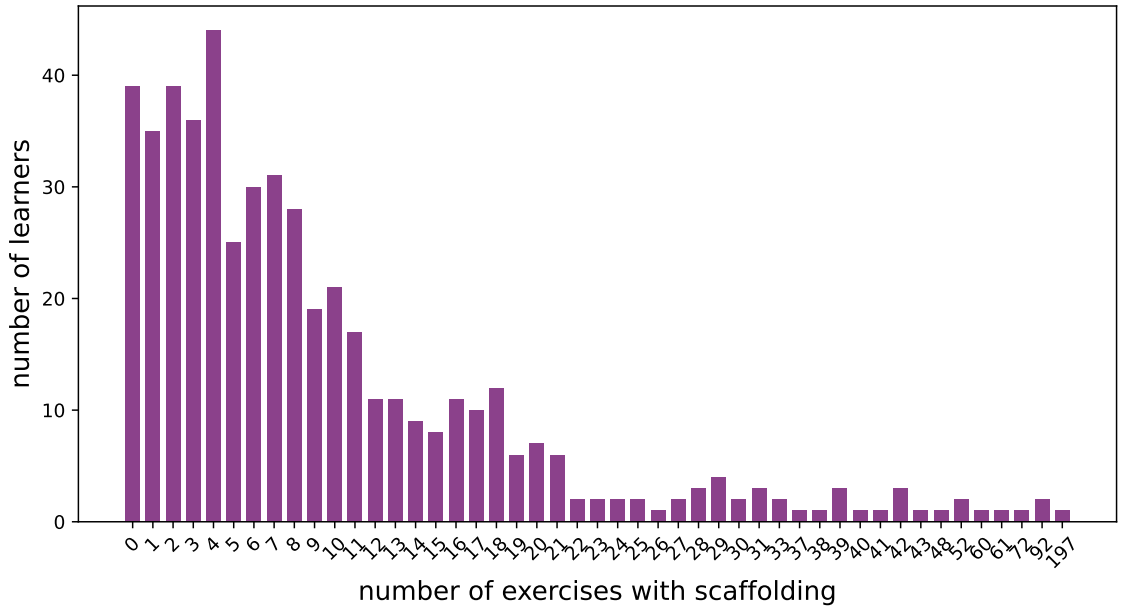
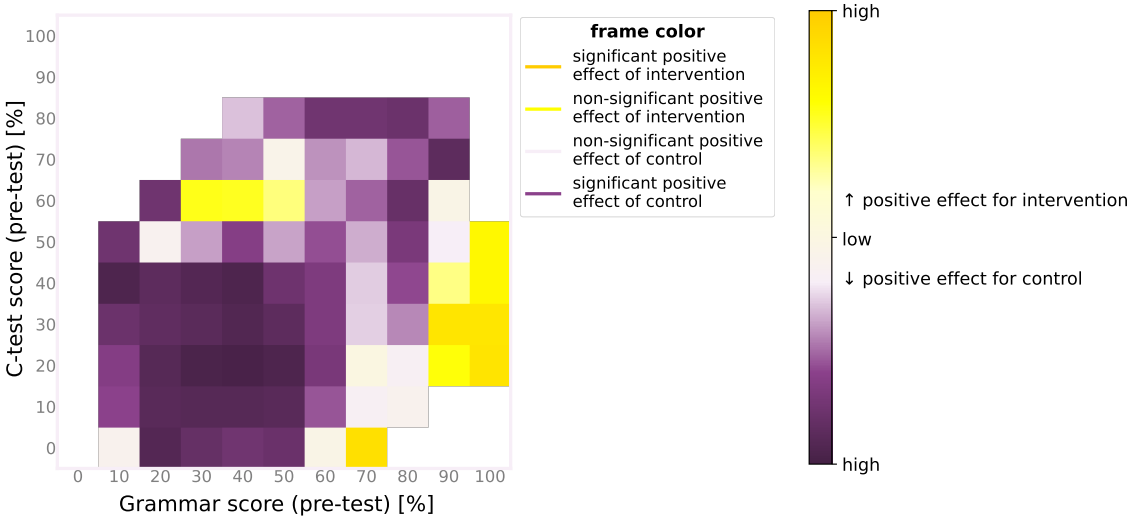


Figure 4.22: Distribution of numbers of scaffolded exercises across learners. Most learners in the intervention condition received meso-adaptive scaffolding in up to 20 exercises.

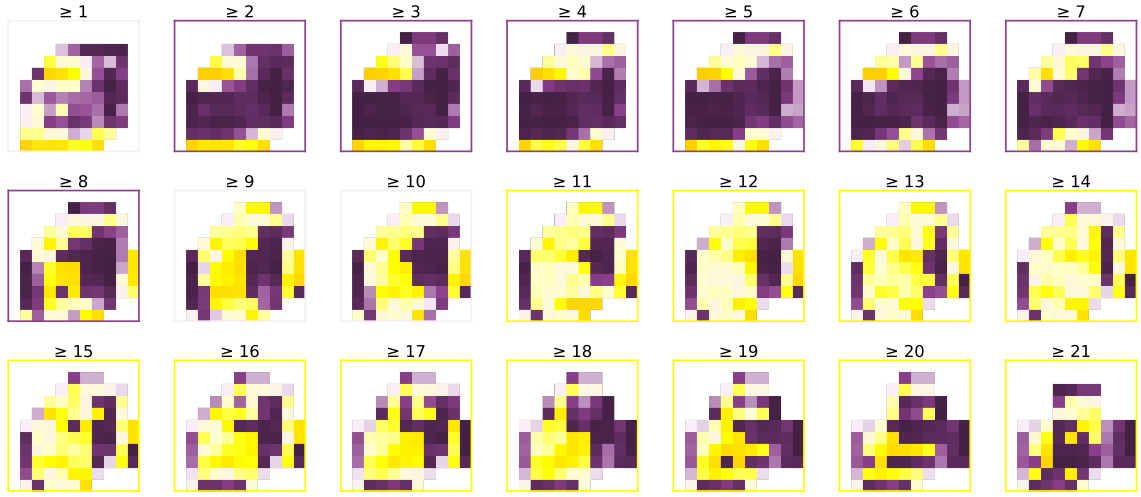
Apart from the C-test scores, we also considered the potential impact of the entry grammar test scores in these evaluations. To this purpose, we incremented both scores (C-test and grammar test) in steps of 10% accuracy. Learners within ± 20% accuracy were considered for each increment step. This resulted in a significance score of either a positive or a negative effect of the intervention for each combination of C-test score, grammar test score, and treatment threshold. Significance was determined based on an ANCOVA analysis controlling for the number of submitted exercises by the learner.

The results are visualized in the heatmaps in Figure 4.23, which attribute different colors depending on whether the intervention was beneficial (yellow) or not (purple), and of increasing richness with increasing significance of the effect. The color of the frame surrounding each heatmap indicates the effect for the entire learner sample with the treatment threshold indicated above the plot. A dark color indicates a significant effect whereas a light color indicates a non-significant effect. The plots illustrate that while the treatment threshold is low, high C-test scores favor a beneficial effect of the treatment. The C-test score loses in importance with increasing number of scaffolded exercises as learners become familiar with the provided assistance. Beneficial effects of the treatment are then distributed across the entire spectrum of C-test scores. Heatmaps for the individual `learning targets` are provided in Figures A.1–A.3 of Appendix A.3.2. The overall effect of the treatment is beneficial, although not significantly, across all `learning targets` within a certain range of treatment thresholds. This range includes the lowest number of scaffolded exercises for *Tenses* (10) whereas *Questions* and *Conditionals* require slightly more scaffolded exercises (11) in order for a beneficial effect of the treatment to manifest. Simple grammar topics such as tenses with a primary focus on morphology thus seem to require less treatment in order for meso-adaptivity to be beneficial for learning gains than more complex topics that require lexical, morphological, and syntactic knowledge, such as questions and conditionals. Significant positive effects of the treatment can be observed for a small number of C-test and grammar score ranges. Overall significant effects, this time in favor of the control condition, are observed for treatment thresholds between two and eight exercises.

In this analysis, we considered students as not treated if they received less exercises with meso-adaptive scaffolding than the treatment threshold. Despite our efforts to



(a) As randomized.



(b) As treated (per number of scaffolded exercises).

Figure 4.23: Overall significance of the treatment for learner subgroups. High C-test scores promote a beneficial effect of the treatment (yellow) when learners have been exposed to little scaffolding. Overall significant effects (dark borders) appear mainly with high treatment levels and in benefit of the control group (purple).

control for a bias by only including learners who submitted as many exercises as the treatment condition and controlling for the number of submitted exercises, this might place unmotivated students with potentially lower learning gains in the control condition. In addition, a single treatment threshold does not take into consideration the possibility that treatment might only be beneficial when applied at a certain extent. Too many scaffolded exercises might have a negative effect on learning gains

(Vanderhyde, 2019). We therefore repeated the analysis, this time replacing the treatment thresholds with a fixed number of exercises in which the learner received meso-adaptivity. We applied a tolerance of 2 exercises more or less than this number in order to supply the analyses with sufficient learner data. In the control condition, we only considered learners who would have received meso-adaptivity in the same number of exercises had they been in the intervention condition. Significance of the effect of the treatment was determined based on a mixed-effects analysis controlling for the number of exercises submitted by the learner. We conducted the analysis for the overall dataset as well as for the three `learning targets` *Tenses*, *Questions*, and *Conditionals*, for which we only considered exercises submitted in the corresponding practice `learning targets`.

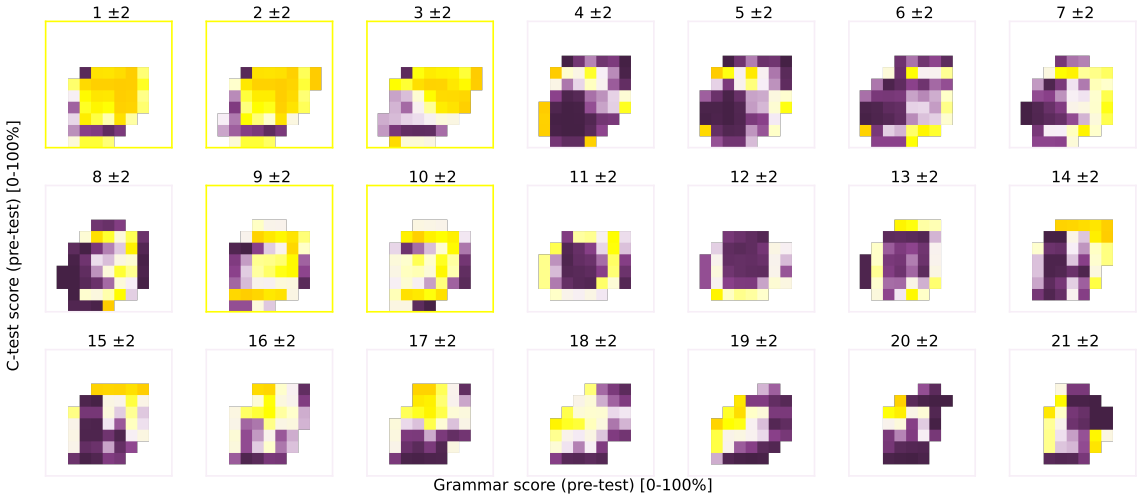
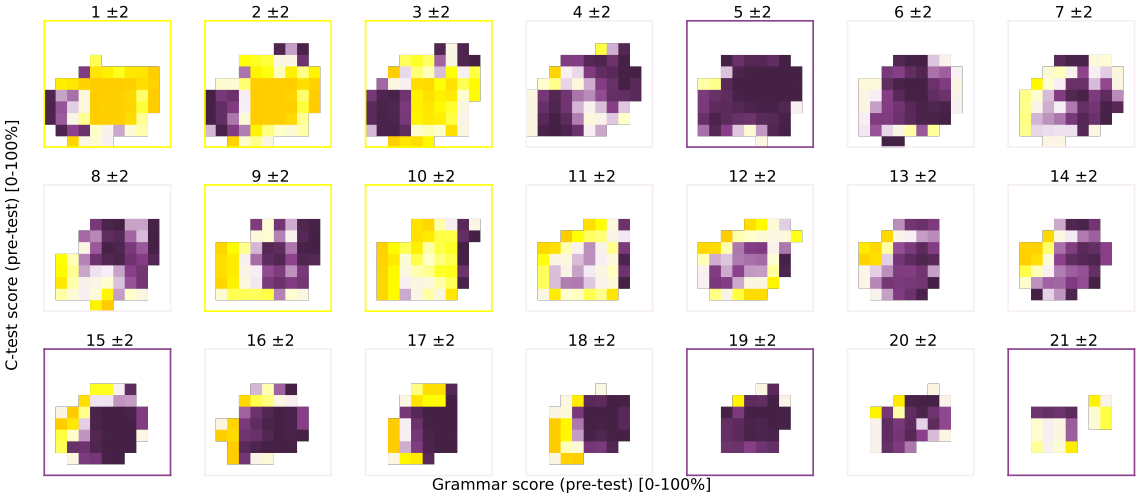
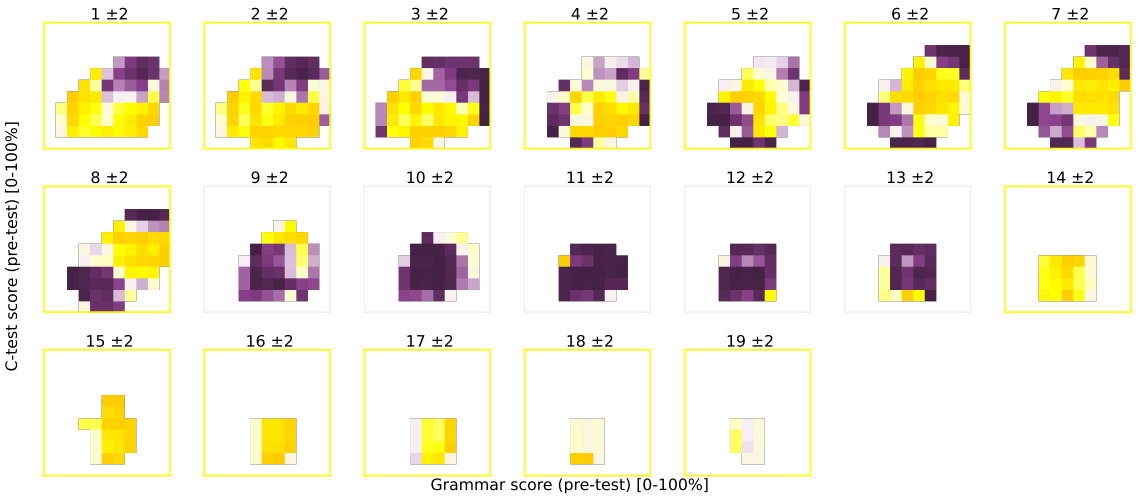


Figure 4.24: Overall significance of the treatment for learner subgroups by number of scaffolded exercises.
Treatment was beneficial (yellow) for learners when they received meso-adaptivity in up to three or in 9 to 10 exercises. None of the effects are significant for the entire learner population (dark borders).

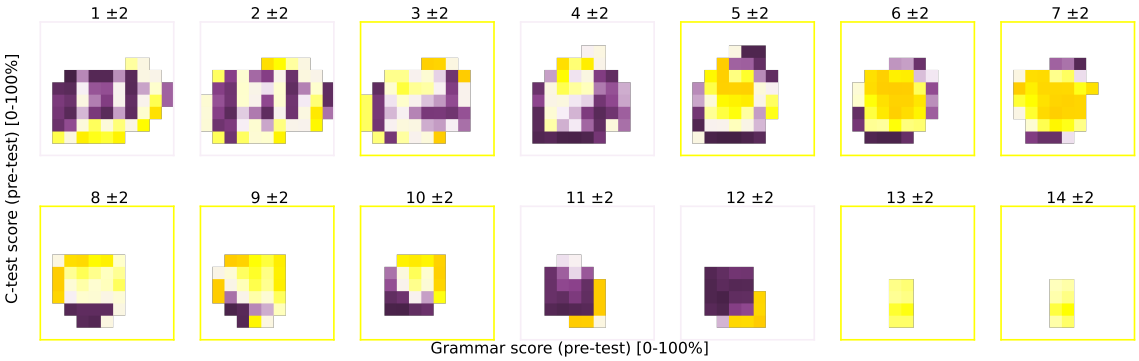
For this analysis, the heatmaps in Figure 4.24 illustrating the results indicate that meso-adaptivity was overall beneficial for learners who received this type of scaffolding in up to 3 or in 9 to 10 exercises. There are, however, no significant effects without considering C-test scores and scores on the grammar pre-test. This time, we observe considerable differences across `learning targets`, as can be seen in Figure 4.25. Treatment was beneficial with up to 3 or 9 to 10 scaffolded exercises on *Tenses*, with up to 8 or 14 to 19 scaffolded exercises on *Questions*, and with 3, 5 to



(a) Tenses.



(b) Questions.

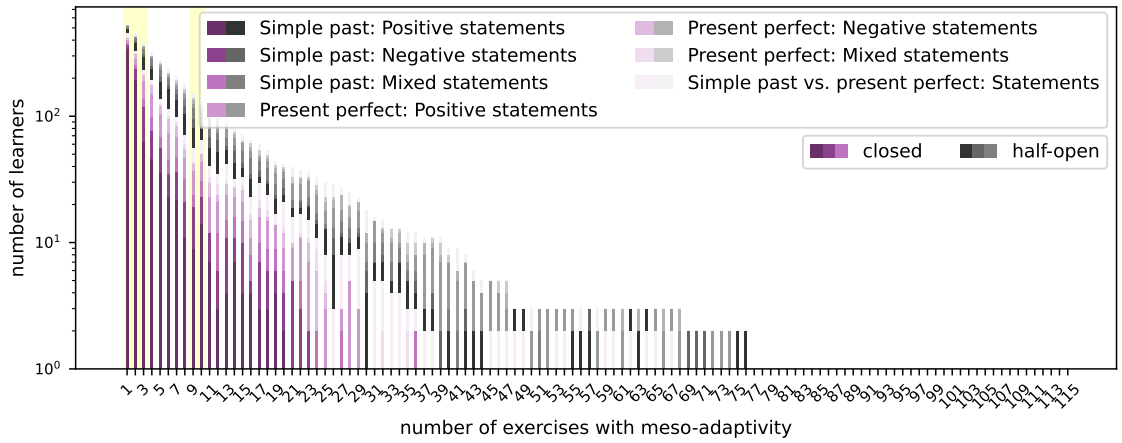


(c) Conditionals.

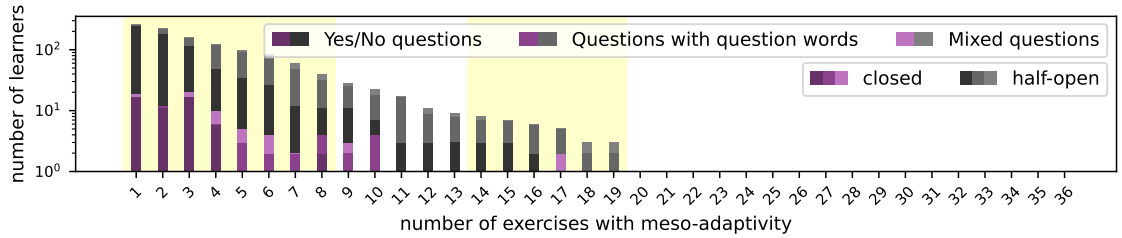
Figure 4.25: Significance of the treatment for learner subgroups by number of scaffolded exercises per `learning target`.
The effect of meso-adaptive treatment differed considerably across `learning targets`.

10, or 13 to 14 scaffolded exercise on *Conditionals*. Not considering the isolated beneficial treatment intensity of 3 scaffolded exercises on *Conditionals*, there are thus two main ranges of treatment intensity for each `learning target` within which treatment was overall beneficial. Overall significant effects for individual `learning targets` appear only with *Tenses* and in favor of the control condition with 5, 15, 19, and 21 scaffolded exercises.

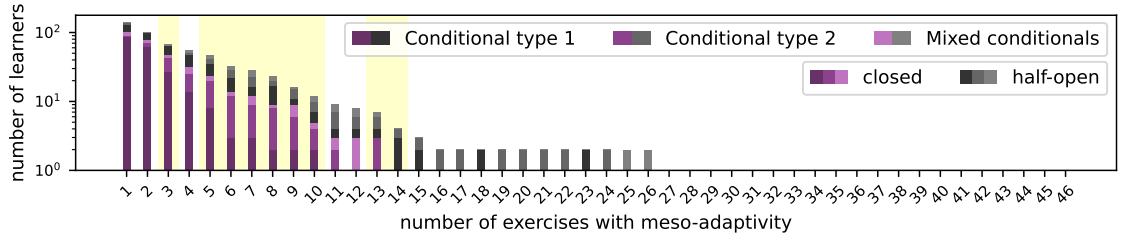
We hypothesized that low numbers of scaffolded exercises correspond to closed exercises while higher numbers correspond to half-open exercises. Alternatively, it is also possible that different ranges within which meso-adaptivity tended to have a positive effect on learning gains represent different `realizations` of the `learning target`. In order to test these two hypotheses, we examined the exercise type in terms of closed or open exercises, and the `realization` of the exercises where learners received meso-adaptive scaffolding. The bar plots in Figure 4.26 indicate the number of learners who received exercises in these distinctions at the corresponding point within their sequence of scaffolded exercises. In the plot for *Tenses*, there are no peculiar changes in the bars at the first threshold identified in the heatmaps of the previous analysis between 3 and 4 scaffolded exercises. However, the bar plot indicates an increase in exercises on *Mixed statements* within the beneficial range of exercises with meso-adaptive adjustments identified in the heatmaps. The bars representing less than 9 and those representing more than 10 scaffolded exercises indicate higher numbers of exercises in `realizations` on positive and negative statements. In the plot for *Questions*, there is a noticeable change with more than 10 scaffolded exercises, after which almost all learners received scaffolding only in half-open exercises. This threshold corresponds roughly to that observed in the heatmaps, namely 9, from which on meso-adaptivity resulted in an overall negative effect. It could indicate that a positive effect of meso-adaptivity on learning gains requires multiple half-open exercises in which this treatment is provided, so that an overall beneficial effect of meso-adaptivity appeared in the heatmaps only after several half-open exercises had been scaffolded. In the plot for *Conditionals*, this effect is reversed. More than 13 scaffolded exercises correspond to most learners submitting only half-open exercises. Within this range, the heatmaps of the previous analysis indicated an overall beneficial effect of meso-adaptivity. In the range between 5 and 10 scaffolded exercises, within which meso-adaptivity was, in the heatmaps, identified to in addition have a positive effect, most learners submitted



(a) Tenses.



(b) Questions.



(c) Conditionals.

Figure 4.26: Type and `realization` of exercises with meso-adaptivity. Ranges of scaffolded exercises within which treatment was beneficial (yellow background) correspond either to more challenging `realizations of learning targets` or to half-open exercises, although the beneficial effect is in some cases delayed.

exercises in the `realization` *Conditionals type 2*. Submissions before that threshold covered primarily the `realization` *Conditional type 1*.

We hypothesize that for closed exercises, receiving meso-adaptive adjustments initially introduces a novelty effect that draws the learners' attention to the exercises so that they engage more with the practice material. As this effect wears off, the

previously discussed demotivating effect sets in. For more complex `realizations of learning targets` such as *Mixed statements* or *Conditional type 2*, however, learners need more assistance to better understand and solve the exercises, and meso-adaptivity can provide it in a form that improves learning gains. In closed exercises, meso-adaptivity thus seems to have a positive effect on learning gains if other parameters increase practice difficulty. In half-open exercises, on the other hand, the supportive adjustments seem to be less intuitive so that beneficial effects of the treatment appear only after learners have become familiar with the adjustments.

When also considering C-test scores and performance on the entry grammar test, we also notice differences across `learning targets`. For *Tenses*, students with high previous knowledge (reflected by higher grammar test scores) benefited more from treatment in only few exercises while learners with low previous knowledge benefited from the intervention if they received meso-adaptive support in more exercises. Minimal treatment was thus sufficient to result in a positive effect of the intervention if the learner was already rather proficient in the `learning target`, whereas less proficient learners needed more extensive treatment. This trend is reversed for the more complex `learning target` *Questions*. Here, little treatment was beneficial for all learners, but only learners with high previous knowledge benefited from a higher degree of intervention. Although there is no intuitive explanation for this pattern, the beneficial effect of the intervention with higher previous knowledge on *Questions* correlates with higher C-test scores. Viewed across all `learning targets`, the effect of the treatment was beneficial for learners with either very low or rather high scores in the entry C-test.

Further research is needed to investigate what personal trait C-tests assess that impacts the suitability of meso-adaptivity. It does not seem to be general language proficiency, as usually assessed through C-tests, which would also explain why C-tests could not successfully resolve the cold-start problem in Section 3.1.3. Performance on C-tests has been associated with working memory capacity in past research (e.g., Daller et al., 2021; Babaii and Fatahi-Majd, 2014). Since no working memory data was collected in the AI2Teach study, we cannot confirm whether such a relationship exists here. However, correlations of C-test scores and working memory test scores were very low in the I4S ($\rho = .2456$) and Didi ($\rho = .1188$) studies. Similarly, Pearson's correlation was low ($.0076 \leq \rho \leq .2775$) between C-test scores

and motivational metrics collected in the I4S study. The metrics, aggregated by Deininger et al. (2024), comprised intrinsic, attainment, utility, and cost value as well as L2 motivation, effort, frustration resistance, task persistence, and conscientiousness. Since C-tests seem to capture none of the evaluated learner parameters, we hypothesize that they measure two distinct learner characteristics. On the one hand, they probably indeed measure a learner's general language proficiency, indicating the learner's need for increased support with practice exercises when the score is very low. This hypothesis is supported by the differences we observed between individual `learning targets`, as a beneficial effect of the treatment did not show for learners with low C-test scores in the `learning target` *Conditionals*, but only with *Tenses* and, to some degree, with *Questions*. Since conditional sentences are the most challenging `learning target` of the three for learners of English (Larsen-Freeman et al., 2016), meso-adaptivity might not have been able to bridge the gap between exercise difficulty and the proficiency of very weak learners. On the other hand, C-tests might provide a measure for a learner's strategic competence in terms of the ability to make use of distributed and changing supportive evidence in language exercises (Bachman, 1990). Learners with high C-test scores thus benefit more from meso-adaptivity than most learners with lower C-test scores. Contrary to the grammar pre-test, C-tests in this case do not assess previous language knowledge.

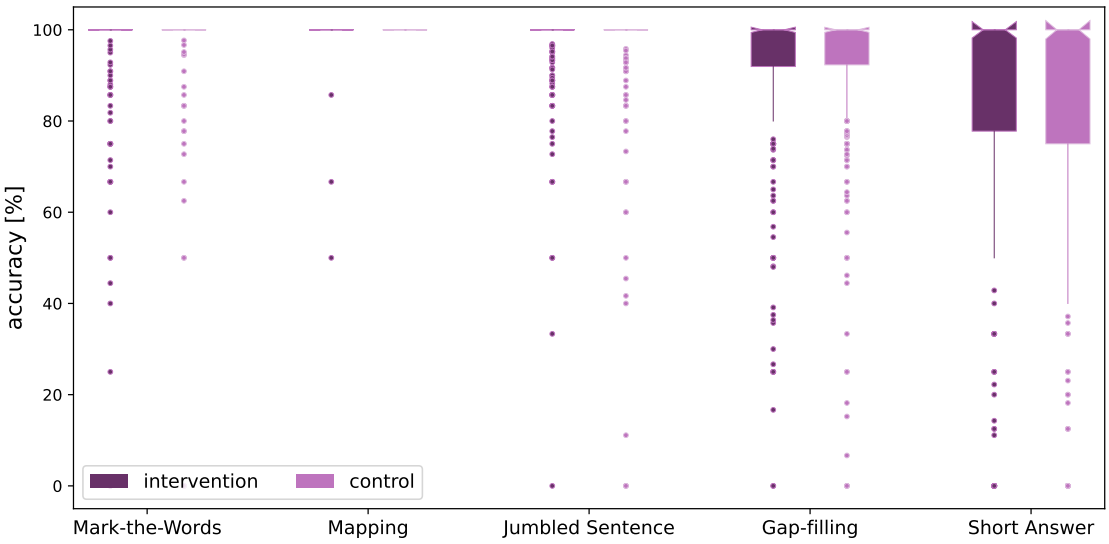


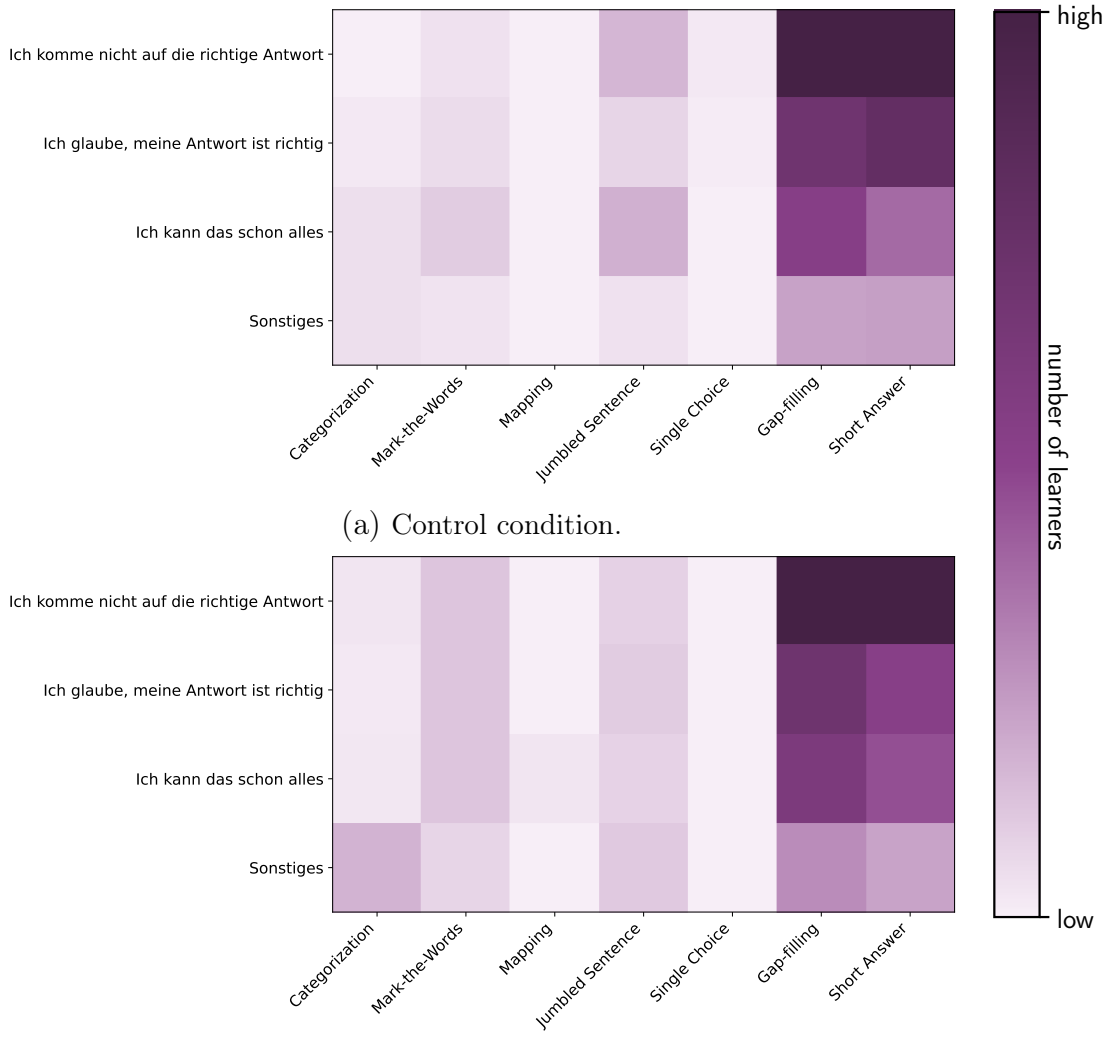
Figure 4.27: Accuracy on practice exercises at submission.
Accuracy at submission was rather high for all learners on all exercise types.

Offering meso-adaptivity is, however, only necessary if learners cannot eventually come to the correct solution by themselves and thus learn from these exercises. We therefore need to take a closer look at the final submissions of answers to actionable elements in order to determine whether, for the two study conditions, they were in the learners' ZPD, thus allowing them to find the correct solution with the provided help. This help was only minimal for the control group in form of micro-adaptivity provided through meta-linguistic feedback if available, and much more extensive for the intervention condition in the form of meso-adaptive adjustments. Figure 4.27 illustrates that the final submissions indeed had much higher accuracies than the first attempts, with the lowest scores for Short Answer exercises at $\mu = 82.67\%$ ($\sigma = 26.99$) for the control group and $\mu = 83.21\%$ ($\sigma = 27.34$) for the intervention group. Interestingly, the effect of the study condition is not significant for any exercise type, so that meso-adaptivity did not significantly increase the learners' probability of getting all actionable elements correct before submitting the exercise.

However, the heatmaps in Figure 4.28, illustrating the results of the poll issued after each incorrect submission, indicate that not being able to find the correct solution was the most frequent reason for incorrect submissions, regardless of whether meso-adaptivity was available or not. Yet, the results suggest that, for Jumbled Sentence exercises in particular, the number of exercises that were submitted incorrectly for this reason was lower in the intervention condition.

The scores of the final submission are in the required range of at least 60% accuracy for most learners across all exercise types so that the available exercise pool and macro-adaptive algorithm are suitable for the target group of 7^{th} grade Gymnasium students. For Gemeinschaftsschule students, on the other hand, mean accuracies were still very low at submission, especially for Short Answer exercises with $\mu = 67.01\%$ ($\sigma = 34.94$), and even Gap-filling exercises showed a mean accuracy of only $\mu = 77.51\%$ ($\sigma = 27.11$), even though their accuracies at first try were comparable to those of Gymnasium students. It should also be noted that all Gemeinschaftsschule students received meso-adaptivity. The easier closed exercises got high accuracy scores for Gemeinschaftsschule students and Gymnasium students alike. For learners of school types other than Gymnasium, easier exercises are thus required throughout, or else more help needs to be provided.

In conclusion, our evaluations showed that the exercises available in the study were



(a) Control condition.

(b) Intervention condition.

Figure 4.28: Distribution of reasons for incorrect submissions.
Darker colors indicate higher numbers of learners selecting the respective answer. The answers were elicited in a popup poll after each incorrect submission. The reasons for incorrect submission included that the student believed that their answer was correct (*Ich glaube, meine Antwort ist richtig*), that the learner could not find the correct answer (*Ich komme nicht auf die richtige Antwort*), that the student was convinced that they were already proficient in the exercise's learning goal (*Ich kann das schon alles*), and other reasons (*Sonstiges*). The reasons for incorrectly submitting an exercise were distributed similarly for both study conditions and across all `learning targets`. For Jumbled Sentence exercises, the number of incorrect submissions due to not finding the answer was lower for the intervention condition.

suitable for Gymnasium students, but not for Gemeinschaftsschule students. Meso-adaptivity was not able to provide enough help to make the exercises suitable for

these learners as well. The treatment also turned out to not be beneficial for all learners. While it had a positive impact on most learners with very low or moderately high C-test scores, our implementation negatively impacted learning outcomes for learners with moderately low C-test scores and also when the intensity of the treatment was too high in easy `realizations of learning targets`. It is, however, possible that a larger group of learners would benefit from meso-adaptivity if engagement with the ITS was increased, or if the implementation of the scaffolds or the timing of providing them was adjusted. Triggering the first meso-adaptive adjustment after the second instead of the first incorrect attempt (Shea, 2002), for instance, or offering meso-adaptivity only when the exercise is within the learner's ZPD (Schulze Bernd and Chounta, 2024), might make it beneficial for either additional or other subsets of the learner population. It is, however, necessary to provide different exercises and adjust sequencing thresholds depending on learner abilities in macro-adaptive sequencing for learners who do not benefit from meso-adaptivity. To this purpose, it is sufficient to differentiate learners at the granularity of stereotypes, as learners at the same class level of the same school type generally performed similar in our study. It is thus necessary to have stereotypes-based exercise pools for different school classes and school types.

We acknowledge that our evaluations concerning meso-adaptivity were biased in that the intervention condition differed from the control condition not only in receiving meso-adaptive adjustments, but also in adjustments to the macro-adaptive exercise selection. While this allowed us to assess whether a reduced macro-adaptive focus on exercise difficulty can be mitigated through meso-adaptivity, it makes it difficult to determine the effect of meso-adaptivity on practice difficulty and learning gains by itself. However, our analyses indicate that meso-adaptivity can be a useful feature if applied carefully.

4.3 Summary

In this chapter, we investigated whether curricular structures and linguistic variability can successfully be considered in exercise selection by shifting the focus from exercise difficulty to these parameters. The resulting implementation both considers pedagogically induced requirements on exercises and adjusts linguistic variability to learner proficiency in a cascade of exercise filters and rankers. The focus on exercise

difficulty is reduced to merely considering the openness of the exercise type. To mitigate this reduced account for exercise difficulty, we introduced meso-adaptivity as a type of adaptivity in-between macro- and micro-adaptivity, which successively and dynamically renders an exercise easier after each incorrect attempt. The presented implementation thus substantiates how adaptivity can be included into an ILTS at different levels. Based on our findings that learners are more likely to persist in finding the correct answer in closed than in half-open exercises, our general strategy to meso-adaptive adjustments reduces the space of possible answers. Specific meso-adaptive strategies differ from one exercise type to another.

The answer to RQ 8, whether meso-adaptive adjustments can leverage exercise difficulty sufficiently to enable learners to solve most exercises, is not straightforward. An evaluation of our implementation found that, while meso-adaptivity reduces perceived exercise difficulty and the number of attempts required to find the correct answer, this lower engagement with each exercise resulted in overall poorer learning gains. These were reflected in lower efficiency as well as effectiveness to reach mastery of the Knowledge Components.

We further found that the effect was not similar for all learners but depended on their C-test performance, previous knowledge of the grammar topic, and the amount of meso-adaptive treatment. We observed a tendency towards meso-adaptivity increasing long-term learning gains when applied to closed exercises until a novelty effect wore off or in difficult `realizations of learning targets`, and when applied to half-open exercises after learners had become accustomed to the adjustments. For different degrees of previous knowledge, the beneficial effect of meso-adaptivity depended on the `learning target` and the extent of the treatment. In addition, learners with either very low or rather high C-test scores benefited more from the treatment. We hypothesize that C-test scores capture the underlying learner characteristics of general language proficiency on the one hand, and of a learner's strategic competence on the other hand. Since the correlation between practice performance and C-test scores is low, there is a learner population that performs poorly on practice exercises but has very low or moderately high C-test scores. For these learners, meso-adaptivity is beneficial as it supports them while working on the exercises and at the same time increases long-term learning gains. Adjusting meso-adaptive parameters might further increase the learner population that benefits from meso-

adaptivity.

However, we also saw that exercise difficulty was similar for both study conditions and that accuracy on the final submissions of practice exercises was at least 60% on all exercise types for most Gymnasium students, including when no meso-adaptivity was available. Meso-adaptivity is thus not necessary to mitigate exercise difficulty for this learner population. For Gemeinschaftsschule students, on the other hand, accuracy values were much lower. Macro-adaptive ILTSs that do not otherwise control exercise difficulty therefore need to maintain different exercise pools with differing complexity ranges for different learner populations in order to account, for instance, for various school types and class levels.

Chapter 5

Selecting Linguistically Varied Exercises Compliant with Curricular Constraints in the Zone of Proximal Development

Parts of this chapter are based on Publications 4 and 5.

Combining all findings from the previous chapters, we acknowledge (1) curricular constraints imposed on macro-adaptive exercise selection in teacher-orchestrated instruction; (2) the need for considering linguistic variability in practice depending on the learner's mastery of the `learning target`; and (3) the need for a minimum of attention to exercise complexity.

In this chapter, we outline the resulting implementation of an adaptivity algorithm selecting exercises matching a learner's profile that conforms to all three requirements. In order to account for linguistic and `skill` variability, consider curricular constraints, and at the same time make sure that the exercises are in the learner's ZPD we propose a Knowledge Tracing (KT) approach in combination with a cascade of filters and rankers. In this approach, exercise selection relies on (1) Bayesian Knowledge Tracing (BKT) to determine when to move on to the next KC, (2) an exercise filter accounting for scope restrictions and exercise difficulty, and (3) an adjustment module considering linguistic variability. We recommend to complement

this macro-adaptivity approach with meso-adaptivity as introduced in Section 4.2 for half-open exercises of easy `realizations of learning targets` and for closed as well as half-open exercises in difficult `realizations`. We suggest, however, to offer the adjustments only after the second incorrect attempt and to apply them only if the learner actively accepts this offer.

Note

Although parts of the algorithm have already been implemented and integrated into the most current version of the FeedBook, we do not yet have user data on which to evaluate the comprehensive approach.

5.1 Tracking Mastery of Knowledge Components

The calculations and decisions about a learner's mastery of a KC are based on an overlay learner model for the domain model presented in Section 3.3.1. This learner model is continuously updated when the learner interacts with an exercise. Although the learner can correct an answer after receiving feedback from the system, only the first attempt is considered for learner model updates.

Pelánek and Řihák (2017) found that simple *n consecutive correct* or *moving average* methods are just as effective as BKT approaches to determine mastery while convincing through their simplicity. These alternative methods do, however, have one major disadvantage. The authors highlight the importance of setting mastery thresholds that correctly identify a learner's mastery of a component. Jansen (2000) point out that the definition of mastery criteria is often very vague. It might be intuitive to use emergence criteria for initial mastery, according to which a linguistic form is acquired when the learner first produces it (Pallotti, 2007). In fact, Pienemann (2015) advocates the use of emergence criteria instead of accuracy criteria, which he deems arbitrary. However, emergence criteria are based on free text production and might not transfer to rather closed exercise types. In order to determine a learner's mastery of a KC, we therefore recommend a simple BKT model if training data is available. These models typically apply a mastery threshold of 95% probability of having mastered the KC, which is assumed in the literature to be a representative probability in systems with fine-grained KCs (Pelánek, 2018). If KCs are defined on a less detailed level for practical reasons, knowledge is usually not

binary (*unknown* vs. *known*) so that a 95% probability of mastery does not necessarily indicate mastery of the entire KC. Although our model tracks each learner's knowledge states of KCs at the level of `properties`, which constitute aggregates of multiple `linguistic constructs`, grammatical knowledge may transfer between the `linguistic constructs` of a `property`. We therefore apply the threshold of 95% probability to determine mastery. Thanks to the RCT conducted within the scope of the AI2Teach project, the model parameters for each KC can be calibrated on the data collected in that study with the implementation presented by Hawkins et al. (2014). The authors found their empirical calibration approach to yield better results with less computation effort than the more widely applied, statistical Expectation Maximization approach (Yudelson et al., 2013). The calculations are based on knowledge sequences created from learner answers by interpreting correct answers as *known* and incorrect answers as *unknown* knowledge states. The approach assumes that there is a single point in the learning curve from which on the KC is mastered. It therefore considers all knowledge states in the knowledge sequence as potential transition points from *unknown* to *known*. For each potential transition point, the algorithm then determines how many of the knowledge states before the transition point match the *unknown* state and how many of the knowledge states after the transition point match the *known* state. The transition point with the highest number of matching knowledge states is assumed to be the time point when the learner reached mastery. The four parameters that the approach aims to calibrate consist of probabilities for a KC's initial knowledge (P_{init}), guessing the correct answer (P_{guess}), accidental provision of an incorrect answer despite having the knowledge (P_{slip}), and transition from a state of not knowing the component to knowing it ($P_{transition}$). Their calculations are given in Formulas 5.1–5.4, where S denotes the set of all learners, $K_s(i)$ denotes the knowledge state of learner s at the relative timepoint i within the exercise sequence I, and $C(i, s)$ denotes the correctness of the answer where a correct answer corresponds to *1* and an incorrect answer corresponds to *0*.

$$P_{init} = \frac{\sum_{s \in S} K_s(0)}{\mid S \mid} \tag{5.1}$$

$$P_{guess} = \frac{\sum_{s \in S} \sum_{i \in I} (1 - K_s(i)) \cdot C_s(i)}{\sum_{s \in S} \sum_{i \in I} (1 - K_s(i))} \tag{5.2}$$

$$P_{slip} = \frac{\sum_{s \in S} \sum_{i \in I} (1 - C_s(i)) \cdot K_s(i)}{\sum_{s \in S} \sum_{i \in I} (K_s(i))} \tag{5.3}$$

$$P_{transision} = \frac{\sum_{s \in S} \sum_{i \in I, i \neq 0} (1 - K_s(i-1)) \cdot K_s(i)}{\sum_{s \in S} \sum_{i \in I, i \neq 0} (1 - K_s(i-1))} \tag{5.4}$$

Note

In order to further refine and improve mastery determination, it is possible to continuously update the BKT model by re-calibrating the parameters with the learner information being constantly recorded in the learner model.

We defined KCs as combinations of `properties` and `skills` in Section 3.3. Macro-adaptive exercise selection using KCs at this level of granularity would require learners to complete large numbers of exercises if all KCs should be mastered in the end. It is, however, easy to imagine that a learner who has mastered the productive formation of irregular verbs in a Short Answer exercise can also correctly identify irregular verbs in a Mark-the-Words exercise. Pelánek (2019) points out that it remains an open research question how the relationship between `properties` and `skills` should be modeled. We therefore suggest to track knowledge based on fine-grained KCs, but determine mastery based on aggregated, higher-level `properties` and `skills`.

To this purpose, the algorithm re-calculates the probability of knowing a KC (P_{known}) based on the calculation given in Formula 5.5 and depending on whether the answer was correct (Yudelson et al., 2013). P_{known} is particular to a specific learner and a specific KC. Although P_{known} is calculated for a specific KC, we define the previous

$P_{Known}(i-1)$ as the highest knowledge state that the `property` of the KC currently being updated has achieved in any KC of the system, if there is any. This allows us to model inter-dependencies between KCs. The `skill` is thus not considered for this parameter. If no KC composed of this `property` has been practiced before, we use P_{init} of the KC that we are currently updating.

$$P_{known}(i) = P_{cond}(i) + (1 - P_{cond}(i) \cdot P_{transition}$$

with

$$P_{cond}(i) = \begin{cases} \frac{P_{known}(i-1)\cdot(1-P_{slip})}{P_{known}(i-1)\cdot(1-P_{slip})+(1-P_{known}(i-1)\cdot P_{guess}}, & \text{for answer correct} \\ \frac{P_{known}(i-1)\cdot P_{slip}}{P_{known}(i-1)\cdot P_{slip}+(1-P_{known}(i-1)\cdot(1-P_{guess})}, & \text{otherwise} \end{cases} \tag{5.5}$$

Note

The presented algorithm constitutes a very basic implementation. State of the art implementations integrating item difficulty into mastery prediction (Bulut et al., 2023) might improve prediction accuracy. Considering the number of attempts, correctness at submission, and response times might further improve updates of knowledge states. In addition, forgetting may be incorporated into mastery calculation. Re-calculation of mastery for all students on a nightly basis could ensure that forgetting is tracked correctly and on a regular basis for all KCs.

The algorithm updating mastery checks for all `properties` that the submitted actionable element practices whether the learner has now mastered the KCs associated with that `property` and the `skills` the exercise practices. 'Submitted' in this case means that a learner's attempt on an actionable element has been registered in the back-end. While the exercises are designed for a specific `realization`, the `properties` determined through linguistic annotations may also link to other `realizations`. Updating mastery for those as well allows to acknowledge that some `realizations` and `learning targets` are prerequisites for other `realizations` and `learning targets` and that the probability of having mastered a prerequisite increases with the probability of having mastered a higher-level `learning target`.

`Properties` relevant to the `learning target` *Simple past*, for instance, constitute prerequisite KCs for the `realization` *Conditionals type 2*. Since BKT does not withdraw the mastery status once diagnosed (Shen et al., 2024), it is sufficient to only update KC masteries if the current KC status changes from *not mastered* to *mastered*. Although we follow the most widely applied probability threshold of 95% for knowledge estimation to determine mastery (Doroudi, 2020), our approach makes it possible to set different thresholds for P_{known} for different KCs and learners in order to define when mastery is reached and the learner may progress to the next KC. It is thus rather easy to extend the target group using the system to, for example, other school types with generally lower expectations on performance.

In an ideal world, an incorrect answer to an actionable element practicing a `property` would still be considered positive evidence for mastery of that `property` if the learner's error is unrelated to the specific `property`. In a realistic world, however, we are faced with a number of difficulties. For one, error diagnosis is not always accurate. This considerably affects the reliability of this idealized approach to determining `property` mastery. For another, it is not always trivial to determine whether an error is related to a `property`. Looking at Example 5.1a it seems rather obvious that the misspelling of the verb *shines* is merely a typo. Yet whether the same could be said of Example 5.1b is highly disputable. This raises the question where to draw the line. One might even question whether incorrect formation of the verb should be counted as an error related to `properties` of the `learning target` *Conditional sentences* at all. After all, conditional sentences constitute higher-level constructs that only require using verbs in a specific tense and aspect. The learning goal for conditionals might thus be defined as using verbs in correct tense and aspect under the assumption that the prerequisite of correct formation has already been mastered. Yet again, the learner's intentions concerning the produced form are unclear as one cannot for certain determine which tense and aspect the learner intended to use if the form is incorrect. Example 5.1c might showcase incorrect formation of either simple present or simple past tense, and establishing a target hypothesis for this error is challenging not only for an algorithm but for human annotators as well. This issue again hints at the non-optimal accuracy of error diagnosis. For this reason, we consider any incorrect answers of an actionable element practicing a `property` as negative evidence.

5.1 a. If the sun **shinse** tomorrow, we will go hiking.
b. If the sun **shins** tomorrow, we will go hiking.
c. If the sun **shinet** tomorrow, we will go hiking.

In order to prevent overloading learners with too many KCs to master, we use higher level KC definitions to determine the milestones a learner needs to master. These encompass high-level `skill` as well as `property` KCs. Each of these high-level KCs is expected to be mastered in closed as well as in half-open exercises. A high-level `skill` KC is specific not to a `property` but to a `realization`. It is mastered once the learner has obtained mastery of the `skill` in combination with an arbitrary `property` of the `realization`. We hereby assume that the `properties` within a `realization` are similar enough to justify that each `skill` needs to be mastered in combination with a single `property` of each `realization` only. A high-level `property` KC is in turn mastered if the learner has obtained mastery for a KC encompassing that `property` and an arbitrary `skill`. When determining mastery of an entire `realization of a learning target`, mastery levels of these high-level KCs need to be aggregated. In this aggregation, `properties` should be weighted higher than `skills` if the focus of mastery is to be on linguistic aspects rather than on `skills`. This can be achieved by aggregating all high-level `skill` KCs into a single, artificial high-level KC. A `realization` is then mastered if all its high-level `property` KCs and the artificial `skills` KC have been mastered.

Thanks to the tree structure in which the domain model is organized, it would in principle be sufficient to persist only knowledge states of the fine-grained KCs to the overlay learner model and dynamically determine `realization` mastery when an exercise is selected. However, in order to facilitate processing and speed up performance, we maintain aggregate progress metrics for the high-level KCs on the `property`, `skill`, and `realization` levels as well.

In Section 2.2.2 we argued for the benefits of maintaining a bug model in addition to an overlay model. This model stores positive and negative evidence of a learner's mastery of linguistic forms. These linguistic forms do not necessarily correspond to `properties` of KCs. The evidence is determined based on an analysis of the learner's answer. In our implementation, error diagnosis is integrated into the system's feedback module. Fuzzy matching tries to identify a fitting answer hypothesis

when a learner provides an incorrect answer. If an answer hypothesis matching a learner's answer can be found, the associated misconception is provided to the adaptivity algorithm, which consequently increases the negative evidence score for that misconception in the bug model. If the answer is correct, the positive evidence score for each misconception for which the actionable element has an answer hypothesis is incremented by *1*.

The thus populated learner model can then be used for exercise selection.

5.2 Selecting Exercise Candidates

Storing exercises as fixed groupings of items considerably increases the challenge of selecting the best exercise for a learner. When selecting exercises based on their difficulty fit, it raises the issue of aggregating item-level difficulties into exercise-level difficulty. It is unlikely that all items of an exercise are equally suitable for a learner, so that the most suitable item's difficulty fit is mitigated by the less optimal fit of the other items. In a variability-based approach, similar issues arise since combinations of linguistic features resulting in optimal variability differ from one learner to another depending on which `linguistic constructs` they have already practiced. The pool of exercises is finite to that no exercise might be available that perfectly matches the learner's needs.

We therefore use an exercise design in which exercise items are stored as individual items that are independent of each other. In Section 4.2.2, we argued that static exercise definitions are necessary in order to support global elements, such as bags of words containing the lemmas of the forms to insert into blanks of a Gap-filling exercise. Distractor lemmas in a bag of words must not constitute lemmas of any alternative correct target answer in order to be compatible with all exercise items. In addition, each lemma must fit only into a single gap. We address this issue by specifying all candidates for such global elements that are compatible with an exercise item in the item itself. The global elements of the entire exercise then consist of the intersection of the global element candidates of all exercise items that are presented to the learner. For example, one exercise item with the target answer lemma *take* might specify *run*, *sing*, *cook*, *fly*, and *give* as possible semantic distractors. Another item with the target answer lemma *give* might specify *take*, *run*,

cook, *fly*, and *dance* as distractor options. Since the distractor options of both items contain the other item's target answer lemma, the two items can be incorporated into the same exercise. An exercise comprised of only these two items would then list *run*, *cook*, and *fly* as semantic distractors. In order to reduce redundancy of data management, we introduce a meta-exercise that aggregates all items of the same exercise type and practicing the same KC. Meta-information such as exercise instructions can thus be specified for the meta-exercise instead of redundantly for each item.

In order to better cater to teachers' needs, the assembly of `learning targets` into learning units is also flexible. Teachers may thus compile custom learning units. They may either include the entire `learning target`, or only certain `realizations` and `properties` thereof.

Within this setup, the approach we present adopts large parts of the macro-adaptivity algorithm described in Section 4.1. Integrating the findings of all our evaluations results in the following adaptivity algorithm.

The first step consists in selecting the next `property` to practice. In Mastery Learning, one KC is practiced after the other (Apfelbaum et al., 2019). A new KC is practiced once the previous one has been mastered (Pelánek, 2017). Other approaches have, however, been explored and discussed in the literature. Empirical research on interleaved vs. blocked practice has yielded contradictory results (e.g., Suzuki et al., 2022; Nakata and Suzuki, 2019; DeKeyser, 2018). Pienemann's (1998) Processability Theory provides an alternative means for defining exercise sequencing based on developmental sequences of SLA (Ipek, 2009). Mansouri and Duffy (2005) present empirical evidence for the effectiveness of approaches that are grounded in this theory based on evaluations of the acquisition of syntactic structures. Shea (2002) suggests that, while interleaved practice is more effective for simple skills, blocked practice is better for practicing complex skills. Since `properties` can be mastered in a single practice session (a characteristic of simple skills), but have multiple degrees of freedom in terms of their concrete realization (a characteristic of complex skills), it is not entirely clear whether they should be considered a simple or a complex skill. Moreover, interleaved practice of `properties` would result in very late progress changes in a student dashboard and rather quick full mastery once progress changes have set in. In order to make progress updates more steady,

we therefore focus on one `property` at a time. However, `properties` are not the only high-level KCs considered for mastery. `Skill` mastery also constitutes a factor relevant to progress. Since all `skills` together are treated as a single artificial KC in the determination of mastery, rotating them does not have the effect on progress update frequency that we identified for `properties`. Defining an optimal `skill` sequence would not be trivial either since, for example, the order in which morphology (*formation* `skill`) and syntax (*word order* `skill`) are best acquired is controversial (Jensen et al., 2020). In addition, since `skills` are related to the exercise type, blocked practice of `skills` would increase monotony in exercises. We therefore rotate practice of not yet mastered `skills` but practice `properties` one at a time. The only exception where interleaved practice is applied to `properties` occurs in the last `realization` of most `learning targets`. These *mixed* `realizations` are pedagogically defined to enforce joint practice of all supported `realizations` of the `learning target` in the same exercise. Interleaved practice is thus also incorporated at the level of `realizations of learning targets`, but only after first introducing each `realization` in a blocked practice format.

`Property` selection thus starts with filtering and ranking the potential `realizations`

Algorithm 5.1: Ranking the `realizations`

Require: learningTarget.includedRealizations ▷ ordered as in the domain model

```
realizationCandidates ← [ ]
if userSelectedRealization = NULL then
    masteredRCs ← [ ]
    notMasteredRCs ← [ ]
    for each realization ∈ learningTarget.includedRealizations do
        if realization ∈ learner.masteredRealizations then
            masteredRCs ← masteredRCs + realization
        else
            notMasteredRCs ← notMasteredRCs + realization
        end if
    end for
    realizationCandidates ← notMasteredRCs + masteredRCs
else
    realizationCandidates ← userSelectedRealization
end if
```

of the `learning target` as outlined in Algorithm 5.1. The potential `realizations` encompass the user selected `realization` if that entry point was used, or all included `realizations` of the `learning target` otherwise. If multiple `realizations`

Algorithm 5.2: Selecting the next `property`

Require: realization.includedProperties ▷ ordered as in the domain model

propertyCandiates ← [ ]
if *learner.proficiency* = *advanced* **then**
  redundantProperties ← [ ]
  for each *realization* ∈ *realizationCandiates* **do**
    for each *p*1 ∈ *realization.includedProperties* **do**
      if ∃*p*2 ∈ *realization.includedProperties* : *p*2.*position* > *p*1.*position* & ∃*c*1 ∈ *p*1.*linguisticConstructs* : ∃*c*2 ∈ *p*2.*linguisticConstructs* : ∀*a* ∈ (*c*1.*coreAnnotations* + *c*1.*environmentAnnotations*) : *a* ∈ (*c*2.*coreAnnotations* + *c*2.*environmentAnnotations*) **then**
        redundantProperties ← *propertyCandidates* + *p*1
      else
        propertyCandidates ← *propertyCandidates* + *p*1
      end if
    end for
  end for
  propertyCandidates ← *propertyCandidates* + *redundantProperties*
else
  for each *realization* ∈ *realizationCandiates* **do**
    propertyCandidates ← *propertyCandidates* + *realization.includedProperties*
  end for
end if

for each *property* ∈ *propertyCandidates* **do**
  if *property* ∉ *learner.masteredProperties* **then**
    return *property*, *realization*
  end if
end for

selectedProperties ← [ ]
for each *realization* ∈ *realizationCandidates* **do**
  selectedProperties ← *selectedProperties* + *realization.includedProperties*
end for
return *selectedProperties*, *realizationCandiates*

result from this step as possible candidates for being practiced next, they are ranked according to whether the learner has passed the respective diagnostic exercise (*masteredRC*). `Realizations` without passed diagnostic exercise (*notMasteredRC*) are prioritized. Within each of these two groups, `realizations` are ranked by the order specified in the domain model.

The next step, illustrated in Algorithm 5.2, determines the `property` to practice. Candidates encompass all included `properties` that have not yet been mastered of the potential `realizations`. They are ranked according to the order specified in the domain model within the previously determined `realization` ranking. However, for advanced learners, higher ranking positions are occupied by `properties` that cannot be practiced by any succeeding `property`. The remaining `properties` are organized in the lower part of the ranking. Figure 5.1 illustrates that a `property` can be practiced by a succeeding `property` if all core annotations and all environment annotations of at least one of its `linguistic constructs` are contained in the core and environment annotations of at least one `linguistic construct` of a `property` that constitutes a right-hand side sibling in the domain model. For all learners, the `property` in the highest ranking position and its `realization` are practiced next if any `properties` can be found. Otherwise, all `properties` of all `realization` candidates are considered.

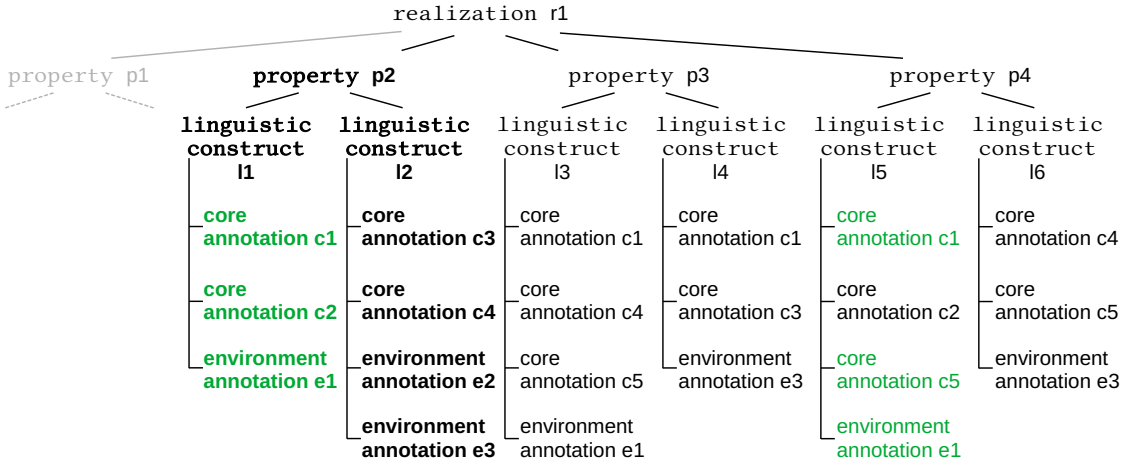


Figure 5.1: Skipping of `properties`.
`Property` p2 can be skipped for advanced learners because `property` p4 constitutes a right-hand sibling whose `linguistic construct` l5 contains all core and environment annotations (c1, c2, e1) of p2's `linguistic construct` l1.

The subsequent exercise selection at the same time filters meta-exercises that contain items fulfilling the selection requirements, and items within a meta-exercise. The initial pool of candidates encompasses all meta-exercises created for the learner's stereotype profile. The filters encompass obligatory filters as well as optional filters. The obligatory filters, outlined in Algorithm 5.3, constitute hard requirements. They first discard all meta-exercises that were not created for any of the `realization` candidates. They then filter out exercise items that were included in the exercise practiced most recently by the learner. The third filter is only applied if the learner has not yet mastered all `properties` of the `realization` candidates. This filter keeps only exercise items that practice the selected `property`.

Algorithm 5.3: Obligatory exercise filters

Require: *orderedSkills* ▷ most recently practiced `skill` last

```
filteredCandidates ← [ ]
for exercise ∈ exerciseCandidates do
    if exercise.realization ∈ selectedRealizations then
        filteredItems ← [ ]
        for item ∈ exercise.items do
            if item ∉ learner.lastpracticedExercise.items then
                if allPropertiesMastered = False then
                    if selectedProperty ∈ item.practicedProperties then
                        filteredItems ← filteredItems + item
                    end if
                else
                    filteredItems ← filteredItems + item
                end if
            end if

            if |exercise.items| ≥ 5 then
                filteredCandidates ← filteredCandidates + exercise
            end if
        end for
        exercise.items ← filteredItems
        filteredItems ← [ ]
    end if
end for
exerciseCandidates ← filteredCandidates
```

All subsequent filters constitute soft requirements imposed on meta-exercises and their items. They may be reverted if their output does not return any exercise items. If a meta-exercise contains less then 5 remaining items after a filter, the meta-exercise is rejected. If no meta-exercise remains after a filter, all candidates remaining before that filter are considered again and the filter is thus bypassed. The

Algorithm 5.4: Optional exercise filters for not mastered `realizations`

Require: allPropertiesMastered = $False$
Require: exerciseCandidates

$filteredCandidates \leftarrow [\ ]$
for $skill \in orderedSkills$ **do**
  if $filteredCandidates = \emptyset$ **then**
    for $exercise \in exerciseCandidates$ **do**
      if $skill \in exercise.skills$ **then**
        $filteredCandidates \leftarrow filteredCandidates + exercise$
      end if
    end for
  end if
end for
if $filteredCandidates \neq \emptyset$ **then**
  $exerciseCandidates \leftarrow filteredCandidates$
  $filteredCandidates \leftarrow [\ ]$
end if

for $exercise \in exerciseCandidates$ **do**
  if $learner.masteredClosedTypes_{realization} = True$ **then**
    if $exercise.type \in halfOpenTypes$ **then**
      $filteredCandidates \leftarrow filteredCandidates + exercise$
    end if
  else
    if $exercise.type \in closedTypes$ **then**
      $filteredCandidates \leftarrow filteredCandidates + exercise$
    end if
  end if
end for
if $filteredCandidates \neq \emptyset$ **then**
  $exerciseCandidates \leftarrow filteredCandidates$
  $filteredCandidates \leftarrow [\ ]$
end if

filters outlined in Algorithm 5.4 are only applied if the learner has not mastered the entire `realization`. The `skill` filter rejects all meta-exercises that do not practice the next `skill` in the rotating `skill` sequence for which candidates are available. The algorithm keeps only meta-exercises of closed types in case the learner has not yet mastered closed exercises for the selected `skill`, or meta-exercises of half-open types otherwise.

The filter outlined in Algorithm 5.5 is applied in the alternative case. If the learner has mastered the entire `realization`, exercise items of all meta-exercises are prioritized that practice misconceptions of a `linguistic construct` for which the learner's accuracy is below an 80% threshold.

Algorithm 5.5: Optional exercise filters for mastered `realizations`

Require: allPropertiesMastered = $True$
Require: exerciseCandidates

for $exercise \in exerciseCandidates$ **do**
  for $item \in exercise.items$ **do**
    for $annotation \in item.linguisticAnnotations$ **do**
      if $annotation \in learner.misconceptions$ **then**
        $filteredItems \leftarrow filteredItems + item$
      end if
    end for
  end for
  if $|filteredItems| \geq 5$ **then**
    $exercise.items \leftarrow filteredItems$
    $filteredItems \leftarrow [\ ]$
  end if
end for

The last filters, presented in Algorithm 5.6, are applied regardless of whether the learner has already mastered all `properties`. An item is only kept if it has not been practiced before. For the items passing that filter, their item set is considered. Items belonging to an item set of which at least one item has already been practiced are filtered out. If any of the remaining candidates are mandatory exercise items that the learner has not yet practiced, the candidate pool is reduced to those. If the filter concerning the items' practice states was rejected, the candidate list contains already practiced exercises. It is therefore important to again ensure that only mandatory exercises are considered that have not yet been practiced. The

Algorithm 5.6: Optional exercise filters for all `realizations`

Require: exerciseCandidates

```
for exercise ∈ exerciseCandidates do
    for item ∈ exercise.items do
        if item ∉ learner.practicedItemSets then
            filteredItems ← filteredItems + item
        end if
    end for
    if |filteredItems| ≥ 5 then
        exercise.items ← filteredItems
        filteredItems ← [ ]
    end if

    for item ∈ exercise.items do
        if item ∉ learner.practicedItems then
            filteredItems ← filteredItems + item
        end if
    end for
    if |filteredItems| ≥ 5 then
        exercise.items ← filteredItems
        filteredItems ← [ ]
    end if

    for item ∈ exercise.items do
        if item.isMandatory = True then
            if item ∉ learner.practicedItems then
                filteredItems ← filteredItems + item
            end if
        end if
    end for
    if |filteredItems| ≥ 5 then
        exercise.items ← filteredItems
        filteredItems ← [ ]
    end if

    for item ∈ exercise.items do
        if ∃c ∈ item.constructs : c ∉ learner.practicedConstructs then
            filteredItems ← filteredItems + item
        end if
    end for
    if |filteredItems| ≥ 5 then
        exercise.items ← filteredItems
    end if
end for
```

final filter accounts for across-exercise variability. Regardless of whether the learner has already mastered the entire `realization`, we aim for high linguistic variability across all practiced exercises. To this purpose, the algorithm determines for each exercise item the number of contained `linguistic constructs` that the learner has practiced previously and the number of `linguistic constructs` the learner has not yet practiced. The algorithm here considers core and discriminative annotations of `linguistic constructs`, but not environment annotations. In order to determine those items that increase overall variability across all exercises, the algorithm filters out all items that practice only `linguistic constructs` that were already practiced before.

The remaining candidates all constitute suitable meta-exercises with items that may support the learner in their goal of mastering the `realization of a learning target`. In order to improve the learner's progress towards that goal, we next consider linguistic variability within each exercise with the aim of assembling an exercise whose variability best supports the learner.

5.3 Selecting and Configuring the Next Exercise

If the learner has already mastered all `properties` of the selected `realization` – and thus also the `realization` –, we aim for high linguistic variability within the practice exercise that we select in addition to high variability across all exercises. Otherwise, we aim for low within-exercise variability. This way, we account for our findings in Section 3.2 that learners of higher proficiency benefited from high variability only when the exercise context was similar to the practice conditions while learners of lower proficiency benefited from high variability only when the exercise context was different from the practice conditions.

When aiming for low within-exercise variability, the number of distinct `linguistic constructs` of the selected exercise should be low. The items within each meta-exercise are ranked by this number of distinct `linguistic constructs`.

When aiming for high variability, we consider both the `linguistic constructs` of the currently selected exercise and those of the previously practiced ones. The number of previously practiced `linguistic constructs` in the selected exercise should be low, whereas the number of distinct, not yet practiced `linguistic constructs`

should be high. Borrowing from evaluation metrics applied in machine learning, we calculate a variability score with the F_1-score formula given in Equation 5.6.

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{5.6}$$

We define TP as the `constructs` present in the item that were not practiced before, FP as the `constructs` present in the item that were practiced before, and FN as the `constructs` of the `realization` not present in the item that were not practiced before. Within each meta-exercise, the items are ranked according to this score.

Algorithm 5.7: Exercise item selection

```
exercise.items ← [ ]
while |exercise.items| < 5 do
    referenceItem ← metaExercise.items[0]
    exercise.items ← exercise.items + referenceItem
    metaExercise.items ← metaExercise.items − referenceItem
    if exercise.semanticDistractors then
        semanticDistractors ← [ ]
        for distractor ∈ item.semanticDistractors do
            if distractor ∈ exercise.semanticDistractors then
                semanticDistractors ← semanticDistractors + distractor
            end if
        end for
        exercise.semanticDistractors ← semanticDistractors
    else
        exercise.semanticDistractors ← item.semanticDistractors
    end if

    environmentAnnotations ← [ ]
    for c ∈ referenceItems.constructs do
        if c ∈ metaExercise.property.constructs then
            environmentAnnotations ← environmentAnnotations + c
        end if
    end for

    filteredItems ← [ ]
    APPLYFILTER(Algorithm 5.8)
    APPLYFILTER(Algorithm 5.9)
end while
```

In order to select 5 items of the meta-exercise and rank the entire exercise, Algorithm 5.7 is applied. It selects the first item (with the highest score) and adds it to the target exercise. It then uses this item as reference item and applies a number of filters to the remaining items. The first filter differs depending on whether the learner has mastered the `realization`. If this is not the case, Algorithm 5.8 is applied. Aiming for low within-exercise variability, it keeps only those items that practice the same `linguistic construct` instantiating the meta-exercise's `property` as the reference item.

Algorithm 5.8: Item filters for not mastered `realizations`

for $item \in metaExercise.items$ **do**
  if $\exists c \in item.constructs : c \in metaExercise.property.constructs \,\&\, c \in referenceItem.constructs$ **then**
    $filteredItems \leftarrow filteredItems + item$
  end if
  if $|filteredItems| \geq 5$ **then**
    $metaExercise.items \leftarrow filteredItems$
    $filteredItems \leftarrow [\,]$
  end if
end for

In the inverse case, Algorithm 5.9 is applied that aims for high variability. It keeps those items that practice another `linguistic construct` of the meta-exercise's `property` as the reference item.

Algorithm 5.9: Item filters for mastered `realizations`

for $item \in metaExercise.items$ **do**
  if $\exists c \in item.constructs : c \in metaExercise.property.constructs \,\&\, c \notin referenceItem.constructs$ **then**
    $filteredItems \leftarrow filteredItems + item$
  end if
  if $|filteredItems| \geq 5$ **then**
    $metaExercise.items \leftarrow filteredItems$
    $filteredItems \leftarrow [\,]$
  end if
end for

Regardless of mastery, the two filters represented in Algorithm 5.10 are applied to all items. If the `linguistic construct` of the reference item contains environment annotations, each iteration only considers exercise items whose core or discriminative

Algorithm 5.10: Item filters for all `realizations`

for $item \in metaExercise.items$ **do**
  if $|environmentAnnotations| > 0$ **then**
    if $\exists c \in item.constructs : \exists a \in environmentAnnotations : a \in (c.coreAnnotations + c.discriminativeAnnotations)$ **then**
      $filteredItems \leftarrow filteredItems + item$
    end if
  else
    $filteredItems \leftarrow filteredItems + item$
  end if
  if $|filteredItems| \geq 5$ **then**
    $metaExercise.items \leftarrow filteredItems$
    $filteredItems \leftarrow [\,]$
  end if
end for

for $item \in metaExercise.items$ **do**
  if $item.lemma \in exercise.semanticDistractors \;\&\; \forall addedItem \in exercise.items : addedItem.lemma \in item.semanticDistractors \;\&\; \exists w \in item.semanticDistractors : w \in exercise.semanticDistractors$ **then**
    $filteredItems \leftarrow filteredItems + item$
    $semanticDistractors \leftarrow [\,]$
    for $d \in item.semanticDistractors$ **do**
      if $d \in exercise.semanticDistractors$ **then**
        $semanticDistractors \leftarrow semanticDistractors + d$
      end if
    end for
    $exercise.semanticDistractors \leftarrow semanticDistractors$
  end if
  if $|filteredItems| \geq 5$ **then**
    $metaExercise.items \leftarrow filteredItems$
    $filteredItems \leftarrow [\,]$
  end if
end for

annotations contain at least one of these annotations. Since environment annotations target primarily *mixed* `realizations of learning targets`, this condition ensures that the compiled exercise contains diverse `linguistic constructs` of the `learning target` in focus. The second filter concerns bags of words with lemmas for Gap-filling exercises, which constitute global exercise elements. Each exercise

item maintains a list of semantic distractors that are suitable for its actionable elements. Only those items are considered (1) whose target answer lemmas are already contained in the bag of words of the target exercise and (2) that support at least one of the semantic distractors supported by all items already added to that exercise. This ensures that (1) the lemmas in the bag of words fit only into a single gap and that (2) the selected semantic distractors are compatible with all items of the exercise.

The order of the items that have passed a filter is not changed so that the ranking obtained before applying the filters is maintained. The first item of the remaining candidates is then added to the target exercise and the filtering is repeated with this item as the new reference item. The item selection algorithm within each meta-exercise terminates when five items have been selected.

The resulting list consists of exercises with five items each. In order to select the best exercise from this list, we aggregate the ranking scores of the five items of each exercise and select the one with the highest sum.

5.4 Summary

In this chapter, we presented an approach to consider linguistic variability, curricular constraints, and to some degree exercise difficulty when selecting exercises in a user-adaptive ILTS.

In order to decide when a KC has been mastered, the learner's progress for each KC must be tracked. We outlined how mastery learning can be applied to KCs defined by linguistic and `skill` dimensions. In this vein, we illustrated how to apply a bug model populated with the help of error diagnosis as the basis for remedial practice, and separate tracking of mastery based on exercise performance in a learner model.

We presented an exercise structure where meta-exercises group items practicing the same KC in order to facilitate learner-adaptive exercise selection. This selection in a first step identifies meta-exercises from the exercise pool that match the learner's stereotypes, are within the pedagogical scope, and are suitable to practice the KC in focus. In a second step, it assembles an exercise that best matches the learner's current need for more or less variability. The need for within-exercise variability

is lower when mastery has not yet been reached and higher for revision exercises. Across-exercise variability is high for all learners in order to foster varied learner productions. We thus outlined an implementation that considers curricular constraints, exercise difficulty at a minimal degree, and linguistic and `skill` variability.

Chapter 6

Exercise Generation to Provide Linguistically Varied Content

Parts of this chapter are based on Publications 1, 2, 4 and 6.

Macro-adaptive systems require huge amounts of practice material with slightly varying properties in order to cover the vast space of possible ability levels a student can have across a range of `linguistic constructs` (Katinskaia et al., 2018; Kerres et al., 2023). Diagnostic exercises need to be provided in addition. In order for them to be used by learners to demonstrate continued mastery and counteract forgetting, multiple such exercises should be available for each `realization of a learning target`. Completing the exact same diagnostic exercise might not yield valid representations of a learner's knowledge state as learners might perform better on repeated completions of the same exercise for which they receive binary feedback (Barkaoui, 2017). Instead, new exercises should be issued that target the same concepts. Manually compiling 'rich and meaningful' material supporting successful learning (Madlener, 2016) in all these variations is just as unfeasible as providing individualized human instruction to each learner. Automatic generation of practice material is therefore a necessary prerequisite for large-scale user-adaptive systems.

Many researchers have argued for grammar practice in context as opposed to mechanical drills in order to improve the perceived practice experience (Doughty and Long, 2003). While entirely open exercises are difficult to automatically generate and score, half-open and closed exercises lend themselves well to automatic generation

(Perez-Beltrachini et al., 2012). Such exercises come with additional advantages. While spontaneous speech productions retrieve the alternative from a collection of linguistic variations that the learner knows best, planned speech selects the best fitting alternative from the learner's systematic knowledge (Ellis, 1985). It is thus important to systematically practice less well known variants in more closed exercises, which elicit planned text production.

The task of automatic exercise generation has received considerable attention in the past two decades, with a particular focus on language exercises. In their literature review, Kurdi et al. (2020) attribute this to the increased need for exercises in this domain due to standardized tests, as well as to the relatively low complexity of exercise generation in the language domain – as opposed to content domains – as automatic generation of language exercises usually requires no or only shallow semantic processing.

Exercise generation can be considered in the wider context of question generation, although some aspects of exercise generation are particular to the intended usage as practice exercises. In the domain of language learning, there are three types of exercises that require slightly different processing (Kurdi et al., 2020): (1) Reading comprehension exercise generation typically employs machine learning to generate pairs of questions and target answers (Rathod et al., 2022). (2) For vocabulary exercises, most approaches first identify target words based on word frequency, learner interests and/or part-of-speech (POS) before generating distractors based on semantic relations (Knoop and Wilske, 2013). (3) Grammar exercise generation relies on syntactic and morphological processing in order to identify target forms (Perez-Beltrachini et al., 2012).

For these three types of language exercises, we acknowledge that vocabulary and reading comprehension exercises seem indeed rather straightforward since they focus on a single dimension of exercise properties each, namely vocabulary and semantics, respectively. For grammar exercises, however, the focus lies on morpho-syntactic properties. It would be possible to only consider this one dimension of grammar exercises. However, such a reductionist approach neglects important other dimensions. Quite notably, we have identified `skills` targeted by an exercise to be relevant in the context of varied practice as well in order for practice to be transfer-appropriate and less monotonous. Since different exercise types practice different `skills`, supporting

practice of different `skills` requires exercises of different types. In addition, the lexical and semantic dimensions of vocabulary and reading comprehension exercises are also relevant for grammar exercises. Neither lexis nor semantics constitute learning goals of grammar exercises. However, they open up opportunities to integrate grammar exercises into meaning-focused language instruction such as TSLT by aligning the semantic content of the exercises with the topic of the lesson (Rosell-Aguilar, 2007). These considerations leave us with three dimensions of exercise properties relevant to grammar exercise generation: semantic topic, morpho-syntax, and `skill`. An approach accounting for all three dimensions needs to systematically generate a variety of exercises from the same text basis, introduce manipulations to vary morpho-syntactic contexts, and finally transform them into exercises of different types to support various `skills`.

Previous implementations for grammar exercises do, however, not include such a comprehensive approach. Authentic ICALL focuses on using authentic texts in language learning (Meurers, 2012). In particular, automatically generating grammar exercises from authentic texts has received considerable attention in the past as a means to meet the demand for practice material in ILTSs (Malafeev, 2015). Closed activity types such as Multiple Choice are especially popular because they can be automatically scored given the very restricted space of possible learner answers (Tafazoli et al., 2019). Yet, supported exercise types vary from one system to the other. Since different learners struggle with different exercise types (Heift and Rimrott, 2012), it is important to offer practice for a `linguistic construct` in different exercise types. As outlined in Table 6.1, a number of tools integrate a variety of different types.

While these systems can generate multiple exercises for a linguistic form from the same source document, the actual number of exercises is usually quite limited. By varying exercise parameters such as the number of distractors, hints in parentheses, or the span of the target form, variability across the generated exercises can be increased. There are some notable examples that make use of such parametrizations. *MIRTO* provides parameters for the choice of target forms, parentheses with hints for the blanks of Fill-in-the-Blanks exercises, and support elements such as reference pages (Antoniadis et al., 2004). The assistant system *Sakumon* requires users to manually select target items and distractors from automatically generated

	Fill-in-the-Blanks	Mark-the-Words	Multiple Choice	Error Detection	Word Formation	Memory	Jumbled Sentence	Categorization	True/False	Tense transposition
MIRTO (Antoniadis et al., 2004)	X	X								
ArikIturri (Aldabe et al., 2006)	X		X	X	X					
FLAIR[1] extension (Heck and Meurers, 2022)	X	X	X			X	X	X		
Sakumon (Hoshino and Nakagawa, 2007)	X		X							
ClozeFox (Colpaert and Sevinc, 2010)	X		X							
View (Meurers et al., 2010; Reynolds et al., 2014)	X	X	X							
Language Exercise App (Pérez and Cuadros, 2017)	X		X				X			
Verb Tenses System (Ferreira and Pereira Jr., 2018)	X		X						X	X

Table 6.1: Exercise types in existing, grammar-focused ILTSs.
The supported exercise types (X) vary from one system to another.

suggestions (Hoshino and Nakagawa, 2007), thus allowing to manually manipulate these exercise parameters. In the *Language Exercise App*, target forms, distractors, and parentheses of Fill-in-the-Blanks exercises are parametrizable (Pérez and Cuadros, 2017). *FLAIR*'s exercise generation functionality provides parameters for target forms, distractors and parentheses, and in addition allows users to influence the specificity of the exercise instructions (Heck and Meurers, 2022). However, these systems require exercise creators to specify the parameter configuration individually for each exercise, so that generating large numbers of parametrized exercises involves considerable configuration effort. Moreover, none of them allows to vary linguistic contexts.

Support for varied morpho-syntactic contexts is even more limited among these systems. In fact, while Baptista et al. (2016) for instance acknowledge very fine-grained

distinctions between five different passive sentences, their *REAP.PT* tool only supports a single one of them. For exercises with a more semantic focus, some tools annotate a variety of linguistic forms in the text selected or provided by the exercise creator, and transform the forms into exercise targets. *AGReE*, for example, generates a Single Choice exercise for each sentence with one of its supported linguistic forms (Chan et al., 2022). Yet these different forms are unrelated and do not constitute morpho-syntactic variations of the same `realization of a learning target`. To find texts containing specific morpho-syntactic variants of the `learning target` in focus, the systems select sentences for exercises based on whether they contain a target form corresponding to one of the targeted linguistic constructions. They do not consider additional morpho-syntactic properties such as, for instance, the clause order in conditional sentences. Controlling linguistic forms at this level is, however, necessary to support macro-adaptivity focusing on linguistic variability as described in Chapter 5. *Sakumon* (Hoshino and Nakagawa, 2007) and the extension to *FLAIR* (Heck and Meurers, 2022) integrate text filtering mechanisms based on contained linguistic forms, which in principle offers the means to find texts for exercise generation with specific linguistic properties also beyond the forms targeted by the exercise. However, the linguistic annotations provided in the two systems are not fine-grained enough to allow filtering for specific morpho-syntactic properties.

Not taking authentic texts as a starting point but instead relying on Natural Language Generation (NLG), *I-FLEG* does in principle support this kind of variety. Integrated into a serious game, the exercise generation is parametrizable in terms of target form, semantic content, and morpho-syntactic properties. The exercise generation algorithm retrieves facts from a knowledge base that match the semantic content and target form properties required by the adaptivity algorithm, creates sentences from them that comply with the required morpho-syntactic properties and generates exercises of four different types from the sentences (Amoia et al., 2012). These exercise types include Fill-in-the-Blanks, Name the object, Shuffle, and Transposition exercises. While it would be possible to generate exercises for any conceivable parameter instantiation and combination, the system instantiates the exercise parameters based on the learner's interactions with the tool. When triggered by a learner interaction such as clicking on an object in the 3D-world, the exercise generation algorithm dynamically only generates a single exercise matching the exercise type and linguistic forms that the macro-adaptivity algorithm currently

judges most suitable for the learner.

Most approaches to dynamic exercise generation take multiple learner factors such as their domain knowledge into account and generate exercises ad-hoc (Cui and Sachan, 2023). This has the advantage that the generated exercises can specifically target a learner's misconceptions (Scheutz et al., 2005) and that all exercises are directly relevant to at least one learner. If the same parameters become suitable for another learner, systems that support a store of generated exercises do not need to re-generate the same exercise but instead retrieve the exercise from the store. An issue with dynamic exercise generation consists in the absence of quality control. Automatic grammar exercise generation relies on NLP to extract linguistic features and manipulate certain linguistic constructions. NLP is, by nature, prone to errors (Danenas and Skersys, 2022). Automatically generated language exercises, like those for any other domain, should therefore be checked for correctness and pedagogic validity before being presented to learners (Slavuj et al., 2022). This is particularly relevant for diagnostic exercises, as they are used for assessment purposes. The rise of ChatGPT has seen different levels of trust in the tool's output with two opposing user personas at its extremes (Al-Mughairi and Bhaskar, 2024): (1) Those who do not trust Artificial Intelligence (AI) at all, and (2) those who believe the tool blindly. It is therefore important to on the one hand allow the first user persona type to double-check and, if necessary, correct AI-generated output, and on the other hand make the second type aware of the need to revise the exercises in order to make sure that learners are exposed to correct and pedagogically valid material. While revising exercises is much less time-consuming than generating them, this task still constitutes considerable human effort. It thus needs to be made as efficient as possible.

In this chapter, we first address RQ 9 by presenting an approach to efficient generation of systematically varied exercises. We then turn to RQ 10 and introduce applications of learning analytics to continuously improve exercise generation.

Research Questions

RQ9 How can exercise generation efficiently support the creation of valid and linguistically varied exercises?

RQ10 How can learning analytics support continuous improvement of exercise generation?

6.1 Efficient Review of Generated Exercises

Many exercise generation tools provide support to post-edit the generated exercises, either within the tool (e.g., Toole and Heift, 2001; Hoshino and Nakagawa, 2007) or through an interface to general-purpose authoring tools such as Hot Potatoes[2] or the learning management system Moodle[3] (e.g., Bick, 2000; Aldabe et al., 2006; Pérez and Cuadros, 2017). These interfaces are, however, designed to edit a single exercise at a time. Modifications of elements that affect multiple exercises generated from the same document thus have to be performed on each exercise individually. For systematically varied exercises, the revision progress can be made more efficient by integrating it into the exercise generation workflow on more abstract exercise representations. This minimizes the number of items to review. RQ 9 investigates how such an approach can look.

Kurdi et al. (2020) outline three steps that are common to most exercise generation approaches: (1) pre-processing, (2) question construction, and (3) post-processing. By making this three-step approach transparent to exercise creators and allowing them to perform corrections after each step, errors in the output of automatic processing can be corrected early on before they are propagated into the next generation step. This avoids having to perform modifications on a series of duplicates of the input since any errors from a previous step are passed on to multiple derivatives of the generated exercise material.

Our approach focuses on a selection of morpho-syntactic constructs and `skills` that are relevant to beginner to intermediate learners of English, i.e., conditionals and relative clauses. The current implementation covers relative clauses with a focus on pronoun usage or contact clauses and conditionals practicing verb forms of a given

[2]Hot Potatoes is available at https://hotpot.uvic.ca.
[3]Moodle is available at https://moodle.org.

conditional type or the differentiation between conditional types. Despite this rather narrow focus, the general procedure can easily be applied not only to additional `learning targets`, but also to other proficiency levels and languages. The specific implementation of the three steps is, however, proper to each `realization of a learning target`.

Figure 6.1 shows how we apply Kurdi et al.'s (2020) systematicity of pre-processing, question construction, and post-processing to grammar exercise generation. In each step, we account for one of the three dimensions relevant to grammar exercises. The pre-processing step targets the provision of sentences from which exercises can be generated with a focus on varied linguistic contexts. The question construction step

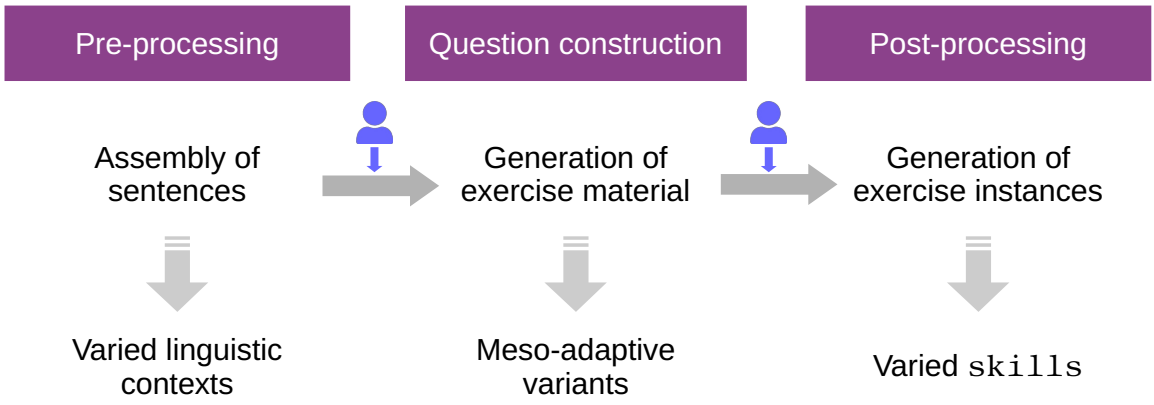


Figure 6.1: 3-step question generation for high-variability grammar exercises. Pre-processing assembles sentences accounting for varied linguistic contexts. Question construction generates exercise material accounting for meso-adaptive variants. Post-processing generates exercise instances accounting for varied `skills`.

As illustrated by the system architecture design in Figure 6.2, we additionally introduce opportunities for exercise creators to revise the output of the previous step after each processing step for more efficient quality assurance. This allows exercise creators to modify aspects of the exercises on the most general layer of abstraction that contains the information to edit.

Since we do not display the final exercises to exercise creators but instead show them more abstract representations at interim steps, there are no readily available

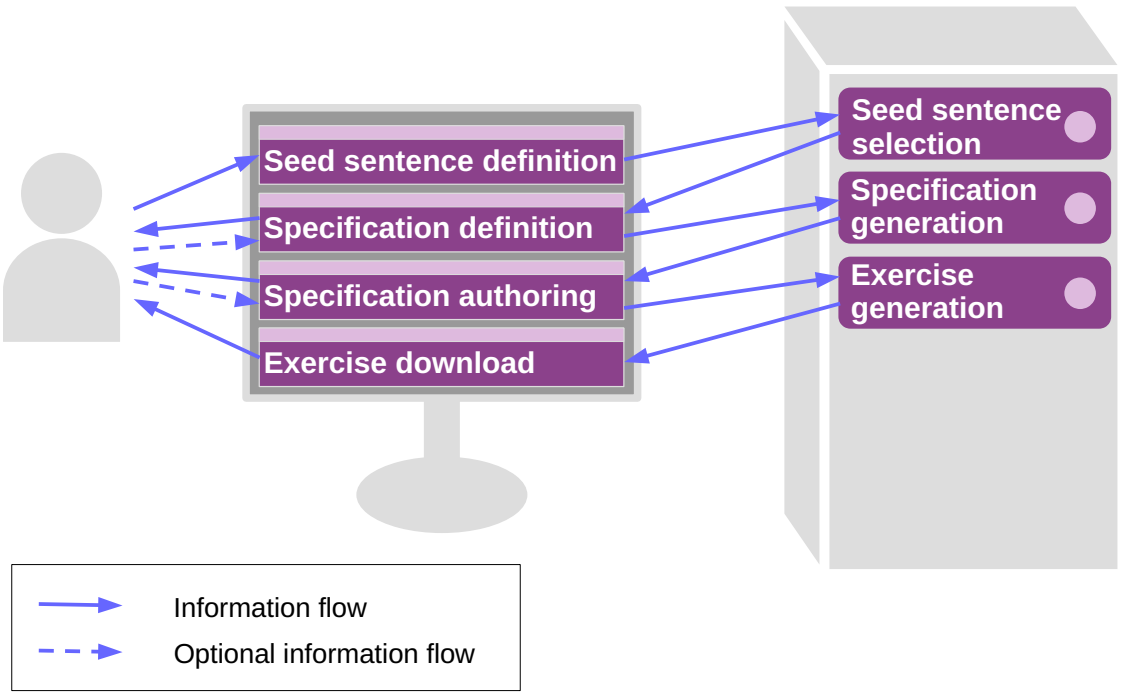


Figure 6.2: Exercise generation architecture.

The user interacts with the front-end, which in turn communicates with the back-end. In the web GUI, exercise creators can define seed sentences, which the algorithm selects based on the definition. Based on a user's exercise specification definitions, the algorithm generates exercise specifications that users can edit in the authoring web interface. The algorithm then transforms the specifications into downloadable exercises.

authoring tools. We therefore implemented a prototypical web-based user interface based on Hypertext Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript, relying on Asynchronous JavaScript and XML (AJAX) for communication with the server. The user interface is used by exercise creators such as pedagogically trained content creators, system administrators, or teachers.

The back-end code is implemented in a micro-service architecture, which supports flexible use of programming languages, thus facilitating the use of best-performing libraries across multiple programming languages.

In the following, we describe the three steps of exercise generation in more detail. We first present how the algorithm selects seed sentences based on definitions that exercise creators configure in the web interface. We then outline how exercise creators can define abstract exercise specifications that the algorithm generates subsequently.

Finally, we introduce the exercise authoring interface designed to edit abstract exercise specifications and describe the generation of downloadable exercise instances from the specifications.

6.1.1 Seed Sentence Selection

Linguistic variability can be considered in exercise generation on two levels. On the first level, categorical differences are proper to the so-called seed sentences. Also referred to as *carrier sentences* or *candidate sentences* in the literature, they constitute natural language sentences from which exercises are generated (Pilán et al., 2017). On the second level, generic variability can be introduced through manipulations of the seed sentence.

The extraction of seed sentences thus provides the first opportunity in the exercise generation workflow to account for varied linguistic contexts in the generated exercises. Language curricula dictate which categorical linguistic variations are relevant. Sentences can either be extracted from a corpus or generated by means of NLG tools (Perez-Beltrachini et al., 2012). Corpus-based sentences have the advantage of offering authentic language material and being contextualized, which has been argued to have positive effects on motivation (Peacock, 1997). Working with corpus texts in general supports two approaches to exercise generation. The first approach annotates all occurrences of the targeted `linguistic construct` in a text (Heck and Meurers, 2022; Meurers et al., 2010). The entire text is then used in the exercise. Contextualized practice, where the focus is more on transmitting specific content rather than on practicing a certain grammar topic, is valuable especially for advanced learners. In addition, it is highly relevant in the context of TBLT (Zúñiga, 2016), which currently is the predominant approach to language instruction (Müller-Hartmann and Schocker-von Ditfurth, 2013). The second approach extracts only single sentences, potentially with some surrounding co-text for basic contextualization, for a specific `learning target` (Volodina et al., 2014). An exercise is then compiled from either a single sentence or multiple, de-coupled sentences that all contain instantiations of the same `learning target`. Accounting for linguistic context becomes harder the more exercise items are considered simultaneously. Working with entire texts is therefore less suitable for the generation of exercise for adaptive systems with a focus on linguistic variability. We therefore

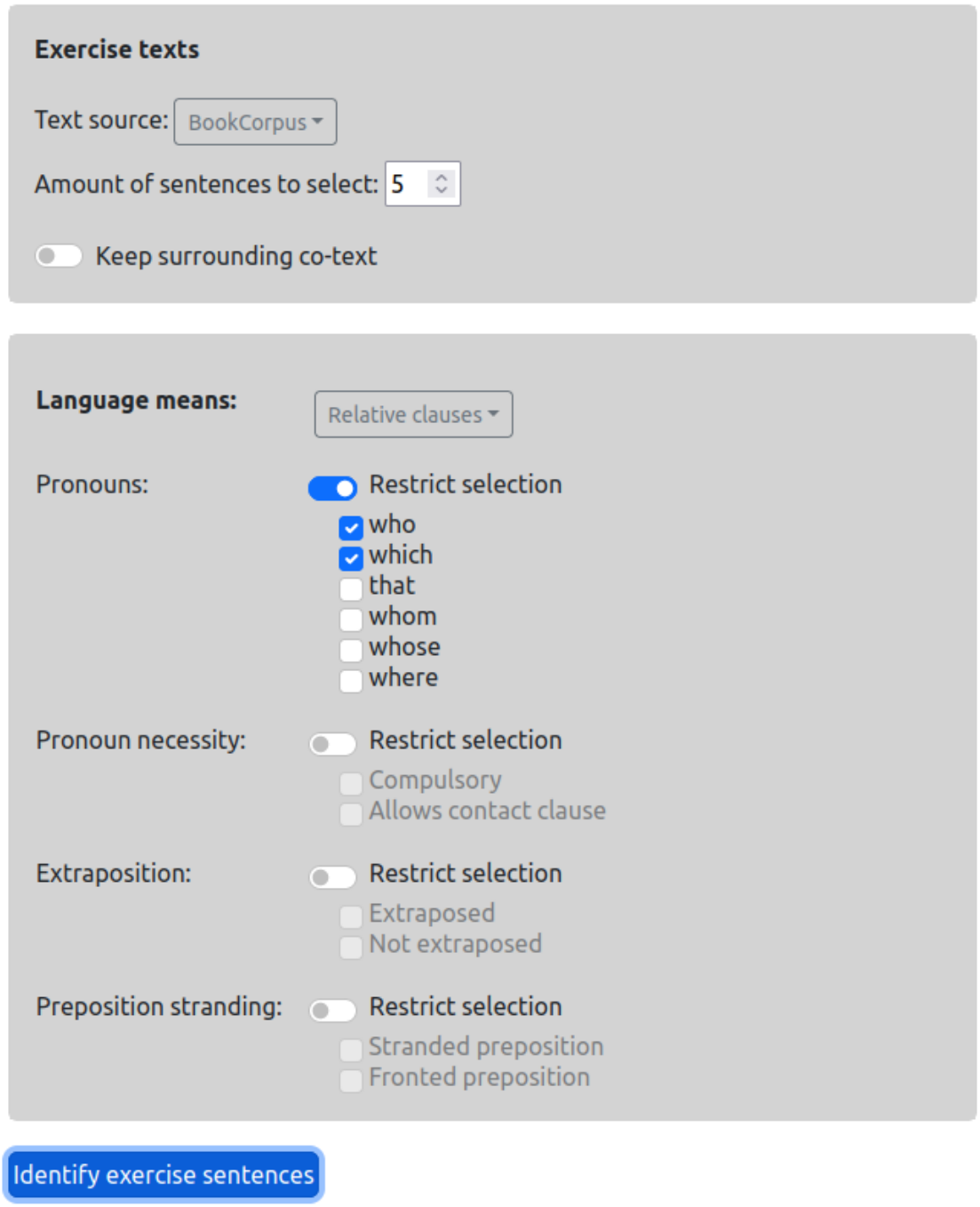


Figure 6.3: Seed sentence definition GUI.

The configuration allows to specify general parameters such as the text source, and parameters specific to the `learning target` that the exercises should practice.

focus on extracting individual sentences.

User interface In our implementation, the selection of suitable sentences starts in the web interface shown in Figure 6.3. It supports four input sources: (1) the BookCorpus[4], (2) the web, (3) NLG, and (4) custom texts. If users want to search the **BookCorpus** for candidate sentences, they need to specify the desired number of sentences. The user interface allows users to restrict the selection of sentences based on properties of the `learning target` in focus and the linguistic context. Since the space of possible parameter combinations in the configuration grows exponentially with the number of parameters, the number of seed sentences to select can only be specified globally and not for specific parameter constellations. Crawling the **Web** in addition allows without further implementation effort to search for sentences that appear in a defined semantic context so that users also need to specify a search term. In addition, we support artificially **generated** sentences for a topic specified by the user. The advantage of this approach is that it allows to control for linguistic properties and the vocabulary of the generated sentences (Perez-Beltrachini et al., 2012). Linguistic properties of the target and co-text forms in these sentences can be specified by the user and are taken into consideration by the NLG implementation. Although it is in principle possible to restrict or else use specific vocabulary (Koraishi, 2023), the current implementation only allows to specify a semantic topic. Besides extracting or generating sentences, users can also upload their **own sentences**. Custom texts must be inserted into the provided input field. They can consist of manually compiled texts or any text copied from an arbitrary source. With this input source, users are responsible themselves for providing appropriate linguistic contexts.

An additional configuration parameter in the GUI determines whether some co-text is extracted from the BookCorpus or the Web along with the seed sentences or only the seed sentences themselves. If the co-text option is activated, the text of the paragraph from which the target sentence is extracted, delimited by line breaks, will be extracted as well.

A final set of configuration parameters allows users to filter the selection of seed sentences that will later be turned into exercise items. Available parameters depend

[4]The corpus based on an implementation by Presser (2020) is available at https://the-eye.eu/public/AI/pile_preliminary_components/books1.tar.gz.

on the `learning target` in focus. For conditionals, they include the conditional type, clause order, polarity, aspect, and sentence form. For relative clauses, the parameters consist of the relative pronoun, whether the pronoun is compulsory or can be left out, extraposition, and preposition stranding.

Algorithm The seed sentence selection algorithm differs from one input source to another. For **corpus texts**, the documents of the BookCorpus are searched until the required number of sentences has been identified. For **web texts**, a Google search is performed instead for the search term and the content of the search results is processed, again until the desired number of seed sentences has been extracted. The Google search results are not further filtered or sorted in the current implementation, rendering the revision step all the more important. Instead of pursuing a rule-based approach for automatic seed sentence **generation** as traditionally done for grammar exercises (Perez-Beltrachini et al., 2012; Almeida et al., 2017), our implementation relies on the GPT-3[5] Python interface. This avoids the cumbersome task of implementing generation rules and makes the approach more transferable to other `learning targets`. The algorithm passes the specified topic on to the GPT-3 interface and obtains the desired number of generated sentences. Since the GPT-3 models perform better in terms of semantics than with respect to grammar on a meta-linguistic level, however, control over grammatical properties of the generated sentences is limited (Brown et al., 2020). **Custom texts** are processed in their entirety in order to identify seed sentences containing `linguistic constructs` of the `learning target` in focus. In order to identify suitable sentences, the text candidates are processed with NLP tools. To this purpose, the Java library Stanford CoreNLP[6], as well as the Python libraries NLTK[7], SpaCy[8], and Stanza[9] were tested. Table 6.2 summarizes the results of the evaluation of their reliability with respect to identifying sentences containing conditional and relative clause constructions. SpaCy and Stanza performed similarly, with SpaCy being considerably faster. Subsequent NLP analyses were therefore implemented based on SpaCy's sentence segmenter, tokenizer, POS tagger, lemmatiser, constituency parser, and dependency

[5]GPT-3 is available at https://github.com/openai/openai-python.
[6]Stanford CoreNLP is available at https://stanfordnlp.github.io/CoreNLP.
[7]NLTK is available at https://www.nltk.org.
[8]SpaCy is available at https://spacy.io.
[9]Stanza is available at https://stanfordnlp.github.io/stanza.

	Precision		Recall	
	RC	C	RC	C
NLTK	.8900	.7000	.7417	.9610
Stanza	.9800	.8100	.8976	.9927
SpaCy	.9400	.8600	.9039	.9902
Stanford CoreNLP	.9600	.7600	.7606	.9683
Number of sentences	100	100	635	410

Table 6.2: Evaluation of NLP libraries.

Precision was computed for a random sample of 100 sentences from the BookCorpus. Recall values were determined for a collection of manually compiled example sentences. All metrics were determined for relative clauses (RC) and conditionals (C).

parser.

The algorithm processes the texts of all input sources in the same manner: A naive form identification rule based on dependency parsing determines whether a sentence could be a potential candidate. For conditionals, it searches for adverb clauses with some additional conditions such as the existence of a token with value *if* and the absence of a verb token contained in a manually compiled list of reported speech markers[10]. For relative clauses, the algorithm searches for relative clauses with a wh-pronoun.

This rough filtering results in a considerable amount of noise in the sentence candidates. Pilán et al. (2017) identify a number of criteria for good seed sentences, including well-formedness, context independence, linguistic complexity, and additional structural and lexical criteria. While, with the parameters that exercise creators may configure, we address most of the structural criteria such as negated or interrogative contexts, we deliberately do not restrict seed sentence selection based on lexical criteria, whose optimal configuration is often user-dependent and better controlled by a macro-adaptivity algorithm in the target ILTS (Gooding and Tragut, 2022). Compliance with context independence will be more likely when the co-text option is activated so that adjoining sentences on which the seed sentence depends are included in the exercise's co-text. Context independence can in addition be ensured manually in the subsequent revision step. In order to account for well-formedness and linguistic complexity, we apply further processing after the naive sentence selec-

[10] Available lists (e.g., Tham and Nhi, 2021; Yilmaz and Özdem Erturk, 2017) contain predominantly affirmative markers. Since only question markers are relevant to confusions with conditional clauses, we compiled a list based on sampled evidence from the BookCorpus.

tion: The sophisticated algorithm extracts all the information relevant to exercise generation. This encompasses the exercise targets and their linguistic properties as well as parameters of the sentences that were configured in the GUI. As soon as one piece of information cannot be extracted or if it does not comply with the configured parameters, it is rejected. This not only ensures the highest possible success rate for exercise generation in the succeeding step, but also filters out most sentences that passed the naive filter but do not actually contain a `linguistic construct` of the the `learning target` in focus. In addition, we hypothesize that the NLP tools' inability to correctly process a sentence would reflect a beginner student's inability to do so. The naive filter thus also eliminates sentences that are too complex for our target group.

The successfully parsed sentences are stored in a result list. If specified by the user, the co-text of the paragraph is also stored in that list as individual elements. Each element in the result list is tagged with its type of either co-text or exercise item. The order of the list elements corresponds to the order in which they appeared in the text from which they were extracted. For seed sentences targeting conditionals, additional filtering is applied if the user has restricted the selection of the conditional type and selected *type 1* and *type 2*. Since a configuration including two conditional types is usually used for exercises targeting the distinction between these two types, the seed sentence selection ensures that both conditional types occur in roughly equal numbers in the result list. If the result list already contains enough seed sentences for one conditional type, any subsequently found occurrences of that type will therefore be treated like sentences with no conditional form. Similarly, a subtopic for relative clauses targets contact clauses, for which students need to learn when the pronoun can be left out. It is therefore important to have seed sentences both with optional and with compulsory relative pronoun. If a user activates the selection restriction targeting pronoun necessity and selects both values in terms of *compulsory* and *optional* pronouns, the algorithm therefore makes sure that sentences with compulsory and sentences with optional pronoun occur with similar frequency in the results.

Evaluation We evaluated precision and recall of candidate sentence selection for corpus texts and for manually compiled texts. To this purpose, we used the naive sentence selection algorithm to extract 100 conditional sentences and 100 relative

sentences from the BookCorpus. The selection was not further restricted. We manually verified whether the extracted sentences were indeed instances of the targeted linguistic form and computed precision values. Since there are no gold standard annotations in the BookCorpus, we did not determine recall based on the BookCorpus. Instead, we manually composed 100 sentences for each of the two `learning targets` and applied the naive selection algorithm to those. We computed recall values for the algorithm's acceptance of the input as seed sentences. We then determined which of the sentences extracted from the BookCorpus with the naive selection were rejected by the sophisticated sentence selection algorithm and computed recall and precision values.

The results are summarized in Table 6.3. For seed sentence selection from the corpus, the naive selection extracted relative clauses with a precision of .9300. The precision of the sophisticated selection, which is also the precision of the overall seed sentence selection, was slightly lower at .8947. The decrease in precision was due to the high number of false negatives, which also resulted in a low recall of .3656. Since a number of sentences that were correctly identified by the naive selection got rejected by the sophisticated algorithm, the percentage of accepted incorrect findings increased relative to the overall number of accepted sentences. The poor recall values indicate that a substantial number of potential seed sentences was lost in the process, which might suggest that the additional filtering should be removed. However, the filtering also serves as a pre-selection to the ensuing exercise generation from the specifications, thus rejecting sentences early on that NLP tools cannot process successfully. Poor recall thus constitutes an accepted shortcoming when parsing large corpora that contain many occurrences of the target phenomena.

	Relative Clauses	**Conditionals**	**Corpus**
Recall (N)	.9100	1.0	M
Precision (N)	.9300	.8900	BC
Recall (S)	.3656	.7528	BC + N
Precision (S)	.8947	.9306	BC + N

Table 6.3: Evaluation of the seed sentence selection.

Recall and precision were calculated for the naive (N) and for the sophisticated (S) algorithm. *Precision (S)* also represents the overall precision of the seed sentence selection. Recall of the naive algorithm was calculated on manually compiled texts (M), the three other metrics on the BookCorpus (BC). For each metric, samples of 100 sentences were considered both for *Relative clauses* and *Conditionals*.

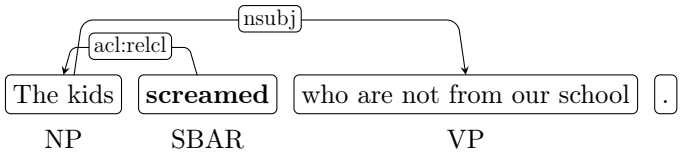
Extracting the required number of seed sentences is unproblematic in large corpora even if the selection restrictions are stricter, although the process may take longer. Considering the trade-off between fast performance of the algorithm and finding sentences that lend themselves well to exercise generation, we put a focus on the latter criterion. However, for custom texts, where the target phenomena occur in limited numbers, we do not apply the sophisticated selection algorithm in order to maximize the number of target forms in the exercise.

Results for conditionals are more in line with the expected behavior in that precision was higher with the sophisticated than with the naive algorithm. Precision on the corpus was already high (.8900) for the naive sentence selection and increased further to .9306 with the sophisticated sentence selection. Recall of the sophisticated selection was also considerably higher than for relative clauses (.7528). Of the 89 sentences that the sophisticated algorithm accepted as conditional sentences, 44 were actually not stereotypical conditional sentences taught in introductory language classes. They deviated in tense such as using simple present instead of will-future in the main clause (Example 6.1a) or in sentence structure such as using elliptical conditionals (Example 6.1b). This highlights the relevance of parameters for restricting the selection of seed sentences allowing users to only select sentences with textbook properties.

6.1 a. If I can't spoil my only daughter on her birthday, I'**m** not much of a father, now am I?
b. **What if** someone sees us?

On the manually compiled sentences, the naive algorithm achieved recall values of 1.0 for conditionals and .9200 for relative clauses. The recall values indicate that all conditional sentences were recognized by the algorithm whereas some relative clauses were rejected. These constituted either extraposed relative clauses such as Example 6.2a or sentences with the pronoun *whom* as in Example 6.2b. We traced the issues back to incorrect parsing outputs obtained from the employed NLP tools.

6.2 a.



b. My parents called my teacher **whom** I saw today.

Applying the sophisticated in addition to the naive selection algorithm thus successfully eliminates false positives of complex `learning targets` such as conditionals, and it reduces issues in subsequent NLP processing steps for complex as well as for more simple `learning targets` such as relative clauses. Allowing users to specify selection restrictions can be crucial for the usability of the tool in classroom instruction to support the identification of pedagogically suitable sentences.

Result review The result list generated by the algorithm is used on the client device to populate the upper part of the user interface for the configuration of exercise specification parameters shown in Figure 6.4.

It initially contains the exercise and co-text items extracted by the seed sentence selection algorithm. The extracted seed sentences can be edited, deleted, or their type changed from co-text to exercise item or vice versa. Additional items can be added manually. The order of all items can be changed through drag and drop operations.

6.1.2 Exercise Specification Generation

Once the seed sentences have been selected, they can be processed into exercises. In this step, the second level of linguistic variability, generic variability, can be introduced into the exercises. Contrary to categorical differences, variability at this level is not subject to a curriculum. Instead, linguistic variations can freely be generated that are relevant to learners in order to improve proceduralization and the development of more varied language.

In order to make exercise generation more efficient, Gierl et al. (2012) present an approach to generate large numbers of Multiple Choice exercises for the medical domain from a single specification. They establish a cognitive model that represents

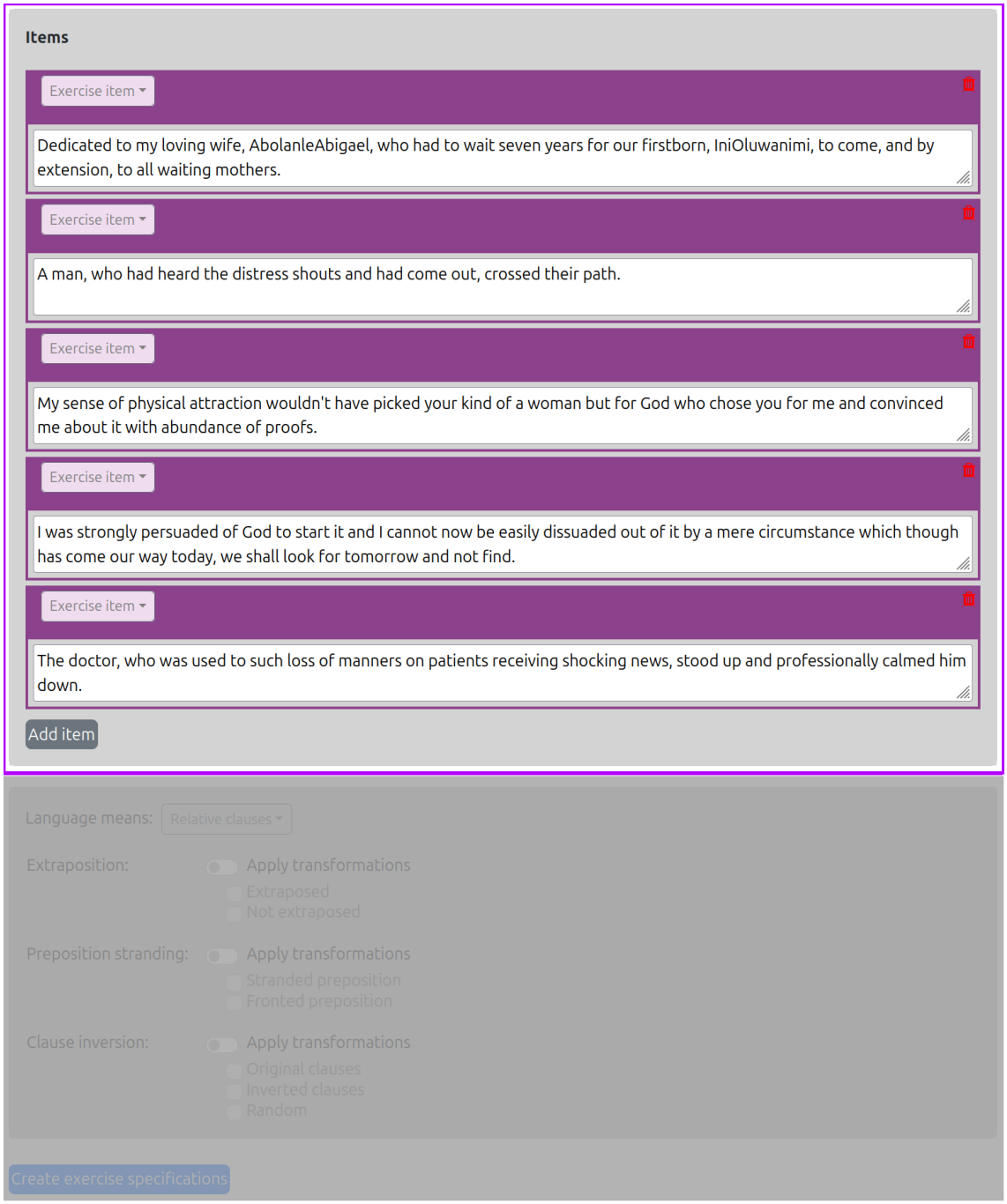


Figure 6.4: Seed sentence editor in the exercise specification definition GUI. The upper part of the GUI comprises an editor for the selected seed sentences.

the knowledge and skill of the domain in a hierarchical structure of concepts with features. These features can have varied instantiations. From the cognitive model,

Gierl et al. (2012) then instantiate item models, each of which contains all necessary information to generate items from it, including information on which features can be varied. Items for all different feature variations of an item model are then automatically generated.

Applied to our approach, we define a cognitive model in the form of exercise specification templates. Each template defines what information is required to generate multiple exercise instances. Exercise specifications – our item models – are then created by populating these templates with content. An exercise specification contains all material relevant to the exercises, such as distractors for Single Choice exercises, parentheses for Fill-in-the-Blanks exercises, and chunks for Jumbled Sentence exercises. Knowledge and skills that can be varied correspond to exercise parameters such as alternative clause orders and the exercise type in our approach. This manipulation of the initial seed sentences provides an additional means for the introduction of varied linguistic contexts. While it usually makes sense to vary contextual properties through either seed sentence selection or sentence manipulations, both strategies can also be applied simultaneously.

User interface If the list of seed sentences does not contain any co-text items, users can set additional parameters that will lead to the creation of linguistic transformations of the seed sentences. As shown in Figure 6.5, these parameters can be configured in the lower part of the exercise specification definition GUI. Whether and which context variants are generated thus depends on how the user configures the parameters and on the `learning target` for which exercises are being configured. Variations target predominantly parameters that are characteristic of developmental sequences discussed in the SLA literature (e.g., Pienemann et al., 1988), such as negation or interrogative structures. Transformations for conditional sentences thus comprise the conditional type (Example 6.3a), polarity (Example 6.3b), sentence form (Example 6.3c), and clause order (Example 6.3d).

6.3 He will win the race if he runs faster.

a. He would win the race if he ran faster.
b. He won't win the race if he runs faster.
c. Will he win the race if he runs faster?
d. If he runs faster, he will win the race.

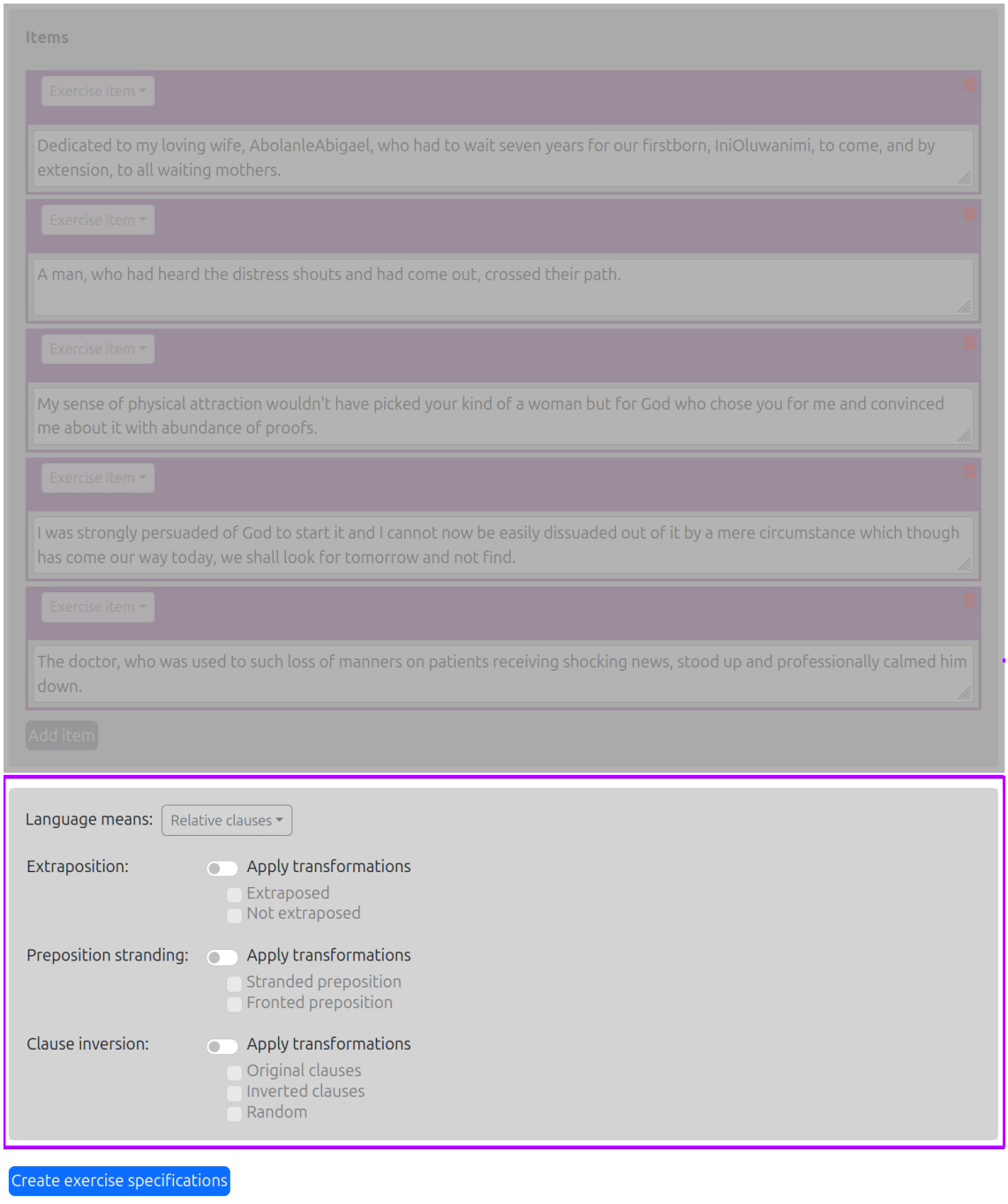


Figure 6.5: Parameter configuration in the exercise specification definition GUI.

The lower part of the GUI comprises a configuration interface specific to the `learning target` to specify desired linguistic transformations.

For relative clauses, preposition stranding (Example 6.4a), extraposition (Example 6.4b), and clause inversion (Example 6.4c) are supported. The latter parameter

transforms the original relative clause into a main clause and the original main clause into a relative clause, if possible.

6.4 The exam for which he studied was hard.

a. The exam which he studied for was hard.
b. The exam was hard for which he studied.
c. He studied for the exam which was hard.

Algorithm Based on the exercise specification configurations, the algorithm processes the texts declared as exercise items while keeping the co-text elements unchanged. Since it has been established in the previous step that the processed sentences must contain an occurrence of the targeted `linguistic constructs`, the algorithm this time does not reject sentences for which not all features corresponding to the configuration parameters can be extracted but instead uses default values in this case. Both steps rely on the same NLP pipeline so that they can use the same code.

The extracted features are used to generate abstract exercise specifications. Whether a transformation results in a separate exercise specification or merely in an alternative sentence of the same specification depends on whether the target tokens, i.e., the pronoun of a relative clause or the verbs of conditional sentences, are affected by the transformation. For example, negating the main clause of the conditional sentence given in Example 6.5a changes the verb from *will go* to *will not go* in Example 6.5b, thus requiring a new specification. Reversing the clause order (Example 6.5c) does not affect the verb form, therefore resulting in alternative sentences of the same specification. All transformations that result in a separate specification also offer the option in the configuration GUI to apply either one of the two realizations. In this case, the exercise specification generation algorithm randomly applies one of the realizations of the transformation to each exercise item in the list of seed sentences while at the same time making sure that each realization is applied approximately the same number of times. This allows to generate exercises that practice a variety of `properties` and `linguistic constructs` of a `learning target`.

6.5 a. If he **gets** better, he **will go** to school.
b. If he **gets** better, he **will not go** to school.

c. He **will go** to school if he **gets** better.

In order to support a range of exercise types encompassing Fill-in-the-Blanks, Single Choice, Mapping, Jumbled Sentence, Short Answer, Mark-the-Words, and Categorization exercises, the specifications are enriched with exercise elements such as distractors for Single Choice exercises and parentheses for Fill-in-the-Blanks exercises. Generating exercise elements mostly relies on linguistic annotations obtained from NLP, with distractor generation in addition requiring NLG functionalities. We use the Java-based SimpleNlg[11] library to this purpose, which can be easily integrated into our Python project thanks to the micro-service architecture.

We make one important distinction regarding distractors. In addition to grammatical distractors, semantic distractors can on top increase exercise difficulty and introduce an incidental focus on form. They also put more focus on meaning, which makes the exercises more suitable for TBLT settings. Considering that research on vocabulary learning provides a range of approaches to select valid semantic distractors (e.g., Zesch and Melamud, 2014), we will not explore this research area further. Contrary to common practice in distractor selection (Zesch and Melamud, 2014), we suggest to not attempt to identify a limited number of best suited distractors, but to instead store all suitable candidates with the respective exercise item. This way, exercise generation can support semantic distractors for exercise-global bags of words containing target forms or lemmas, as described in Section 5.3. In order to increase the overlap of global distractors for all items of an exercise so as to make them suitable for distractors in a global bag of words, we suggest to manually precompile lists of distractor candidates for each `learning target` that are suitable for the target group. For *questions*, for example, they constitute a list of question words. For *tenses*, the list contains verbs that are known to learners for whom the exercises are intended. Storing the distractor candidates in the order of their suitability makes it possible to use the highest-ranked semantic distractor in the local scope of an actionable element, specifying it in parentheses behind a blank along with the lemma of the target answer.

The exercise specifications for different `realizations` of the `learning target` *Relative clauses* differ slightly from each other due to differing requirements of the

[11] SimpleNlg is available at http://github.com/simplenlg/simplenlg.

resulting exercises. The specifications for the `learning target` *Conditionals* can in principle be used for multiple `realizations` if the grouping of specification items is altered in such a way that all items of a group belong to the same (for the `realizations` *Conditional type 1* and *Conditional type 2*) or to different (for the `realization` *Conditional type 1 vs. type 2*) conditional types.

In order to support the kind of meta-exercises used for an adaptivity algorithm such as presented in Section 5.2 as well as variability-focused exercise selection, we in addition determine the `linguistic constructs` and `properties` of each exercise specification item. The `linguistic constructs` are determined as described in Section 3.3.1 based on core, discriminative, and environment annotations: The linguistic annotations obtained from the employed POLKE micro-service are refined to support the domain model `constructs` using an additional NLP layer. The `property` that a specification item practices depends on these `linguistic constructs` of the actionable elements. If all core annotations relevant for a `property` are present in an exercise specification item, the `property` is annotated.

Evaluation We manually considered the suitability of the generated exercise specifications with the aim of identifying weaknesses of the implementation. Limitations of the current approach include the handling of discourse phenomena such as coreference and cohesion. Example 6.6 illustrates that for conditional sentences, the semantic validity is challenged if the first clause contains pronouns referring to an antecedent in the second clause. Also related to clause order, Example 6.7 showcases how cohesion can become an issue with de-contextualized relative clauses if the clauses of the relative sentence extracted from the text have a natural order or a causal relationship, like in Example 6.8, that is violated by the transformed relative sentence. Further issues were identified in the generated main clauses used as exercise prompts. Although not incorrect, Example 6.9a presents a generated prompt for relative clauses that is rather questionable in style. Since it does not consider coreference, the algorithm uses the same subject form for both prompt sentences instead of using a pronoun in the second sentence, as outlined in Example 6.9b.

6.6 a. If Tommy found his glasses, he would wear them.
b. *He would wear them if Tommy found his glasses.
b' Tommy would wear his glasses if he found them.

6.7 a. The boy who came in greeted everyone.
b. *The boy who greeted everyone came in.

6.8 a. Students who are faced with debt leave college.
b. *Students who leave college are faced with debt.

6.9 a. The boy came in. The boy greeted everyone.
b. *The boy came in. He greeted everyone.*

We therefore integrated further NLP analyses to overcome these limitations. For conditionals, we added a coreference annotator to the NLP pipeline and now extract all pronominal coreferences with their referent from the annotator's results. This allows us to identify whether a seed sentence uses a pronoun to refer to an antecedent, and to consequently interchange pronoun and referent when reversing the clause order. Thus, sentences like Example 6.6b are automatically replaced by sentences like Example 6.6b'. Similar replacements can be implemented for relative clauses as well. In addition, exercise generation targeting relative clauses should replace the subject of the second prompt sentence with a pronoun if it is identical to the subject of the first prompt sentence. A morphological annotator can inform the selection of the concrete pronoun by identifying the grammatical gender of the subject.

Result review The primary purpose of the graphical interface shown in Figure 6.6 is to present the results of the previous step to review the generated exercise parameters such as target forms, chunking, distractors, and hints in parentheses. Some exercise types share elements generated in the specification generation step. The GUI therefore presents the exercise elements and parameters for all exercise types, stored in the exercise specification, in a cumulative representation. This and the grouping of multiple transformations ① into a single specification reduces revision effort to a minimum. When conducting revisions at this stage of the exercise generation process, one only needs to check a single abstract specification instead of dozens of spelled out exercises. The transformations can be edited or deleted individually and additional transformations can be added as well.

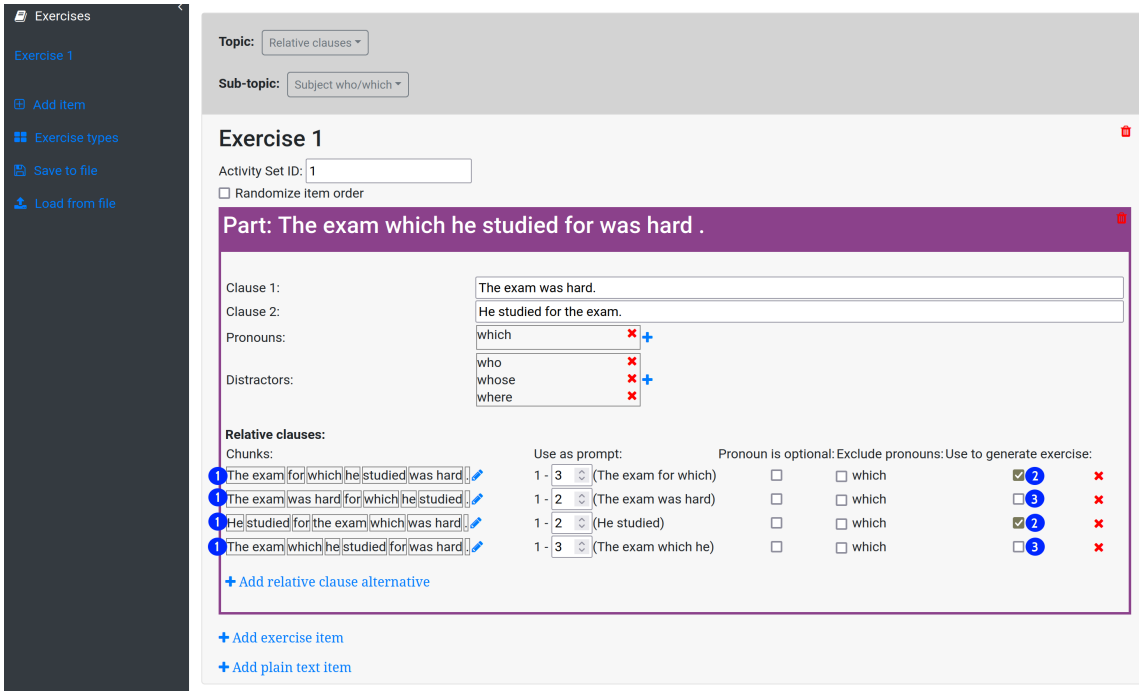


Figure 6.6: Specification authoring GUI.

The editor here presents the specification parameters for the `learning target` *Relative clauses*. It supports adding, modifying, and deleting specifications and their parameters. Transformations ① can be used to generate exercise instances ② or may only constitute accepted target answers ③.

6.1.3 Exercise Generation

Since each exercise type encodes a different set of `skills` and different learners may struggle with different `skills`, it makes sense to generate exercises of all supported exercise types and transformations. Even if all learners initially practice the same exercises when a new topic is introduced, additional practice can then specifically target `skills` that are problematic for the individual learner. The data compiled so far supports multiple exercise types with slightly different characteristics. Although our implementation was designed to support exercise generation for the Didi project (see Section 2.1.3), it already takes into account that some of the project's exercise types constitute variants used in meso-adaptive adjustments for a common, overarching type. This results, for example, in a joint exercise type (Gap-filling) for Single Choice and Fill-in-the-Blanks exercises.

User interface The settings to specify the exercises to generate can be configured in the exercise specification authoring interface shown in Figure 6.6. Apart from constituting an authoring interface, it also allows to mark each transformation as exercise seed from which to actually generate an exercise. If this option is not activated, the transformation merely serves as accepted correct answer alternative (provided the exercise context such as given prompts licenses the sentence). For instance, Figure 6.6 presents an exercise specification for relative sentences with four transformations ①. Since for two of the transformations, the flag *Use to generate exercise* is activated ②, two exercise instances will be generated per exercise type from the specification. The two transformations for which the flag is not activated ③ will be accepted as alternative correct target answers. In order to make sure that all resulting exercises have an associated transformation for all items, we maintain a link between the sentences of all exercise specification items that share the same parameter constellation. Deletion of one transformation therefore also deletes the corresponding sentence of all other items of the specification. Although some transformations of the same seed sentence require individual exercise specifications, all specifications associated with the same seed sentence are linked by a common identifier. As explained in Section 5.2, these item set IDs enable adaptive systems using the generated exercises to avoid selecting similar exercise items in succession for the same learner.

An additional web view provides configuration parameters specifying which exercise variations are generated. As can be seen in Figure 6.7, exercises can be generated for all transformations of each item specification, as well as for a random selection of a single transformation of each item. If the specification contains multiple exercise items, the second option results in more varied exercises whereas all items within the exercises resulting from the first option share the configured exercise properties. For relative clauses, the variations for Gap-filling exercises encompass the number of distractors, and the presentation of base words in parentheses or as bags of words above the exercise. Jumbled Sentence exercises on relative clauses may be configured to either keep relative pronouns as individual chunks or to combine them with adjoining chunks. For Short Answer and Mapping exercises, users can define in which order to display the clauses from which to form relative sentences in the prompt. For conditional sentences, the options for Gap-filling exercises are similar to those for relative clauses. In addition, users can configure for all exercise types whether

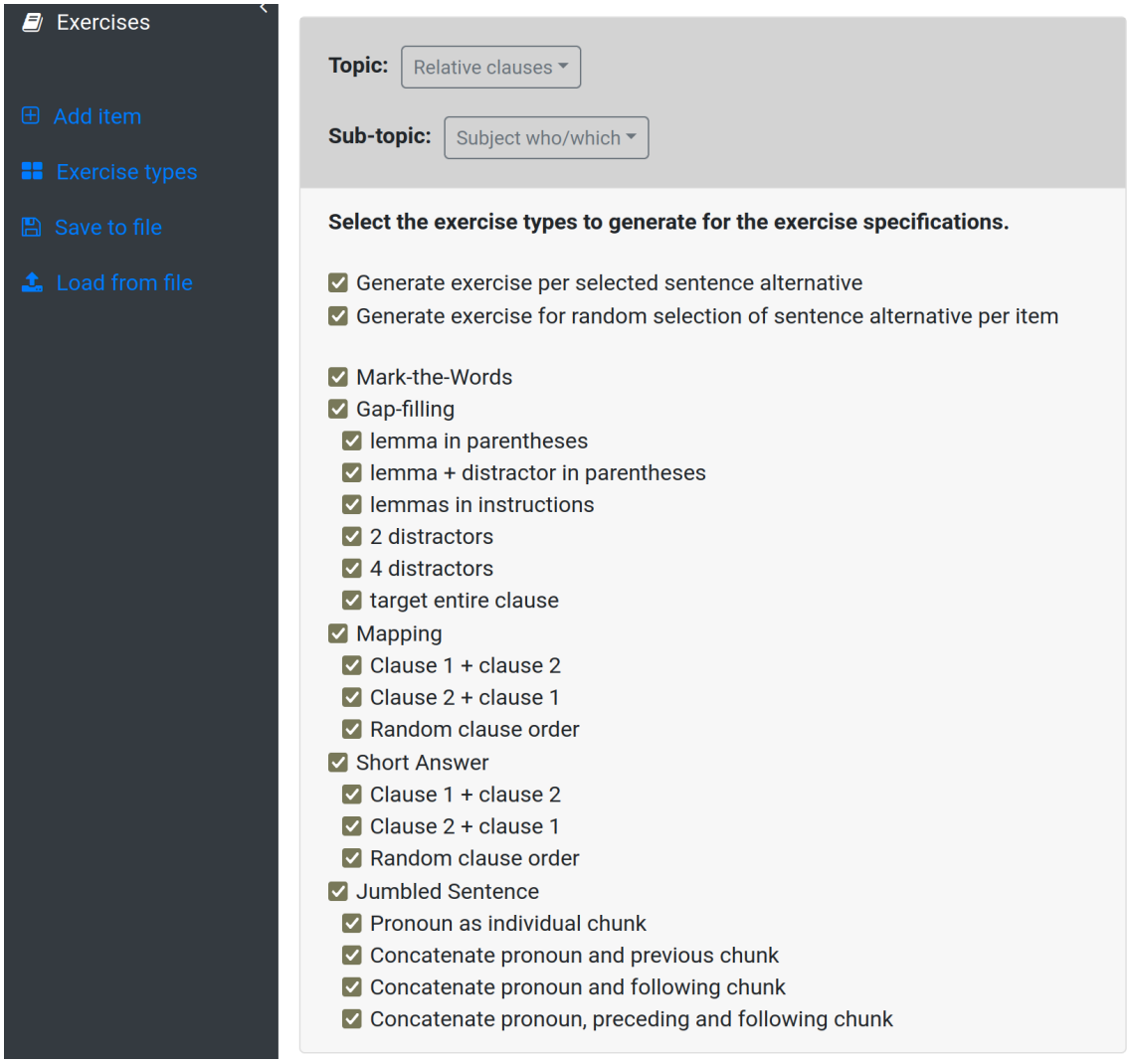


Figure 6.7: Exercise type definition GUI.

Users may specify which exercise types to generate in which variations. The supported types and their variations differ from one `learning target` to the other.

to insert exercise targets in both clauses or only one.

Note

Since the implementation was developed for the Didi project, the exercise type definitions do not support meso-adaptivity. In order to provide the generated exercises with all relevant information for meso-adaptive adjustments, an additional type configuration per exercise type will need to combine all information relevant to that type into a single exercise.

Algorithm Based on the specifications, subsequent exercise generation is straightforward. All necessary information is already contained in the exercise specifications, apart from instructions. These have been defined for each exercise type and linguistic structure by practicing teachers and are available to the algorithm as code resources. Similar to most exercise elements generated in the previous step, multiple instruction variants exist that are equally suitable for the exercises and differ merely in complexity. We store all variants with the exercises and add annotations of their specificity.

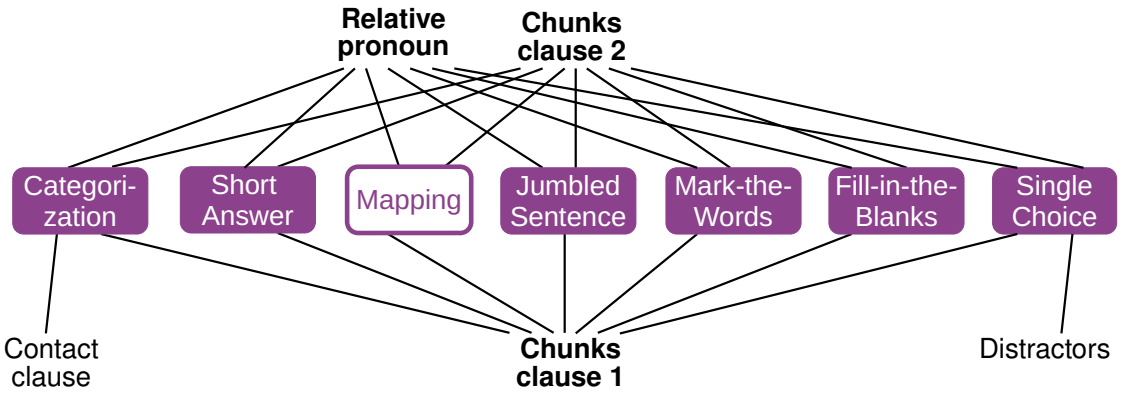


Figure 6.8: Parameter usage of different *Relative clause* exercises.

Different exercise types are characterized by different specification parameters. A specification parameter may be relevant for multiple exercise types.

Figure 6.9 illustrates how this subset of specification components is used to generate four distinct Mapping exercises. The parameters responsible for this variability

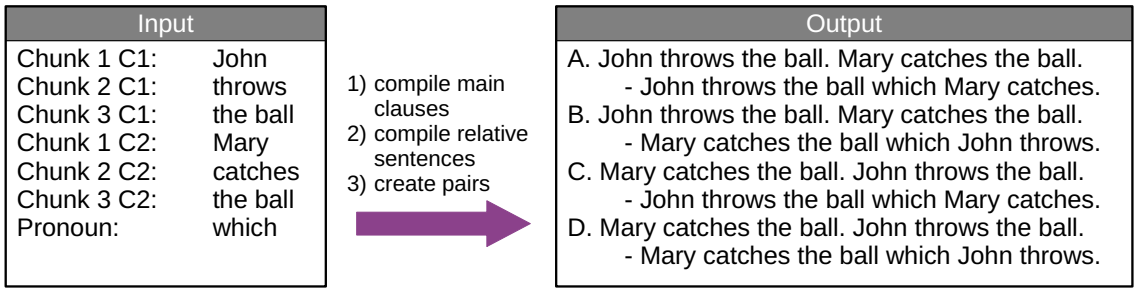


Figure 6.9: Generation of *Relative clause* Mapping exercises from a specification.

In order to generate a Mapping exercise from a specification, the algorithm compiles the two main clauses that can be transformed into a relative sentence, as well as the relative sentence, and associates the main clauses with the relative sentence.

consist in the order of the main clauses as well as in the choice of which clause forms the relative clause in the relative sentence.

Comparing Figures 6.8 and 6.10 highlights that, while some exercise parameters such as the relative pronoun or the conditional type are specific to a `learning target`,

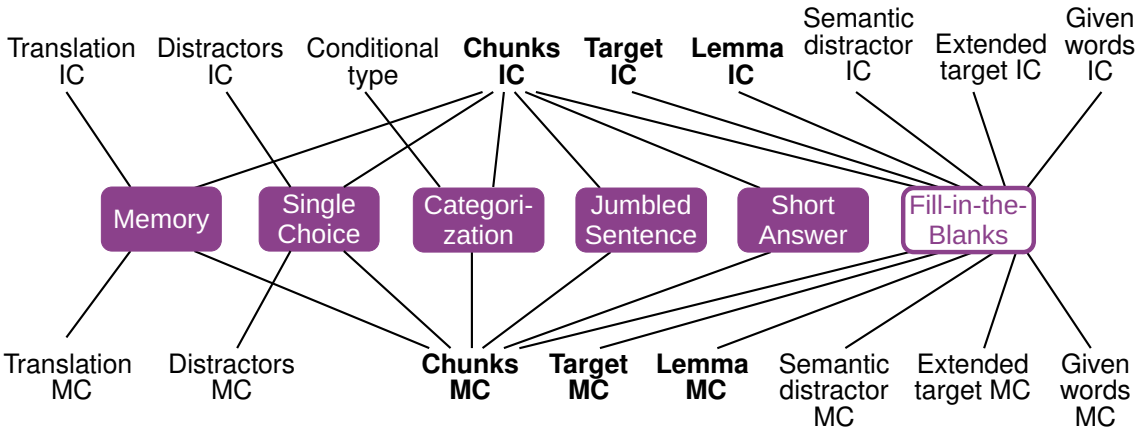


Figure 6.10: Parameter usage of different *Conditionals* exercises.

Exercises for *Conditionals* support the same exercise types as *Relative clauses* except for Mark-the-Words exercises, which are not supported for *Conditionals*. The two `learning targets` require slightly different parameters. Parameters for *Conditionals* must be specified for both the main clause (MC) and the if-clause (IC).

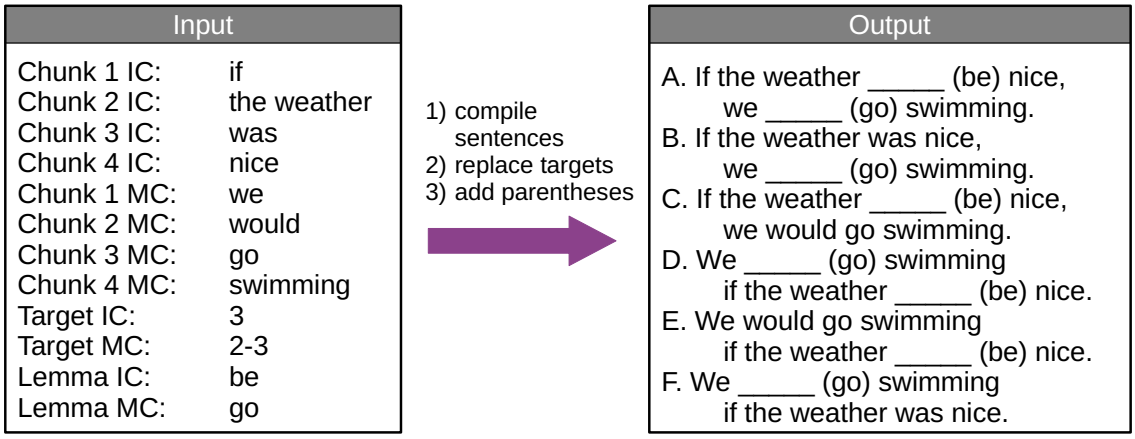


Figure 6.11: Generation of *Conditionals* Gap-filling exercises from a specification.

In order to generate a Gap-filling exercise from a specification, the algorithm first compiles the conditional sentence, then replaces the verb constructs with blanks, and finally adds parentheses with the lemmas of the verb constructs.

Similar to instructions of varying specificity, the element variations are stored in the exercise instantiation with indications of their specificity so as to allow the meso-adaptivity algorithm to dynamically select the appropriate one.

Apart from identifying the relevant exercise parameters, exercise generation consists in converting the specifications into the required output format. The JavaScript Object Notation (JSON) output is compatible with the exercise format used in the Didi system. In addition, H5P[12] exercises can be generated that may be used in any tool supporting this standardized format (López et al., 2021). The generated files are returned to the client where they can be downloaded by the exercise creator.

Evaluation The number of exercise items that can be generated from each seed sentence depends on three factors: (1) the user selections for sentence transformations in the specification definition GUI (Figure 6.4) and for exercise types in the specification authoring GUI (Figure 6.7), (2) the algorithm's success in generating sentence transformations, and (3) the grammar subtopic.

The maximum number of exercise items is given by the formulas in Equation 6.1.

[12]H5P is available at https://h5p.org/.

$$n_{all} = \underbrace{n_{types} * n_{spec}}_{n_{rand}} * n_{trans} \tag{6.1a}$$

with

$$n_{spec} = \prod_{i=1}^{\substack{item- \\ params}} \text{options}_i \tag{6.1b}$$

and

$$n_{trans} = \begin{cases} 1, & \text{if single random transformation} \\ n_{manual\_trans} + \prod_{i=1}^{params} \text{options}_i, & \text{otherwise} \end{cases} \tag{6.1c}$$

The number of generated exercise specification items n_{spec} constitutes the product of the options per activated transformation parameter of those parameters that produce separate exercise items. If all transformations are turned into exercise items, the number of activated transformations per exercise specification item n_{trans} is also considered. The number of transformations constitutes the product of parameters that can be configured in the exercise specification GUI (Figure 6.4) that result in separate transformations of the same specification. Exercise creators may in addition manually vary certain options for a transformation in the specification authoring GUI and thus add additional transformations. For instance, an exercise creator could specify multiple entries in the section *Relative clauses* of Figure 6.6 with identical chunks but different end indices for the prompt. If not all transformations are turned into exercises but only one randomly selected alternative of a transformation is used per specification item, the transformation factor equals *1* and therefore does not impact the result n_{rand} of the equation. The overall number of exercise items n_{all} constitutes the product of the number of exercise types n_{types} and the number of exercise specification items n_{spec} and, if applicable, the number of sentence alternatives per item n_{trans}.

Table 6.4 illustrates that, not considering manually added transformations, applying

this formula to the `realizations` *Conditional sentences*, *Differentiation of conditional types*, *Relative clauses with relative pronouns*, and *Contact clauses* results in up to more than 5,500 exercise items for a single seed sentence.

	C_{sent}	C_{diff}	RC_{pron}	RC_{cont}
Exercise types	34	21	16	3
Exercise specification items	81	81	3	3
$\mathbf{n}_{rand}$	**2,754**	**1,701**	**48**	**9**
Transformations	2	2	4	4
$\mathbf{n}_{all}$	**5,508**	**3,402**	**192**	**36**

Table 6.4: Maximum numbers of generated exercises.

For the random selection of an alternative (n_{rand}), the maximum number of generated exercise items depends on the available exercise types and the number of specification items. If each transformation is turned into an exercise, the number of active transformations per exercise specification item is considered in addition for the overall number of exercises (n_{all}). Available exercise types differ between the `realization` differentiating conditional types (C_{diff}), the remaining `realizations` of the `learning target` *Conditionals* (C_{sent}), the `realization` *Contact clauses* (RC_{cont}), and the remaining `realizations` (RC_{pron}) of the `learning target` *Relative clauses*.

6.2 Continuous Improvement through Learning Analytics

New systems that target learner groups for which performance data is either not available or not shared with the public need to rely on human intuition in order to determine learning goals and exercise parameters. Once learner data has been collected within the system, however, the practice material that the system provides can and should be re-evaluated. As more and more learner data becomes available, the material can then be continuously adapted to the learners' requirements. An approach to thus address RQ 10 can be found in the domain of Business Management. Continuous Improvement (CI) has been aiming to identify waste in companies' systems and processes for decades (Bhuiyan and Baghel, 2005). In an ILTS, waste corresponds to practice material that learners do not need in order to improve their language skills. This is the case if the learning goal of the practice material does not pose a challenge to learners. Grammatical forms that are not challenging for learners do not need excessive practice. On the other hand, forms where learners

make many errors should be practiced in a variety of exercises focusing on remedying existing misconceptions. By eliminating waste in the sense of effort to provide irrelevant practice material that a macro-adaptivity algorithm does not select for any learner, ILTSs can then instead focus on providing more instances of relevant material. In particular, in order to reduce the vast amount of automatically generated exercises of which only a fraction might actually be used in real-world settings, learning analytics can determine those variations that learners require. Exercises that do not practice linguistic forms with which learners struggle do not produce significant learning gains. Practicing word order, for instance, only makes sense if the exercise requires learners to determine the order of constituents that they tend to place in incorrect positions within a sentence. Jumbled Sentence exercises should therefore use sentence chunks that separate the constituents with which learners struggle, but should not require tedious ordering of easy constituents. Similarly, distractors of Single Choice exercises should be challenging, yet they should not expose learners to any misconceptions they would not have come up with on their own (Yamada, 2019). These considerations reveal three aspects of grammar exercises that should be aligned with learners' requirements: linguistic learning goals, target constituents for word order practice in Jumbled Sentence exercises, and distractors for Gap-filling exercises.

For distractor generation, approaches based on learner errors have been found to yield the best candidates (Lee et al., 2016). However, disentangling systematic errors from accidental mistakes and abstracting learner errors into patterns that facilitate the generation of distractors for arbitrary target answers is not a trivial task. Yet, many ILTSs incorporate error diagnosis mechanisms. Approaches that match answer hypotheses to error diagnoses are particularly interesting for distractor generation. The FeedBook, for instance, successfully pursues this approach (Rudzewitz et al., 2018). The answer hypothesis generation process shows strong similarities to distractor generation: The most frequent learner errors constitute the most plausible distractors whereas alternative, correct answers represent unreliable distractors that need to be avoided. Systems generating answer hypotheses for error diagnosis therefore inherently have the means to automatically generate distractors. This is especially valuable if Single Choice exercises are used for remedial practice as it opens the possibility to directly associate Single Choice exercises with learner errors and select exercises that best target a learner's misconceptions. The parallels

between error analysis based on answer hypotheses and distractor generation are striking, yet to the best of our knowledge, these two subfields of NLP have never been approached in tandem.

Although previous approaches to exercise generation have used learner errors solely for distractor generation (e.g., Lee et al., 2016), they can similarly inform chunking of Jumbled Sentence exercises for word order practice, and assist in determining relevant linguistic learning goals for which learners need practice material. We show the feasibility of using a system's error diagnosis mechanism for all three aspects of exercise generation (distractors, chunking, and learning goals) at the example of learner data collected in the I4S study (see Section 2.1.3).

Data The subset of the data used for the evaluations consists of the exercises of the `learning target` *Questions in the simple past*. It comprises 132 exercise items of four exercise types: 27 Jumbled Sentence items, 27 Single Choice items, 58 Fill-in-the-Blanks items, and 20 Short Answer items. 48 Fill-in-the-Blanks items gave lemmas in parentheses behind the blanks. The remaining 10 Fill-in-the-Blanks items presented all correct forms to insert into the blanks as bags of words in the exercise instructions. As this renders them more similar to Single Choice exercises, we treat them as such in this evaluation. Fill-in-the-Blanks and Short Answer exercises constitute half-open exercise types while Single Choice and Jumbled Sentence exercises are closed types. A total of 4,836 incorrect learner answers to an actionable element of the exercises was collected from 199 learners who submitted at least one of the exercises that we assessed for this evaluation. Unlike in previous evaluations, an actionable element of a Jumbled Sentence exercise was not defined as an entire sentence. Instead, since chunks constitute the elements of primary interest in the current evaluations, we defined an actionable element as a chunk. For the other exercise types, actionable elements were defined as usual: the blank of a Fill-in-the-Blanks or Single Choice exercise, or an answer to a Short Answer exercise. All submissions were considered so that there may be multiple answers per learner and actionable element if a learner re-submitted a revised answer.

We further determined misconceptions of interest by asking two annotators to freely annotate the entire dataset without any reference set of potential labels. From the union of the thus compiled error labels, we included those specific to questions in

the simple past in the final label set. In order to develop an annotated learner corpus from the learner answers, we relied on two sources: (a) automatic annotations provided by the system's error diagnosis and feedback module, and (b) manual annotations. The automatic annotations provide the single most relevant error for each learner answer. They were automatically refined into more fine-grained labels if simple string matching allowed more detailed diagnoses. For instance, the automatic annotations identified when an incorrect question word was used, but did not specify the correct wh-word nor the wh-word provided by the learner. The regular expression given in Example 6.10 was used to search the target answer and the learner's answer for a question word. This provided error labels specifying which wh-words were being confused by the learner. We used these annotations whenever available ($n = 1,778$) and manually annotated the remaining learner answers ($n_{answers} = 3,058$, $n_{labels} = 6,576$) if the system could not diagnose the nature of the error. Five annotators with backgrounds in computational linguistics annotated the learner answers independently with an unconstrained number of labels. Inter-Annotator Agreement (IAA) for the multi-label annotations of all annotators was calculated as Krippendorff's Alpha at $\alpha = .2075$. For the evaluations, the union set of manual and automatic annotations was used in order to not miss any potential errors. Although this might introduce some noise, it serves the purpose of identifying distractor candidates best.

6.10 ^.*?([w|W]hat|[w|W]ho|[w|W]hom|[w|W]hose|[w|W]hich|[w|W]here|[w|W]hen|[w|W]hy|[h|H]ow).*?$

In a second step, we contrasted the learner errors against misconceptions judged relevant by human exercise creators. To this purpose, we analyzed the manually created exercises, distractors, and Jumbled Sentence chunks used for these evaluations – representing the material considered relevant by human exercise creators – with respect to the errors for which they provide opportunities. We annotated the exercises with the same labels used for learner error annotations. A label was assigned if it is in principle possible to make the associated error in the exercise.

In order to address RQ 10, we evaluated the dataset with the aim of answering the following questions:

RQ10.1 Can the creation of learning goals, distractors, and Jumbled Sentence chunks be automated through learning analytics?

RQ10.2 Does human perception of relevant misconceptions align with relevant misconceptions derived from learning analytics?

RQ10.3 Do errors made in half-open exercises constitute plausible distractors for closed exercises?

In order to answer RQ10.1, we identified which steps in the process of identifying relevant learning goals, distractors, and chunk boundaries could not be based on automated processing of the data but instead required manual labor. In addition, we determined the ability of the system's error diagnosis module to automatically identify errors relevant to the `learning target` that learners make frequently. This outlines the status quo of possible automatization and indicates future directions for extending the error diagnosis module in order to support the envisioned learning analytics based adaptivity.

In order to answer RQ10.2, we identified the most frequent errors made in half-open exercises and contrasted them to error opportunities provided by the exercises considered in our evaluation.

Errors made in half-open exercises can only inform distractor generation if learners tend to choose the associated distractors in Single Choice exercises. Similarly, Jumbled Sentence exercises only support learning if the sentence constituents are separated into individual chunks in such a way that learners fail to put the chunks into the correct order. In order to answer RQ10.3, we therefore analyzed whether the identified most frequent errors from half-open exercises were also made in Single Choice and Jumbled Sentence exercises if the exercises provided opportunities to make them.

With RQ10.1 to RQ10.3 in mind, we tested our approach to use the system's error diagnosis module to determine relevant learning goals, generate distractors for Single Choice exercises, and create chunks of Jumbled Sentence exercises.

6.2.1 Learning Goal Definition

Covering all grammar topics within the time available for a language course is impossible (Burton, 2022), even if the class does not follow a communication-based curriculum but a grammatical syllabus, where the focus is on grammar. It is therefore necessary to select a subset of grammatical concepts that can be covered in the course. Yet how this selection should be made is unclear.

While high-level grammar topics to be taught are usually dictated by an institution, instructors are often free to select specific learning goals within these topics (Brophy, 1982). In order to maximize learning gains, the practice material should focus on learners' knowledge gaps. It therefore makes sense to determine learning goals based on learner errors.

Learning goals are defined by pedagogically motivated groupings of learner errors. We therefore manually identified linguistically and pedagogically related groups of

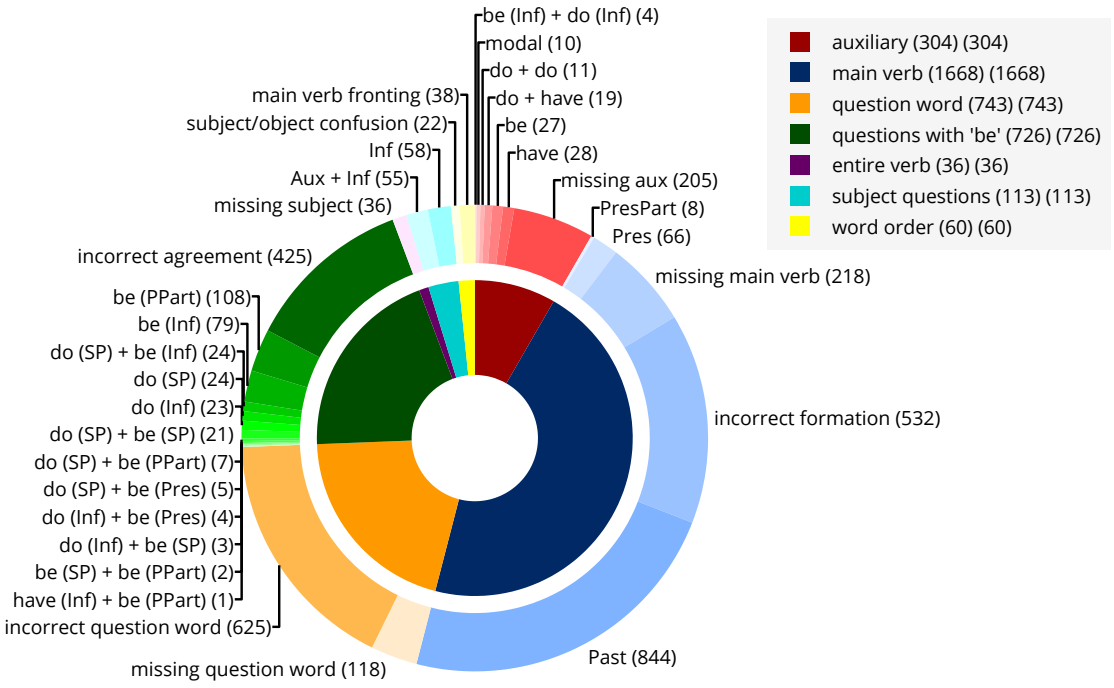


Figure 6.12: Absolute frequencies of error types.

Errors were annotated for the seven error categories *auxiliary*, *main verb*, *question word*, *questions with 'be'*, *entire verb*, *subject questions*, and *word order*. Each error category can be further split into multiple sub-categories.

error labels from our label set.

The resulting seven groups of errors indicate important learning goals. The groups are illustrated in Figure 6.12: Auxiliary errors indicate the need for practice targeting the auxiliary, main verb errors the need for practice targeting the main verb, and errors targeting the entire verb the need for practice of the entire verb form. Question word errors identify a learning goal in the selection of the correct question word. Word order errors indicate the need for word order practice. Errors in questions with 'be', and errors in subject questions constitute interesting special cases since question formation rules for these questions differ from the general syntax of questions. Their relevance as separate learning goals is strikingly emphasized when normalizing the error frequencies by the number of opportunities to make the respective error, as

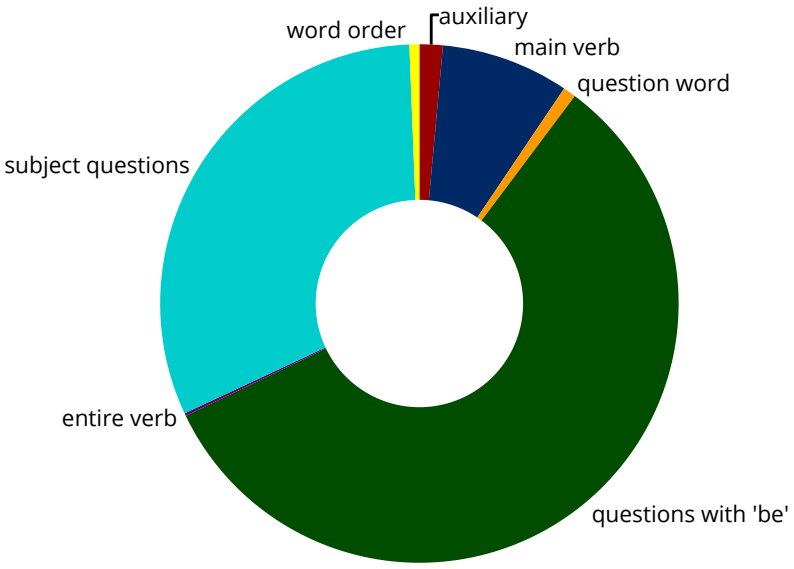


Figure 6.13: Normalized frequencies of error types.

When normalizing error frequencies by the number of opportunities to make them, *questions with 'be'* emerge as the most frequent source of learner errors.

Focusing on RQ10.1, we verified how well the system's automatic error diagnosis detects the errors identified as relevant to exercises on questions in the simple past. To this purpose, we determined the overlap between the labels used in manual and in automatic annotations. In addition, we manually annotated a subset ($n = 491$) of the automatically annotated learner errors and calculated multi-label IAA

between the automatic and the joint manual annotations. The evaluations revealed that the automatic annotations cover 34 of the 63 manual labels found relevant for exercises on questions in the simple past. Although this leaves substantial potential for extensions of the error diagnosis module, it also provides a solid starting point for our analyses. Automatic annotations include only a single label per error, yet human-system IAA was even slightly higher ($\alpha = .2175$) than human-human IAA ($\alpha = .2075$). The error diagnosis module can therefore be used for automatic exercise generation, although both error diagnosis and exercise generation would benefit from extending the coverage and granularity of diagnosed learner errors.

In order to address RQ10.2, we examined the system's exercises practicing questions in the simple past. They evidently provide practice opportunities for all annotated misconceptions as the errors were observed in the ILTS' learner records. Yet the number of opportunities to make the error might differ from one misconception and exercise type to another. In order to evaluate the available exercises' coverage of the identified learning goals, we determined the exercise annotations' coverage of the error labels. The analysis revealed that not all exercise types offered practice for all misconceptions. Figure 6.14 illustrates that not all learning goals that are relevant according to the learner records could be practiced with both closed and half-open exercises. Thus, there was no perfect overlap between learning goals defined by human exercise creators – represented through exercise labels – and those identified through learning analytics. While there was substantial common ground, exercise

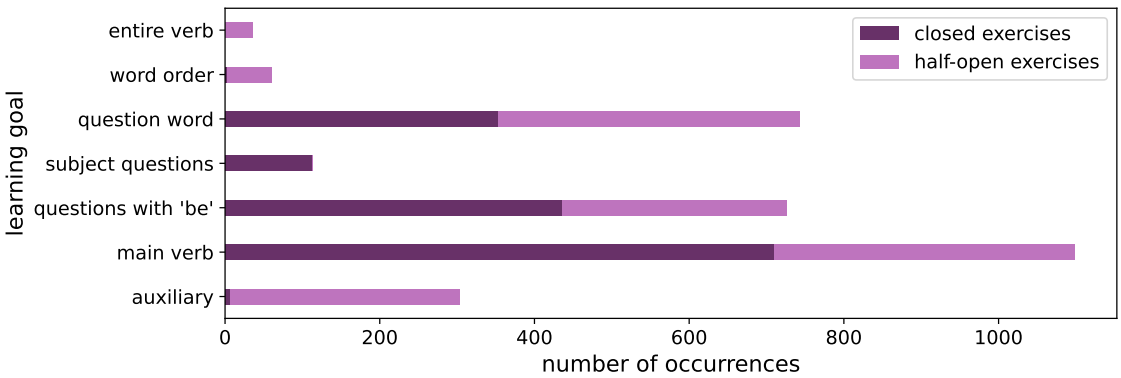


Figure 6.14: Error frequencies per exercise type.

Some error types are present only with closed exercises, some only with half-open exercises, and some with both.

creators and learning analytics differed in the emphasis they put on different misconceptions. Since this emphasis was inconsistent in human output across exercises, human exercise creators might benefit from explicitly specifying the learning goal in a first step. Our evaluations suggest that highest pedagogic validity of exercises can be achieved by relying on human effort to classify learner errors when defining learning goals, and on learning analytics based, automatic processing in order to generate an adequate number of exercises for each learning goal.

6.2.2 Distractor Generation and Selection

While half-open exercises require learners to produce the target form, closed exercise types provide a range of answer alternatives to choose from (Spada and Tomita, 2010). Closed-type Single Choice exercises require special attention as they expose learners to incorrect linguistic material in the form of distractors (Roediger and Marsh, 2005). While distractors should cover developmental misconceptions in order to be sufficiently challenging and thus relevant to learning, they should not introduce misconceptions that the learner does not already have (Yamada, 2019).

Given these considerations, it is not surprising that distractor generation is seen as the most challenging aspect of generating Single Choice exercises (Mitkov et al., 2006). In order to determine pedagogically valid and plausible distractors, human judgment is often deemed best (Susanti et al., 2018), yet even manually created distractors do often not meet the requirements of being challenging but not introducing new misconceptions (Haladyna and Downing, 1993; Patil et al., 2016). In order to automate distractor generation and at the same time increase plausibility and validity, data-driven approaches base distractors on common misconceptions extracted from a learner corpus (Lee et al., 2016). This in addition allows a more learner-centered adaptation of distractors by dynamically selecting those distractors for each learner from a pool of options that target their individual misconceptions.

Distractor generation usually consists of candidate generation and candidate filtering and/or ranking, although filtering and ranking are sometimes executed in a single step. Many implementations combine a number of different filtering and ranking approaches.

Table 6.5 provides a non-comprehensive overview of related work on distractor gen-

eration. Most work on distractor generation is centered around question answering and vocabulary-focused Gap-filling exercises. The approaches differ in the source from which the pool of distractor candidates is compiled, as well as in the filtering and ranking strategies. Table 6.5 outlines that the candidates are either extracted from unstructured data such as text corpora, from structured data such as databases or word lists, or else generated based on machine learning or on transformation rules. The candidate pool then comprises either a subset (Sumita et al., 2005; Stasaski and Hearst, 2017) or all entries (Smith et al., 2010; Pérez and Cuadros, 2017) of the resource, or transformations of the entries (Maritxalar et al., 2011; Quan et al., 2018) or of the target answer (Zesch and Melamud, 2014). Filtering and ranking depend on the intended distractor type such as ungrammatical, nonsensical, and plausible distractors (Mostow and Jang, 2012). The distractor type determines, for example, the usefulness of grammaticality checks (Pino et al., 2008; Moser et al., 2012). For plausible distractors, the desired similarity of the distractors with the target answer constitutes an additional factor. This is on the one hand influenced by the task setup, as for example synonyms may be context-inappropriate and therefore useful distractors for contextualized exercises (Knoop and Wilske, 2013), yet would constitute unreliable distractors if they can correctly replace the target answer (Hill and Simha, 2016). In addition, since exercise difficulty increases with distractor plausibility, target similarity can be adjusted according to the learner's proficiency (Alsubait et al., 2015; Chen et al., 2015; Correia et al., 2012). As can be seen in Table 6.5, similarity can target the surface form, linguistic complexity, phonetics, morphology, syntax, or semantics and be based on NLP tools including POS taggers, Latent Semantic Analysis (LSA) and word embedding models, on external resources such as ontologies, WordNet[13] or FrameNet[14], or else on statistical methods including classification, regression, and deep learning. If the final candidate selection is not based on a ranking, it may be left to the exercise creator (Nikolova, 2009), or done randomly (Araki et al., 2016; Gutl et al., 2011).

[13]WordNet is available at https://wordnet.princeton.edu/.
[14]FrameNet is available at https://framenet.icsi.berkeley.edu/.

Table 6.5

	Candidate source	Candidate pool	Candidate whitelist	Candidate blacklist	Candidate ranking
Papasalouros et al. (2008)	ontology	nodes	relation match		
Stasaski and Hearst (2017)	ontology	nodes	relation match		
Mitkov et al. (2006)	WordNet	nodes	siblings		
Aldabe and Maritxalar (2014)	WordNet	nodes	- siblings - hypernyms - hyponyms	no occurrence in corpus	- LSA similarity - PageRank scores - word family variety
Sumita et al. (2005)	thesaurus	entries	relation match	collocational occurrence in corpus	
Smith et al. (2010)	thesaurus	entries			thesaurus frequencies
Karamanis et al. (2006)	thesaurus	entries	- semantic type match - related words		thesaurus frequencies
Susanti et al. (2018)	- text - thesaurus	- text tokens - thesaurus entries	- difficulty level match - WordNet siblings - POS match	- POS mismatch - length mismatch	- GloVe embedding similarity - collocational frequency
Liang et al. (2017)	word embedding model	vocabulary list			
Kumar et al. (2015)	word embedding model	vocabulary list			embedding similarity
Welbl et al. (2017)	- word embedding model - corpus - exercise corpus	- vocabulary list - phrases - distractors	noun phrases		plausibility classification
Knoop and Wilske (2013)	- text - WordNet	- text tokens - WordNet nodes	- POS match - synonyms - antonyms		
Mitkov et al. (2006)	corpus	tokens	head match		

Continued on next page

Table 6.5: Related work on distractor generation.

Table 6.5 – continued from previous page

	Candidate source	Candidate pool	Candidate whitelist	Candidate blacklist	Candidate ranking
Mostow and Jang (2012)	text	tokens	- POS match - collocations match	- ungrammatical sentences - tokens highly relevant in the text - tokens not relevant in the sentence	
Nikolova (2009)	corpus	tokens	pattern match		
Quan et al. (2018)	corpus	domain terminology			LSA similarity
Moser et al. (2012)	corpus	semantic concepts	POS match	ungrammatical sentence	- LSA similarity - stylomentric similarity
Araki et al. (2016)	corpus	event triggers		identical to TA	
Maritxalar et al. (2011)	corpus	tokens (morphologically adjusted)	POS match		LSA similarity
Goto et al. (2010)	- transformation rules - WordNet	- transformations of TA - WordNet nodes	- synonyms - antonyms - affix match	collocational occurrence in corpus	
Lee and Seneff (2007)	annotated learner corpus	learner errors			- corpus-based frequency - collocational frequency
Žitko et al. (2009)	transformation rules	transformations of TA			
Liang et al. (2017)	- word embedding model - Generative Adversarial Network	model output tokens			- plausibility classification - adversarial network score
Jiang and Lee (2017)	exercise corpus	distractors		- POS mismatch - collocational occurrence in corpus + dependency relations match	- Word2Vec embedding similarity - token overlap - corpus-based frequency - collocational frequency
Susanti et al. (2015)	WordNet	nodes	siblings	POS mismatch	WordNet based semantic similarity

Continued on next page

Table 6.5: Related work on distractor generation.

Table 6.5 – continued from previous page

	Candidate source	Candidate pool	Candidate whitelist	Candidate blacklist	Candidate ranking
Gutl et al. (2011)	WordNet	nodes	- siblings - antonyms		
Guo et al. (2016)	corpus	corpus sentences	syntax tree match		- embedding similarity - categorical closeness - corpus-based frequency
Chen et al. (2015)	corpus	tokens	- POS match - semantic category match		WordNet based semantic similarity
Pilán and Volodina (2014)	FrameNet	entries			- frame similarity - FrameNet frequencies
Sakaguchi et al. (2013)	annotated learner corpus	ML model output			
Liu et al. (2005)	dictionary	entries	- POS match - frequency match		reverse co-occurrence frequency with important word in sentence
Yeung et al. (2019)	ontology-based BERT model	- related words - model output		Word2Vec embedding similarity	BERT probabilities
Gao et al. (2020)	dictionary	entries	- POS match - difficulty match - frequency match		plausibility classification
Liu et al. (2017b)	corpus	words			regression-based similarity
Gates (2011)	corpus	TAs of all generated exercises		- high WordNet similarity - high length difference - syntactic pattern mismatch	
Hill and Simha (2016)	corpus	n-grams	POS match	- WordNet synonyms - higher collocational frequency with co-text than TA	inverse corpus frequency
Smith et al. (2009)	thesaurus	entries			thesaurus frequencies

Continued on next page

Table 6.5: Related work on distractor generation.

Table 6.5 – continued from previous page

	Candidate source	Candidate pool	Candidate whitelist	Candidate blacklist	Candidate ranking
Mitkov et al. (2009)	WordNet	nodes	- siblings - hypernyms - hyponyms		- collocational frequency - WordNet similarity - phonetic similarity
Correia et al. (2012)	- word list - transformation rules	- entries - transformations of TA	POS match	- WordNet synonyms - hypernyms - hyponyms - domain match	Levenshtein similarity
Pérez and Cuadros (2017)	word embedding model	vocabulary list			embedding similarity
Zesch and Melamud (2014)	context-insensitive inference rules	transformations of TA		output of context-sensitive rule application	
Coniam (1997)	word list	entries	- POS match - frequency match		
Shei (2001)	word list	entries	- POS match - frequency match		
Brown et al. (2005)	WordNet	nodes	- POS match - frequency match		
Pino et al. (2008)	dictionary	entries	adverbs	- ungrammatical sentence - high WordNet similarity	collocational frequency

Table 6.5: Related work on distractor generation.

Most approaches focus on distractors for target answers (TA) in vocabulary exercises. Research on distractor generation for grammar exercises is scarce.

While automatic distractor generation has been widely explored in research for vocabulary exercises, distractors for grammar exercises have received less attention. For closed class grammatical forms such as prepositions, many of the approaches used for vocabulary distractors are applicable. However, this greatly underrates the importance of linking distractors to the pedagogical learning goal as good distractors characterize the space of options that a learner needs to weigh against each other. Since the core of form-focused grammar exercises is not semantics but form, good grammatical distractors usually rely on ungrammatical forms (Volodina et al., 2014). Goto et al. (2010) illustrate that, for closed class target answers, the initial

candidate pool consists of all types belonging to the class, whereas for open class target answers, transformations may produce suitable distractors. For the closed class of prepositions, Lee et al. (2016) start with the defined set of prepositions as candidates. For ranking, they consider co-occurrence of the candidates with either the prepositional object or the head, and their frequency as annotated errors or learner-corrected tokens in a learner corpus. For open class types, Chen et al. (2006) use distractor generation rules that introduce modifications of the target answer such as morphological or syntactic variants. Aldabe et al. (2007) present an approach to generate morphological transformations of the target answer as distractor candidates and discard those whose morpho-syntactic pattern can be found in a corpus. For verb exercises, Aldabe et al. (2009) filter the verbs from the Academic Word List by transitiveness, tense, and person and rank them according to semantic similarity and distributional data. Heck and Meurers (2022) apply NLP as well as rule-based transformations to generate well- and ill-formed variations of the target answer.

Lee et al. (2016) found distractor generation based on learner errors to yield the most plausible distractors. While their approach is closest to what we suggest, it relies on a manually annotated corpus. The resulting, statically determined distractors may be sufficiently representative for the learner population that provided the error corpus, yet they are likely to be unsuitable when applied to a broader learner population. In addition, Pino and Eskénazi (2009) found that distractor plausibility differed between learners of different native languages. While their distractors included only orthographic, phonemic and morphological variants of the target word, more diverse distractors might result in even greater differences in plausibility among learners of different abilities and characteristics. We therefore investigate how automatic annotations obtained from a system's error diagnosis mechanism can effectively be used to generate and dynamically select valid and plausible distractors.

While feedback generation aims to cover as many learner errors as possible, distractor generation needs to focus on the most frequent learner errors. This requires to filter the answer hypotheses generated for feedback provision. Tversky (1964) found drop downs with three options to result in the most reliable Single Choice exercises. We therefore aimed to determine the two most frequent errors made in half-open exercises in order to identify the best suited distractors for most learners in Single

Choice exercises. Since not all error types can be made in all exercises, we normalized the occurrences of misconceptions by the number of exercise items that provided opportunities to make the error. The results are visualized in Figures 6.15–6.17.

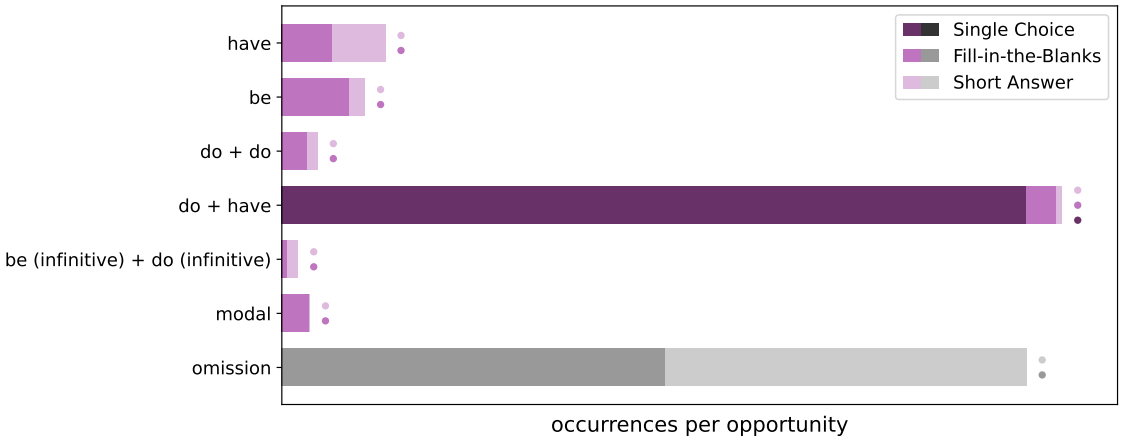


Figure 6.15: Frequencies of errors targeting the auxiliary.

Except for errors concerning *modals* in Short Answer exercises, all error sub-types that we analyzed occurred with both half-open exercise types. Single Choice exercises provided opportunities (indicated as dots next to the bars) for only a single one of these error sub-types with auxiliaries (*do + have*). The most frequent error in half-open exercises (marked in gray color) with respect to auxiliaries consists in omitting the auxiliary.

The most frequent error observed in half-open exercises with respect to **auxiliaries** made by learners (see Figure 6.15) consists in leaving it out (Example 6.11a). Of the remaining errors observed in half-open exercises, using *be* (Example 6.11b) or *have* (Example 6.11c) instead of the auxiliary *do* are most frequent. Combinations of multiple auxiliaries (Example 6.11d) were also observed, but only in occasional submissions of half-open exercises.

6.11 What did Mr. Connor bake?

a. *What baked Mr. Connor?
b. *What was Mr. Connor bake?
c. *What had bake Mr. Connor?
d. *What does Mr. Connor have bake?

With respect to the **main verb** (see Figure 6.16), the most frequent learner error consists in using the simple past or past participle form instead of the infinitive

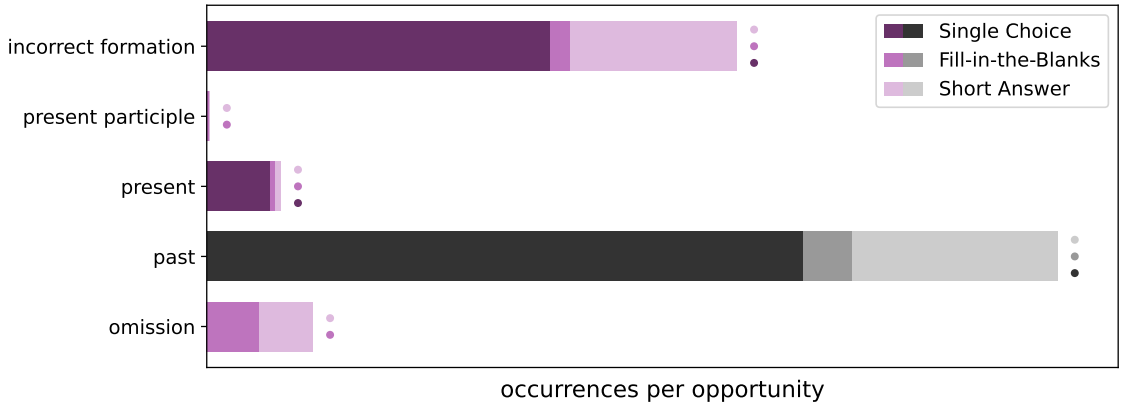


Figure 6.16: Frequencies of errors targeting the main verb.

Single Choice exercises did not provide opportunities (indicated as dots next to the bars) for all sub-categories of errors targeting the main verb. The most frequent sub-category observed in half-open as well as in Single Choice exercises (marked in gray color) consists in the use of a past form instead of the infinitive.

(Example 6.12a). Omitting the main verb altogether (Example 6.12b), using simple present – identifiable through the third person singular 's' – (Example 6.12c), or incorrectly forming the main verb (Example 6.12d) were also observed rather frequently in learner answers. The latter error constitutes a special case in that it occurs only in combination with other misconceptions. This makes sense considering that infinitives do not transform the verb so that learners do not struggle with the correct formation of the main verb if they correctly intend to provide it in this mood. Other misconceptions appeared only occasionally in learner answers.

6.12 Did you enjoy them?

a. *Did you enjoyed them?
b. *Did you them?
c. *Did you enjoys them?
d. *Did you enjoyd them?

Only a single misconception, namely omitting the subject, emerged as relevant to the learning goal of correctly forming the **entire verb**. This error could be found with Fill-in-the-Blanks as well as Short Answer exercises, which constitute the two exercise types providing opportunities for the error. Single Choice exercises, however, did not provide opportunities to make the error.

The most frequent error concerning **questions with 'be'** (see Figure 6.17) by far constitutes incorrect agreement with the person of the subject (Example 6.13a). With Fill-in-the-Blanks exercises, additional do-support (*did be*; Example 6.13b), do-support in combination with an inflected form of 'be' (*did was/were*; Example 6.13c), an inflected form of 'do' instead of 'be' (*did*; Example 6.13d), and *do* (Example 6.13e) were also frequent and should therefore be considered for distractor generation. For the latter error it is unclear whether the learners intended to use an infinitive or a simple present form of the verb.

6.13 Were you scared?

a. *Was you scared?
b. *Did you be scared?
c. *Did you was scared?
d. *Did you scared?
e. *Do you scared?

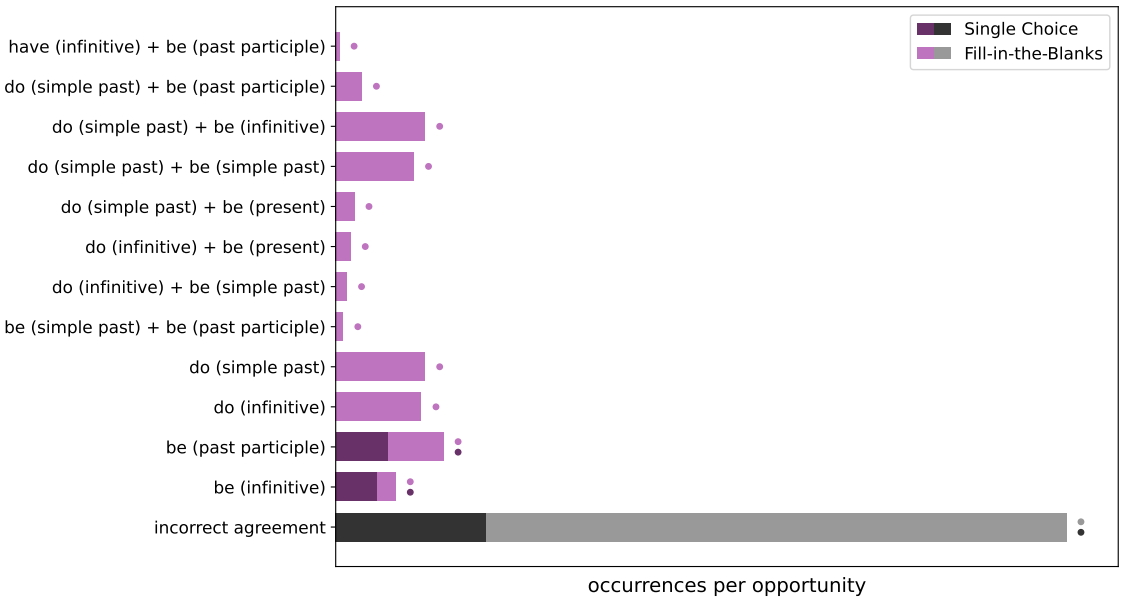


Figure 6.17: Frequencies of errors in questions with 'be'.

Incorrect agreement with the person of the subject constitutes the most frequent subcategory of errors (marked in gray color) in questions with 'be', although both Single Choice and Fill-in-the-Blanks exercises provided opportunities (indicated as dots next to the bars) for the remaining sub-categories as well.

As our data set does not contain half-open exercises for **subject questions**, it is not possible to determine from the data what kind of errors learners would produce on their own. Observed misconceptions are therefore restricted to those offered by the Single Choice distractors. They consist in using only the infinitive of the main verb (Example 6.14a) or else the infinitive with do-support (Example 6.14b).

6.14 Who persuaded you to come to the party?

a. *Who persuade you to come to the party?

b. *Who did persuade you to come to the party?

Exercises on **question words** constitute a special case in that they are of a more semantic rather than a syntactic or morphological nature. Misconceptions do not target word formation, but the choice of the most suitable question word. The bar chart in Figure 6.18 illustrates that although almost all possible question word confusions are present in the dataset, there are clearly discernible, predominant misconceptions in the use of question words. These are, however, not bi-directional. While *where* was often incorrectly substituted by *what* or *when* in the normalized dataset, the most frequently used question words instead of *what* are *how* and *which*, and omitting the question word altogether or using *where* was the most frequent error with *when*. Instead of *why*, learners most often used *how* or *who*, whereas the most frequent question word instead of *how* was *what* or sometimes *why* in the dataset.

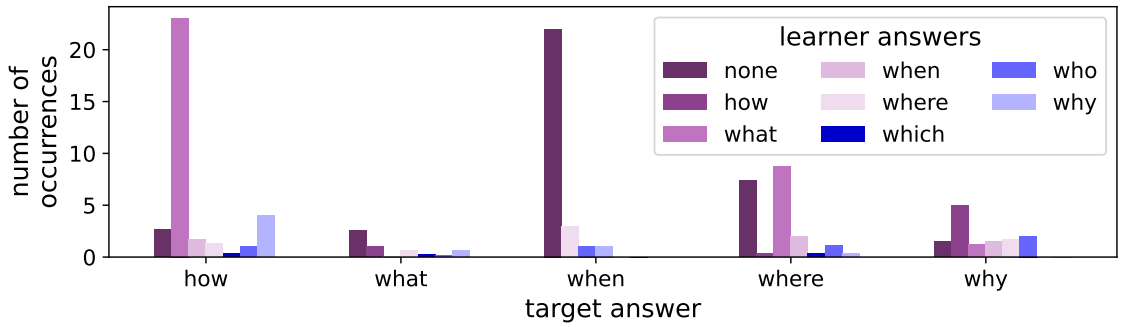


Figure 6.18: Frequencies of question word confusions.

Although incorrect question word usage was focused on certain confusions of question word pairs, the confusions are not bi-directional.

With respect to RQ10.1, we determined whether the automatic annotations cover the labels of the two most frequent errors in half-open exercises. Looking at the indi-

vidual learning goals, the most frequent misconceptions with auxiliaries – omission and the use of *be* – are both covered by the automatic annotations so that the error diagnosis module was able to generate such distractors. The automatic annotations do not yet cover any of the misconceptions concerning the main verb, the entire verb, or questions with 'be'. They do, however, support all labels for question word errors, thus providing the means to automatically generate according distractors.

These results demonstrate the feasibility of using a system's error diagnosis mechanism to automatically annotate relevant learner errors in order to dynamically adapt distractors to a learner's misconceptions. Although not all of the most frequent errors are automatically annotated in our analyzed system, it is possible to extend the error diagnosis module to generate additional answer hypotheses that cover errors that we identified as relevant for learners. Extensions of the error diagnosis in turn allow to further improve distractor generation. The presented evaluations of most frequent learner errors based on the entire learner corpus on questions in the simple past collected in the I4S study were performed offline in order to establish a proof of concept. They can, however, also be applied at runtime based on the errors recorded for a specific learner and thus inform an adaptivity algorithm about which distractors are most suitable for that learner. In addition, the results of the offline analysis may be used to define default values for distractors while the system still lacks learner records for individually adapted exercise configurations.

Focusing on RQ10.2, we compared the distractors introduced by human exercise creators with the most frequent errors in the learner records.

For **auxiliaries**, distractors in the dataset covered only the use of *do + have* out of the identified misconceptions. Although this distractor was selected very frequently in Single Choice exercises, the error appeared only occasionally in half-open exercises. The most frequent error with respect to this learning goal, leaving it out, was not covered by any of the Single Choice distractors in the system. This makes sense considering that according Single Choice exercises focusing only on the auxiliary would require an empty distractor. This exercise type thus does not lend itself well for practice targeting omission errors. However, the system's Single Choice exercises did not cover any of the remaining observed misconceptions related to auxiliary usage either.

The distractors did, however cover three of the most frequent misconceptions iden-

tified in half-open exercises practicing the **main verb**: using a past form, simple present, or incorrectly formed variant of the main verb. Only omitting the main verb altogether, which was also observed rather frequently, was not covered for the same reason why it was not covered for auxiliaries.

The system did not provide any Single Choice exercises to practice the **entire verb**.

Concerning **questions with 'be'**, Single Choice exercises offered distractors targeting the most frequent misconception, incorrect agreement, as well as the use of the past participle (*been*; Example 6.15a) and of the infinitive of 'be' (Example 6.15b) instead of its simple past form. However, the latter two were not among the most frequent learner errors.

6.15 Were you scared?

a. *Been you scared?

b. *Be you scared?

In summary, while the available distractors for Single Choice exercises did not cover all misconceptions found in the learner submissions, coverage of the identified most frequent errors was high. Only misconceptions targeting word order, which is better practiced with Jumbled Sentence exercises, and omission errors were not covered by the distractors. Concerning their pedagogical validity in terms distractors being based on learners' misconceptions, all errors for which Single Choice distractors provided opportunities were also observed in half-open exercises, except errors concerning subject questions, for which half-open exercises did not provide any opportunities. The same holds for co-occurrences of error labels, indicating that the available distractors only integrated combinations of misconceptions that learners also tended to make jointly in production exercises. The manually created distractors therefore seem to be pedagogically valid since the system did not expose learners to misconceptions they would not have developed of their own accord. However, these results need to be considered with caution since some learners may have practiced Single Choice exercises before half-open exercises, thus being introduced to misconceptions in the closed exercises. We thus cannot rule out that they would not have made the errors in half-open exercises otherwise. In addition to the misconceptions covered by the error labels, the distractors encompassed errors that had not been identified

as pedagogically relevant in the manual annotation and selection process. Although both distractor creation and learning goal identification constituted manual processes, they thus put different foci on targeted misconceptions. This might indicate that exercise creators do not intuitively choose distractors that are relevant to the learning goal and that automatic distractor generation based on learner errors can improve distractor generation beyond what is possible through manual labor.

In order to compare manually created distractors to those informed by learning analytics in terms of plausibility, we followed Haladyna and Downing (1993)'s approach, which states that at least 5% of all observed incorrect answers to a question need to correspond to the distractor in order for it to be plausible. We calculated the percentage of a distractor being selected among any of the item's distractors. Distractors obtaining a ratio lower than .05 were thus considered implausible. The evaluation shows that all distractors were selected at least once, although with differing frequencies. Only two instances of distractors were beneath the 5% threshold. While the incorrect form *forgat* may indeed be implausible, there is no clear indication as to why the form *been* was selected so rarely in the distractor group *be - was - been*, which appears in the same constellation in various other items submitted before as well as after that item, where this distractor was selected more frequently.

Turning to RQ10.3, we analyzed whether the identified most frequent errors in half-open exercises appeared in Single Choice exercises as well when according distractors were available. As it turns out, the system's distractors did not cover the most frequent errors for all learning goals. With respect to the main verb, they supported three of the most frequent misconceptions identified in half-open exercises: using a past form, simple present, or an incorrectly formed variant of the main verb. These distractors were selected frequently by learners in Single Choice exercises. Concerning questions with 'be', Single Choice exercises offered distractors targeting the most frequent misconception. Distractors targeting this misconception were also selected frequently by learners. This indicates that errors observed in half-open exercises constitute plausible distractors for Single Choice exercises, although the available distractor coverage was too scarce for any generalizable conclusions. Where no learner data from half-open exercises was available, no conclusions could be drawn about the pedagogical validity of learner errors as distractors. However, the good fit of misconceptions from half-open exercises as distrators that we observed

for the learning goals *main verb formation* and *questions with 'be'* provide promising evidence that the same also holds for all other learning goals. It is also yet unclear to what extent the most frequent misconceptions differ between and within learners over extended periods of time.

6.2.3 Sentence Chunking

Jumbled Sentence exercises are a natural choice of exercise type for controlled practice of word order. In order to constitute useful practice material, the chunks should fulfill two criteria: (1) On the one hand, they should be small enough to separate the challenging constituents that learners may struggle to assemble in the correct order. (2) On the other hand, the chunks should only be as small as necessary so as not to distract from the learning goal. We therefore analyzed word order errors with the goal of identifying constituents that should be separated into individual chunks.

Word order errors particular to questions in the simple past concern fronting of the main verb before the subject (Example 6.16a), as well as interchanging the subject and the object of the sentence (Example 6.16b). Relevant chunks for Jumbled Sentence exercises therefore comprise a chunk for the main verb, for the subject, and for the object.

6.16 Did Mr. Jones see a doctor?

a. *Did see Mr. Jones a doctor?
b. *Did a doctor see Mr. Jones?

With respect to RQ10.1, the automatic annotations did not further distinguish between specific sub-types of word order errors. Thus, our system's error diagnosis could not be used to determine the most appropriate chunking for a learner.

Addressing RQ10.2, we analyzed the Jumbled Sentence exercises in the system. To evaluate the first criterion concerning sentence chunking, we determined whether the exercises provided opportunities for making the word order errors that learners made in half-open exercises. In the exercises, 10 out of the 27 items merged the main verb with the succeeding token, thus not supporting main verb fronting errors. Only 11 items had individual chunks for the subject and the object, while the remaining 16

items had either no object or merged it with the preceding preposition or succeeding main verb. With regard to the second criterion, we determined the number of chunks not corresponding to a constituent involved in any of the errors. To this end, we subtracted the general number of word order relevant constituents from the number of the exercise item's chunks. Allowing for some preceding and succeeding co-text, we defined results greater than 2 as indicative of excessive chunking. We found that the sentences in the system were split into a mean of $\mu = 5.33$ ($\sigma = .88$) chunks so that according to the criterion of $n = 3$ relevant chunks $+2$, the overall number of chunks was only sightly higher than the optimal number of 5 chunks. Considering that most exercise items merged some of the relevant chunks with preceding or succeeding tokens or did not incorporate them at all, however, chunk boundaries were rather frequently set at pedagogically questionable places.

Regarding RQ10.3, the learner error data reveals that while Jumbled Sentence exercises offered opportunities for all word order errors concerning questions in the simple past that learners made in half-open exercises, none of the learners made any main verb fronting errors in these exercises, indicating that this is only an issue in more open exercises. Subject/object confusion, on the other hand, was only observed with Jumbled Sentence and Fill-in-the-Blanks, but not with Short Answer exercises, although all three exercise types offered opportunities for this error. Since this misconception is of a more semantic nature, this could suggest that learners do not put much effort into semantically parsing sentences in less open-ended exercise types, rendering subject/object errors careless mistakes rather than misconceptions. Thus, neither subject/object confusion nor main verb fronting seem to be relevant for Jumbled Sentence exercises. This might suggest that Jumbled Sentence exercises are not relevant for practicing question formation since word order issues seem to arise mostly in combination with formation issues so that learners cannot practice these issues with form-controlling Jumbled Sentence exercises. On the other hand, the fact that we observed word order errors specific to word order of questions in the simple past only in exercises where learners have to focus on multiple linguistic levels at once could also indicate that they lack automatization that would allow them to overcome processing overload. In this case, Jumbled Sentence exercises could provide opportunities to practice the learning goal of word order in isolation.

6.3 Summary

In this chapter, we focused on exercise generation that supports adaptive ILTSs focusing on linguistic variability.

In response to RQ 9, we presented and evaluated an approach to exercise generation that can efficiently create the vast numbers of exercises required for varied and learner-adaptive practice and at the same time ensure high quality of the practice material. This approach separates exercise generation into the three steps of seed sentence selection, exercise specification generation, and exercise generation. By allowing exercise creators to review the result after each step in a web interface integrated into the exercise generation workflow, revision can always be performed at the most abstract level before the result item is split into multiple instantiations in the next step.

Although NLP and NLG make it possible to automatize exercise generation to a large degree, we demonstrated that manual revision steps are important for quality assurance purposes. Since the presented exercise generation is otherwise rule-based, the implementation does not transfer across `learning targets` and languages. It can, however, take pedagogical aspects into account which makes automatically generated exercises suitable for school settings.

In order to answer RQ 10, we tested an approach using learning analytics to continually improve the exercise generation. According to this approach, learning goals for which exercises are generated and exercise type specific parameters such as the distractors generated for Gap-filling exercises or chunking of Jumbled Sentence exercises should be reviewed on a regular basis. We showed how learning analytics based on learner errors can support such reviews and improve exercise generation over their manual creation.

Chapter 7

Conclusion and Future Work

This dissertation has extensively explored the relevance and feasibility of considering linguistic variability in ILTSs employed in school contexts. Revisiting our research questions from the beginning, in Figure 7.1 we are now able to accept or reject the previously uncertain nodes and connections that were marked in gray in Figure 1.2. As we can see, we do not reject any of our assumptions altogether. Connections originating from a research question marked in orange are, however, subject to certain conditions.

Although transfer of linguistic knowledge across linguistic forms benefits from varied practice, this kind of transfer is mostly observed on the lexical and semantic level, whereas syntactic variations are often limited in number and thus do not provide variation at an extent that would allow transfer across the variants (RQ 3). Varied practice has benefits for all learners as we could confirm that it fosters variedness in learner productions (RQ 4). Still, linguistic variability in practice material needs to be manipulated with caution as we saw that it impacts practice difficulty (RQ 6). In addition, varied practice seems to strengthen declarative knowledge but slow down proceduralization. In order to balance these competing effects of varied practice, we suggest to keep within-exercise variability low and across-exercise variability high for low-proficient learners. For more proficient learners, increasing within- as well as across-exercise variability fosters varied language productions. Since, however, we also confirmed that linguistic variability – inherent (RQ 7), deliberate (RQ 5), as well as across-exercise variability (RQ 6) – impacts exercise difficulty, adaptive exercise

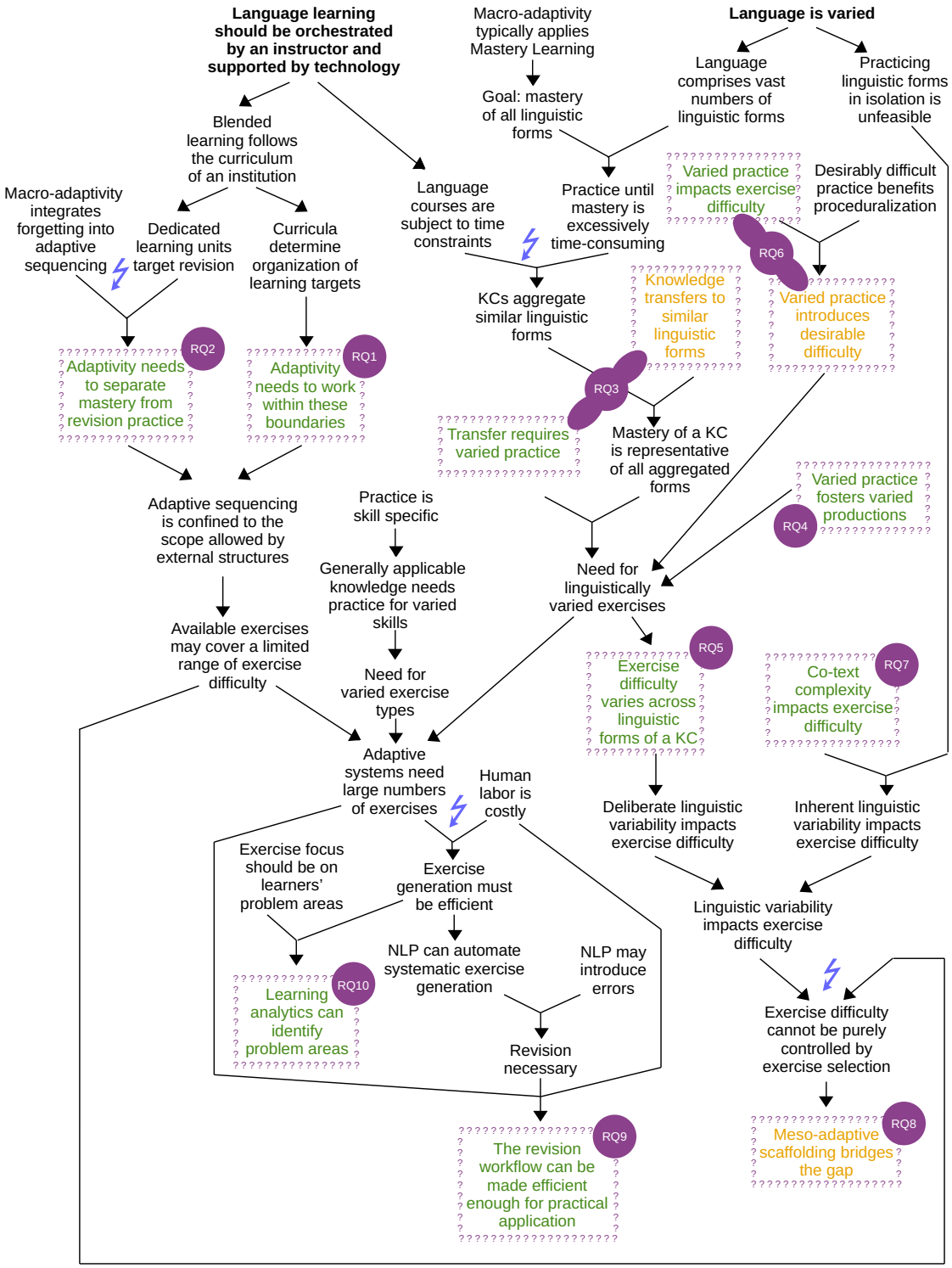


Figure 7.1: Research questions revisited.
Research questions that were answered in the affirmative are colored in green. Orange coloring indicates that the assumption on which the RQ was based only holds under certain conditions.

selection must find means to balance linguistic variability and exercise difficulty. Meso-adaptivity, which dynamically and successively adjusts a selected exercise, can mitigate exercise difficulty in order to support macro-adaptive exercise selection. Yet, such dynamic scaffolding is only beneficial for a sub-population of learners who receive meso-adaptive adjustments in closed exercises when other parameters increase practice difficulty or in half-open exercises after having become familiar with the meso-adaptive adjustments. Very low or rather high C-test scores as well as previous knowledge in combination with a grammar topic dependent extent of the treatment may also indicate that learners would benefit from meso-adaptivity (RQ 8). On the other hand, we found that learner abilities do not differ so much within the same class level and school type to result in significant differences in performance on practice exercises. It is only necessary to control exercise difficulty when for instance including other school types. This in turn allows macro-adaptive exercise selection to primarily focus on providing suitably varied exercises. Macro-adaptivity is, however, constrained by the curriculum of the school setting, which not only reduces the scope within which exercises can be selected (RQ 1), but also introduces the need to separate revision practice from introductory practice (RQ 2).

In order to have large numbers of systematically varied exercises suitable for different school types, it is necessary to automatically generate them. We have illustrated how automatic exercise generation can on the one hand yield the quality required in school contexts if it is paired with our efficient review process (RQ 9) and on the other hand be continuously improved through learning analytics informed by learner errors (RQ 10).

We have thus covered the central aspects of considering linguistic variability in an ILTS in school settings, from assessing its benefits to creating suitable practice material and integrating it into macro-adaptive exercise selection. This opens up vast potential for **future research**. Extending the assessment of levels of variability that are beneficial for certain learner populations to additional `learning targets`, for one, constitutes a crucial task in order to refine the domain model. We have already made an effort to consider multiple levels of linguistic variability by looking into syntax, lexis, and also tapping into semantics. A more comprehensive investigation of such linguistic levels, also encompassing morphology and pragmatic distinctions, can yield further insights.

Additional `learning targets` should also be supported by automatic and systematic exercise generation. In this line of work, a more generic approach would facilitate the extension to additional `learning targets`. The user interface might be even more helpful for teachers if it is improved in terms of design and maintainability and integrated into the ILTS, thus also allowing teachers to use it and also to share the created exercises with other teachers. Quality control might in such a scenario be possible through peer-review processes.

Concerning meso-adaptivity, more learners might benefit from this extensive scaffolding if some of its parameters are adjusted, for example the parameter of what triggers scaffolding. Adaptively adjusting feedback and exercise instructions might also be beneficial for some learners. This, however, needs to be explored further. A good starting point might be to include feedback and instructions in technical and simple versions of the target language, as well as translations into the learner's native language, and example-based feedback. The latter kind of feedback and instructions also opens up further research potential in terms of automatic generation thereof. Whether there is a global complexity associated with each meso-adaptive variant or whether their difficulty is specific to the learner, implying that the variant selection should be informed by the learner model, needs further research as well.

We acknowledge that our system so far covers only one of the four strands introduced by Nation (2007), namely language-focused learning. It should be kept in mind that the system is deliberately intended to be used as supplement to classroom teaching as the use of technology as stand-alone tools for language learning is questionable (Karasimos, 2022) and does not represent learners' preferences (Oh and Reeves, 2013). The ILTS should thus by no means constitute the single mode of practice. The goal is to relieve the teacher of one of the strands where ILTSs can provide individualized material and support beyond what is possible in a classroom setting. Digital systems such as language-aware search engines (e.g., FLAIR; Chinkina and Meurers (2016)) or conversational agents (e.g., Aisla; Bear and Chen (2023)) target the remaining strands of learning through listening and reading, learning through speaking and writing, and becoming fluent in listening, speaking, reading and writing. It would, in principle, be possible to extend our system with additional components such as reading and audio material, more open tasks, and a chat functionality. However, allowing the teacher to orchestrate the instruction makes it

possible to integrate multiple tools into a language class so that no single tool needs to be able to cover all four strands. Instead, each system can optimize those aspects of learning on which they focus. The teacher, in the meantime, can and should on the one hand make sure that all four strands are targeted in equal amounts of time, and on the other hand cover any potentially remaining aspects in class. Since all tools should use the same learner model, a common interface is needed for all tools that are used in a language class (Robson and Barr, 2013).

On a concluding note, we stated in Chapter 1 that ILTSs provide equal opportunities to all learners. This does by no means imply that all learners will indeed eventually be at the same proficiency level. Bloom (1968) stresses the importance of perseverance for successful Mastery Learning. ITSs do, however, provide opportunities for learners to engage with appropriate learning material for the amount of time necessary for them to – in approximately 90% of all cases (Bloom, 1968) – reach mastery. Whether or not they do remains outside the power of the ILTS.

Publications

The following listing comprises the publications on which this thesis is based.

Publication 1: Heck, Tanja and Meurers, Detmar. 2022. Generating and authoring high-variability exercises from authentic texts. In D. Alfter, E. Volodina, T. François, P. Desmet, F. Cornillie, A. Jönsson, and E. Rennes (Eds.), *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, NLP4CALL'22, pages 61–71. Louvain-la-Neuve, Belgium. LiU Electronic Press. https://doi.org/10.3384/ecp190007.

Publication 2: Heck, Tanja and Meurers, Detmar and Nuxoll, Florian. 2022. Automatic exercise generation to support macro-adaptivity in intelligent language tutoring systems. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Fririksdóttir, H. Gararsdóttir, S. Thouësny, and M. J. Whelpton (Eds.), *Intelligent CALL, granular systems and learner data - Short Papers*, EuroCALL'22, pages 162–167. Reykjavik, Iceland (Online). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1452.

Publication 3: Heck, Tanja and Meurers, Detmar. 2023. On the relevance and learner dependence of co-text complexity for exercise difficulty. In D. Alfter, E. Volodina, T. François, A. Jönsson, and E. Rennes (Eds.), *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, NLP4CALL'23, pages 71–84. Tórshavn, Faroe Islands. LiU Electronic Press. https://aclanthology.org/2023.nlp4call-1.9.

Publication 4: Heck, Tanja and Meurers, Detmar. 2023. Exercise generation supporting adaptivity in intelligent tutoring systems. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, and O. C. Santos (Eds.), *Proceedings of the 24th International Conference on Artificial Intelligence in Education - Posters and*

Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, AIED'23, pages 659–665. Tokyo, Japan. Springer Nature. https://doi.org/10.1007/978-3-031-36336-8102.

Publication 5: Colling, Leona and Heck, Tanja and Meurers, Detmar. 2023. Reconciling adaptivity and task orientation in the student dashboard of an intelligent language tutoring system. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'23, pages 288–299. Toronto, Ontario, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.25.

Publication 6: Heck, Tanja and Meurers, Detmar. 2023. Using learning analytics for adaptive exercise generation. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z, Yuan, and T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'23, pages 44–56. Toronto, Ontario, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.4.

Publication 7: Heck, Tanja and Meurers, Detmar. 2023. Exercise parameters influencing exercise difficulty. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Fririksdóttir, H. Gararsdóttir, S. Thouësny, and M. J. Whelpton (Eds.), *CALL for all Languages - Short Papers*, EuroCALL'23, pages 162–167. Reykjavik, Iceland. Editorial Universitat Politécnica de Valéncia. https://doi.org/10.4995/eurocall2023.2023.16921.

Three of the publications were presented as posters at conferences. These posters are included hereafter in the following order.

1) The first poster was used to present Publication 4 at the AIED conference in 2023.

2) The second poster was used to present Publication 5 at the BEA conference in 2022.

3) The third poster was used to present Publication 6 at the BEA conference in 2023.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Faculty of Humanities
University of Tübingen, Department of Linguistics
icall-research.de group

Exercise Generation Supporting Adaptivity in Intelligent Tutoring Systems

Tanja Heck and Detmar Meurers

Macro-Adaptivity

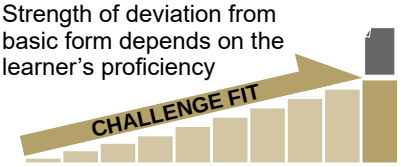
exercises

Learning target & skill
- Same learning target until mastered
- Rotated skills

Semantics
- Topic of the text fits into curriculum
- Co-text complexity matches learner's language proficiency

Syntactic context
- Varied contexts
- Strength of deviation from basic syntax depends on learner's proficiency

Practiced form
Strength of deviation from basic form depends on the learner's proficiency

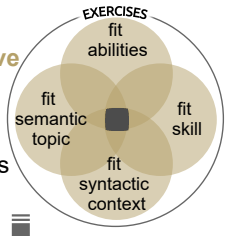


Goal

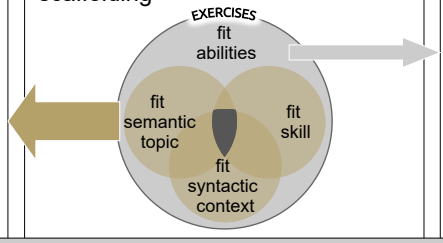
We want to expose learners to ...
- **semantic** topics of a curriculum [1]
- varied **syntactic contexts** [2]
- varied **skills** [3]
- exercises at their ability **level** [4]

This ensures that ...
- practice supports classroom teaching
- knowledge transfers & generalizes
- learners are motivated
- learning outcomes are maximized

Problem:
Macro-adaptive selection criteria reduce the amount of fitting exercises



Solution:
Replace **macro-adaptive** filter for ability fit through **micro-adaptive** scaffolding



Micro-Adaptivity

Adjustment of feedback
domain-specific English → simple language / native language

Adjustment of instructions
unspecific English → specific cues / native language

Adjustment of exercise elements

be, do, have, make, play
The dogs ____ happy.
Children often ____ with them.
They ____ lots of toys.

remove distractor lemma

be, have, play
The dogs ____ happy.
Children often ____ with them.
They ____ lots of toys.

put lemmas behind correct blank

The dogs ____ (*be*) happy.
Children often ____ (*play*) with them.
They ____ (*have*) lots of toys.

give answer & grammatical distractors as dropdown

The dogs *beed / were / was* happy.
Children often *plaid / played / playd* with them.
They *has / haved / had* lots of toys.

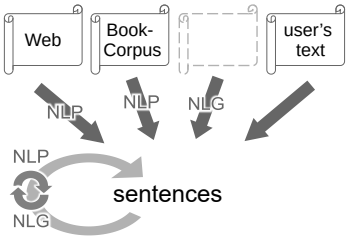
The dogs *beed / were* happy.
Children often *plaid / played* with them.
They *haved / had* lots of toys.

reduce number of distractors

➢ **Balance multiple aspects of adaptivity (exercise complexity, variability, semantics) through a combination of macro- and micro-adaptivity**

Exercise Generation

1) Assembly of sentences



Varied syntax & semantics

2) Generation of exercise material

Chunking
→ vary constituent sizes
The dogs | were | happy.

Hints in parentheses
→ vary meta-information
The dogs ___ (be) happy.

Feedback
→ vary text
The dogs *beed* (be) happy. — 'be' is irregular.

Distractors
→ vary misconceptions
The dogs *beed / were / was* happy.

Instructions
→ vary text
Find the 2 verbs.
The dogs ate and slept.

Micro-adaptive variants

3) Generation of exercise instances

Memory
→ language use

Gap-filling
→ formation
→ semantic parsing
xxx _____ xxx

Jumbled Sentences
→ word order

Mark-the-Words
→ metalinguistic knowledge
xxx xxx xxx xxx

Varied skills

References

[1] Grellet, F., Francoise, G.: Developing Reading Skills: A Practical Guide to Reading Comprehension Exercises. Cambridge University Press (1981)
[2] Kim, T., Wright, D.L., Feng, W.: Commentary: Variability of practice, information processing, and decision making—how much do we know? Front. Psychol. **12** (2021)
[3] Rosell-Aguilar, F.: Top of the Pods — In Search of a Podcasting "Podagogy" for Language Learning. Comput. Assist. Lang. L. **20** (12 2007)
[4] Pelánek, R., Řihák, J.: Experimental Analysis of Mastery Learning Criteria. In: Proceedings of UMAP'17. pp. 156–163. UMAP '17, ACM, New York, USA (2017)

University of Tübingen · Department of Linguistics, Keplerstraße 2· 72074 Tübingen · Germany
{tanja.heck,detmar.meurers}@uni-tuebingen.de

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

LEAD Graduate School & Research Network
University of Tübingen, Department of Linguistics
ICALL Research Group (http://icall-research.de)

Reconciling Adaptivity and Task Orientation in the Student Dashboard of an Intelligent Language Tutoring System

Leona Colling, Tanja Heck and Detmar Meurers

GOAL

Adaptivity (Slavuj et al., 2017):
- Personalize learning to fit individual differences and preferences using:
  - Macro-adaptive exercise selection → no predefined exercise sequences
  - Micro-adaptive exercise adjustments → no static exercise content

Task orientation (Ellis, 2016):
- Practice as preparation for communicative task, functionally relevant language use

→ connect form-focused practice to overarching functional goal to ensure motivation

Ecologically valid setting:
- 7th grade classrooms
- English language learning
- Blended learning with ILTS

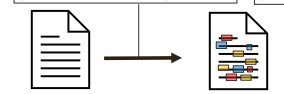
Requirements for a student dashboard in ILTS:
- Learning material needs to fit the curriculum while supporting individual differences
- Comprehensible, target group tailored metrics to enable learners to monitor their progress and performance
- Teachers / students need certain control to be able to:
  - Assign/do homework
  - Discuss exercises in plenary
  - Compare metrics
  - Choose topic to be practiced
  - (Self-)regulate learning

STUDENT DASHBOARD (co-designed with teachers)

Learning content structure:
- Teachers define a functional communicative GOAL for the learning unit → link practice with purpose
- Teachers self-assemble LEARNING TARGETS, skills to be acquired in order to accomplish functional goal, into their learning unit → fit textbook / curriculum
- Each learning target is subdivided in its (pedagogically-defined) language REALIZATIONS → acquisition stepping stones

Segmentation, POS-tagging, dependency parsing, lemmatization, morphological analysis (Rudzewitz et al., 2018)

Rule-based approach to automatically derive metadata annotations incl. linguistic constructions (Quixal et al., 2021)

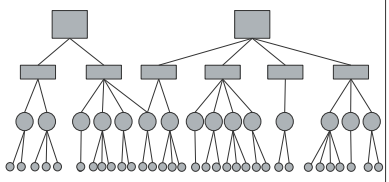


Learning material (exercise text, target answers, instructions)

Annotated learning material

Hierarchy levels	Examples
Learning Targets	Simple Past
Realizations	Affirmative statements
Linguistic Properties	Regular verb forms
Linguistic Constructions	Regular verb forms with infinitive ending in -y

Domain model



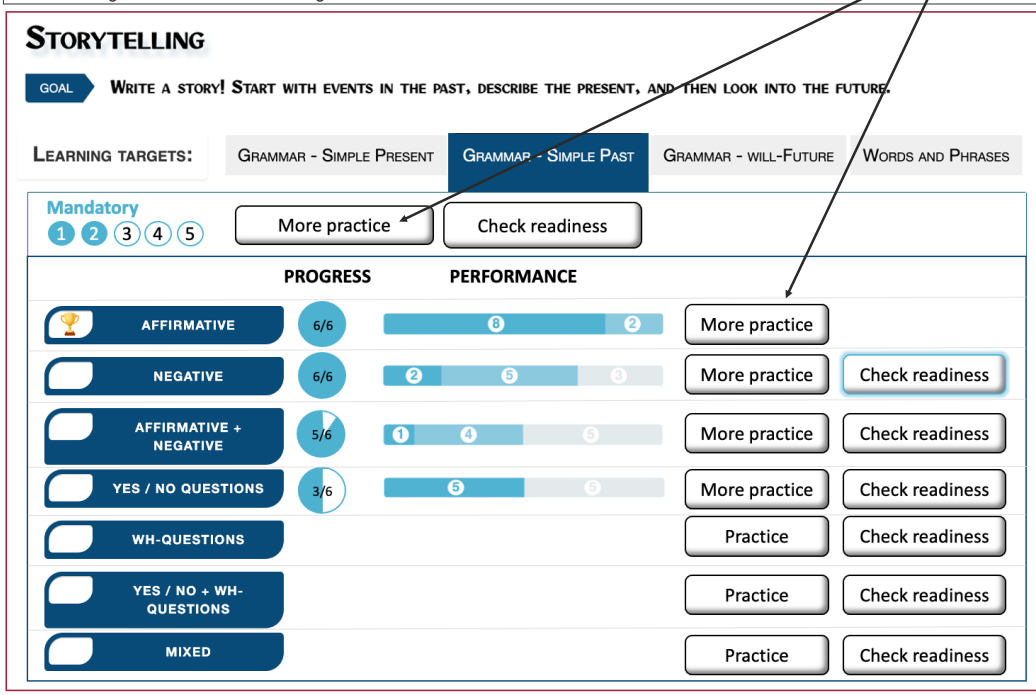
Linking of learning material and learning targets via linguistic annotations → ensure variable and curricular anchored practice

Navigation:
- Options for different navigational preferences: Learners can vary the scope of the adaptive algorithm (and thus the pool of exercise candidates)

→ incorporating less/more learner control

→ flexible integration into different teaching scenarios

System Type	Responsible for choosing		
	Learning target	Realization within learning target	Specific exercise
Non-adaptive systems	USER	USER	USER
Adaptive systems	USER	SYSTEM	SYSTEM
Our approach	USER	SYSTEM / USER	System / USER (for mandatory exercises)



Mandatory exercises:

1 2 3 4 5

- Exercises (selected by the teacher) to be completed by all students
- Offer possibility to discuss exercises in class
- Automatically adaptively sequenced OR explicitly accessed by the student

Mastery:
- Assessed via specific exercises taking only 'first try' attempts into account

Check readiness

- Parallel exercises to re-try
- Completed progress → predicted mastery
- Trophies "gather dust" → incorporate forgetting over time

Active mastery — *Potential forgetting*

Progress:
- Pie chart with discrete numbers of acquired linguistic properties
- Property acquisition is based an accuracy threshold and weighted student attempts on exercise items that carry annotations of the respective property

5/6

Performance:
- Based on 10 most recent items
- Stacked bar charts, that are easy to understand and compare, with discrete numbers of items

1 correct at first try
1 correct after feedback
1 incorrect or not attempted

REFERENCES

Rod Ellis. 2016. Focus on form: A critical review. Language Teaching Research, 20(3):405–428.
Martí Quixal, Björn Rudzewitz, Elizabeth Bear, and Detmar Meurers. 2021. Automatic annotation of curricular language targets to enrich activity models and support both pedagogy and adaptive systems. In *Proceedings of NLP4CALL'21*, 15– 27.
Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, Verena Möller, Florian Nuxoll, and Detmar Meurers. 2018. Generating feedback for English foreign language exercises. In *Proceedings of BEA'18*, pages 127–136.
Vanja Slavuj, Ana Meštrović, and Božidar Kovačić. 2017. Adaptivity in educational systems for language learning: a review. *Computer Assisted Language Learning*, 30(1-2):64–90.

Universität Tübingen · Department of Linguistics · Keplerstraße 2 · 72074 Tübingen · Germany
{leona.colling,tanja.heck,detmar.meurers}@uni-tuebingen.de

metavorhaben digitalisierung im bildungsbereich
Bundesministerium für Bildung und Forschung
aim INNOVATIVE BILDUNG
LEAD

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Faculty of Humanities
University of Tübingen, Department of Linguistics
icall-research.de group

Using Learning Analytics for Adaptive Exercise Generation

Tanja Heck and Detmar Meurers

Exercise Generation

Exercises need to accommodate a learner's proficiency [1]:

receptive ⟶ **productive**

input enhancement → closed exercises → half-open exercises → open exercises

- closed exercises: •Single Choice •Jumbled Sentences
- half-open exercises: •Fill-in-the-Blanks •Short Answers

simple ⟶ **complex**

Challenges:
- **Learning goals** must target a learner's weaknesses
- Single Choice **distractors** must target a learner's misconceptions but not introduce new ones [2]
- Jumbled Sentences must **separate constituents** that learners mix up

Learning Analytics (LA)
based on learner errors from half-open exercises

Randomized controlled field study [3,4]

199 learners
→ German 7th graders
→ English classrooms

1 school year

132 exercises
→ 68 half-open
→ 64 closed

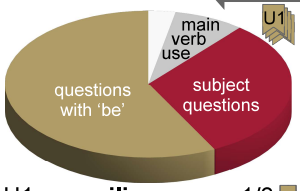
4,836 incorrect answers

Goals:
- Exercises target **learning goals** relevant to the learner
- Single Choice exercises provide plausible & reliable **distractors**
- Jumbled Sentences incorporate **chunking** that fosters learning

Approach

Learning goals

all errors



U1 → **auxiliary use** 1/3
U2 → **main verb use** 2/3

re-ranking

1. main verb use
2. auxiliary use

Complete the questions in the simple past.

1. At what time *came Peter* U1 (Peter, come) to work?
2. *Did the mother kissed* U2 (the mother, kiss) her baby?
3. Why *did Mary stay* (Mary, stay) away for so long?
4. To what city *you did went* W1 U2 (you, go) last year?

Distractors

usage errors U1

U1 *missing auxiliary*

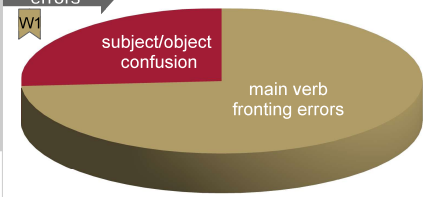
1. __ kissed the mother
2. __ came Peter
3. __ stayed Mary
4. __ went you

U2 *main verb inflection*

did the mother kissed
did Peter came
did Mary stayed
did you went

Chunking

word order errors W1



1. Did | the mother | kiss her baby?
2. At what time | did | Peter | come to work?
3. Why | did | Mary | stay away for so long?
4. To what city | did | you | go last year?

Research Questions and Our Answers

	1) Can the creation of LGs, distractors, and chunking be automated through LA?	2) Does human perception of relevant misconceptions align with relevant misconceptions derived from LA?	3) Do errors made in half-open exercises constitute plausible distractors of closed exercises?
Learning goals (LG)	•**Manually** determine relevant LGs •**Automatically** annotate & rank learner errors corresponding to LGs	•Humans put focus on **different** LGs than LA •**Humans**' focus is **inconsistent**	*not applicable*
Distractors	•Dynamic distractor selection requires error diagnosis to **diagnose all errors** corresponding to LGs	•High overlap between humans' & LA-based distractors for LGs •**Additional human distractors** not included in LGs	•Selected distractors **appear** as errors in half-open exercises •Inconclusive (data sparseness)
Chunking	•Dynamic chunking requires error diagnosis to distinguish **different word order errors**	•Humans' chunking does not cover all LGs •Humans introduce **additional chunking** not relevant to any LG	•Errors occur **mostly in half-open** exercises, not in Jumbled Sentences → in combination with formation errors → isolated word order practice for **proceduralization**

➢ **Learning Analytics can inform generation and dynamic adaptation of exercises**

References

[1] Shabani, K., Khatib, M., & Ebadi, S. (2010). Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. English language teaching, 3(4), 237-248.
[2] Yamada, M. (2019). Language learners and non-professional translators as users. In The Routledge handbook of translation and technology (pp. 183-199). Routledge.
[3] Parrisius, C., Pieronczyk, I., Blume, C., Wendebourg, K., Pili-Moss, D., Assmann, M., Beilharz, S., Bodnar, S., Colling, L., Holz, H., Middelanis, L., Nuxoll, F., Schmidt-Peterson, J., Meurers, D. Nagengast, B., Schmidt, T., & Trautwein, U. (2022). Using an Intelligent Tutoring System within a Task-Based Learning Approach in English as a Foreign Language Classes to Foster Motivation and Learning Outcome (Interact4School): Pre-registration of the Study Design.
[4] Parrisius, C., Wendebourg, K., Rieger, S., Loll, I., Pili-Moss, D., Colling, L., Blume, C., Pieronczyk, I., Holz, H. Bodnar, S., Schmidt, T., Trautwein, U., Meurers, D., & Nagengast, B. (2022). Effective Features of Feedback in an Intelligent Tutoring System-A Randomized Controlled Field Trial (Pre-Registration).

University of Tübingen · Department of Linguistics, Keplerstraße 2· 72074 Tübingen · Germany {tanja.heck,detmar.meurers}@uni-tuebingen.de

Bibliography

Abdi, Solmaz and Khosravi, Hassan and Sadiq, Shazia and Gasevic, Dragan. 2019. A multivariate Elo-based learner model for adaptive educational systems. In Michel C. Desmarais, Collin F. Lynch, Agathe Merceron, and Roger Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining*, EDM'19. Montréal, Quebec, Canada. International Educational Data Mining Society (IEDMS). doi:10.48550/arXiv.1910.12581.

Abraham, Roberta G. and Chapelle, Carol A. 1992. The meaning of cloze test scores: An item difficulty perspective. *Modern Language Journal*, 76(4):468–479. doi:10.2307/330047.

Aftab, Jaweria. 2016. Teachers' beliefs about differentiated instructions in mixed ability classrooms: A case of time limitation. *Journal of Education and Educational Development*, 2(2):94–114. doi:10.22555/joeed.v2i2.441.

Al-Jadaa, Ali A. and Abu-Issa, Abdallatif S. and Ghanem, Wasel T. and Hussein, Mohammed S. 2017. Enhancing the intelligence of web tutoring systems using a multi-entry based open learner model. In *Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing*, ICC'17. Cambridge, England, UK. Association for Computing Machinery. doi:10.1145/3018896.3036389.

Al-Mughairi, Habiba and Bhaskar, Preeti. 2024. Exploring the factors affecting the adoption AI techniques in higher education: insights from teachers' perspectives on ChatGPT. *Journal of Research in Innovative Teaching & Learning*. doi:10.1108/JRIT-09-2023-0129.

Aldabe, Itziar and Lopez de Lacalle, Maddalen and Maritxalar, Montse and Martínez, Edurne and Uria, Larraitz. 2006. ArikIturri: An automatic question generator based on corpora and NLP techniques. In M. Ikeda, K. Ashley, and T.-W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, ITS'06, pages 584–594. Jhongli, Taiwan. Springer Nature. doi:10.1007/11774303_58.

Aldabe, Itziar and Maritxalar, Montse. 2014. Semantic similarity measures for the generation of science tests in Basque. *IEEE Transactions on Learning Technologies*, 7(4):375–387. doi:10.1109/TLT.2014.2355831.

Aldabe, Itziar and Maritxalar, Montse and Martinez, Edurne. 2007. Evaluating and improving the distractor-generating heuristics. In N. Ezeiza, M. Maritxalar, and M. Schulze (Eds.), *Workshop on NLP for Educational Resources. In conjuction with RANLP07*, RANLP'07, pages 7–13. Borovetz, Bulgaria.

Aldabe, Itziar and Maritxalar, Montse and Mitkov, Ruslan. 2009. A study on the automatic selection of candidate sentences distractors. In Vania Dimitrova, Riichiro Mizoguchi, du Benedict Boulay, and Art Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, volume 200 of *AIED'09*, pages 656–658. Brighton, England, UK. IOS Press. doi:10.3233/978-1-60750-028-5-656.

Aleven, Vincent and Mclaren, Bruce and Roll, Ido and Koedinger, Kenneth. 2006. Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*, 16(2):101–128.

Aleven, Vincent and Mclaughlin, Elizabeth A. and Glenn, R. Amos and Koedinger, Kenneth R. 2016. Instruction based on adaptive learning technologies. In R. E. Mayer and P. Alexander (Eds.), *Handbook of Research on Learning and Instruction*, chapter 24, pages 522–560. doi:10.4324/9781315736419.

Almeida, J. João and Grande, Eliana and Smirnov, Georgi. 2017. Exercise generation on language specification. In Álvaro Rocha, Ana Maria Correia, Hojjat Adeli, Luís Paulo Reis, and Sandra Costanzo (Eds.), *Proceedings of the 2017 World Conference on Information Systems and Technologies - Recent Advances in Information Systems and Technologies*, WorldCIST'17, pages 277–286. Porto Santo Island, Portugal. Springer Nature. doi:10.1007/978-3-319-56535-4_28.

Alsubait, Tahani and Parsia, Bijan and Sattler, Uli. 2015. Generating multiple choice questions from ontologies: How far can we go? In Patrick Lambrix, Eero Hyvönen, Eva Blomqvist, Valentina Presutti, Uli Qi, Guilinand Sattler, Ying Ding, and Chiara Ghidini (Eds.), *Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management*, EKAW'14, pages 66–79. Linköping, Sweden. Springer Nature. doi:10.1007/978-3-319-17966-7_7.

Amaral, Luiz A. and Meurers, Detmar. 2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *The Journal of the European Association for Computer Assisted Language Learning*, 23(1):4–24. doi:10.1017/S0958344010000261.

Amoia, Marilisa and Bretaudiere, Treveur and Denis, Alexandre and Gardent, Claire and Perez-Beltrachini, Laura. 2012. A serious game for second language acquisition in a virtual environment. *Journal on Systemics, Cybernetics and Informatics*, 10(1):24–34.

Anderson, John R. 1983. Knowledge compilation : The general learning mechanism. In J. R. Anderson, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*, volume 2 of *Machine Learning*, pages 289–369.

Antal, Margit. 2013. On the use of Elo rating for adaptive assessment. *Studia Universitatis Babes-Bolyai, Informatica*, 58(1):29–41.

Antoniadis, Georges and Echinard, Sandra and Kraif, Olivier and Lebarbé, Thomas and Loiseau, Mathieu and Ponton, Claude. 2004. NLP-based scripting for CALL activities. In *Proceedings of the Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning*, eLearn'04, pages 18–25. Geneva, Switzerland. COLING. doi:10.3115/1610028.1610031.

Apfelbaum, Keith and Brown, Carolyn and Zimmermann, Jerry. 2019. Foundations learning system: The varied practice model. Technical report, Foundations in Learning.

Applefield, James. M. and Huber, Richard and Moallem, Mahnaz. 2000. Constructivism in theory and practice: Toward a better understanding. *The High School Journal*, 84(2):35–53.

Araki, Jun and Rajagopal, Dheeraj and Sankaranarayanan, Sreecharan and Holm, Susan and Yamakawa, Yukari and Mitamura, Teruko. 2016. Generating questions and multiple-choice answers using semantic analysis of texts. In Yuji Matsumoto and Rashmi Prasad (Eds.), *Proceedings of the 26th International Conference on Computational Linguistics - Technical Papers*, COLING'16, pages 1125–1136. Osaka, Japan. Coling 2016 Organizing Committee.

Astete, José Antonio Fajardo. 2015. The communicative language approach and the task-based approach. Master's thesis, Universidad del Pacífico, April.

Attali, Yigal and Bar-Hillel, Maya. 2003. Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, 40(2):109–128. doi:10.1111/j.1745-3984.2003.tb01099.x.

Babaii, Esmat and Fatahi-Majd, Mosayeb. 2014. Failed restoration in the C-test: Types, sources, and implications for C-test processing. In Rüdiger Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends*, volume 34 of *Language Testing and Evaluation*, pages 263–276. doi:10.3726/978-3-653-04578-9.

Bachman, Lyle F. 1990. *Fundamental Considerations in Language Testing*. Oxford Applied Linguistics. doi:10.2307/329499.

Baptista, Jorge and Lourenco, Sandra and Mamede, Nuno J. 2016. Automatic generation of exercises on passive transformation in Portuguese. In *Proceedings of the IEEE Congress on Evolutionary Computation*, CEC'16, pages 4965–4972. Vancouver, British Columbia, Canada. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/CEC.2016.7744427.

Barkaoui, Khaled. 2017. Examining repeaters' performance on second language proficiency tests: A review and a call for research. *Language Assessment Quarterly*, 14(4):420–431. doi:10.1080/15434303.2017.1347790.

Barnes, Tiffany and Stamper, John. 2010. Automatic hint generation for logic proof tutoring using historical data. *Educational Technology & Society*, 13(1):3–12.

Bates, Elizabeth and MacWhinney, Brian. 1987. *Competition, Variation, and Language Learning*. 2010.

Bear, Elizabeth and Chen, Xiaobin. 2023. Evaluating a conversational agent for second language learning aligned with the school curriculum. In Ning Wang, Genaro Rebolledo-Mendez, Vania Dimitrova, Noboru Matsuda, and Olga C. Santos (Eds.), *Proceedings of the 24th International Conference on Artificial Intelligence in Education - Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, AIED'23, pages 142–147. Tokyo, Japan. Springer Nature. doi:10.1007/978-3-031-36336-8_22.

Becker, Carmen and Roos, Jana. 2016. An approach to creative speaking activities in the young learners' classroom. *Education Inquiry*, 7(1). doi:10.3402/edui.v7.27613.

Beinborn, Lisa. 2016. *Predicting and manipulating the difficulty of text-completion exercises for language learning*. Ph.D. thesis, Technische Universität Darmstadt.

Beinborn, Lisa and Zesch, Torsten and Gurevych, Iryna. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2:517–530. doi:10.1162/tacl_a_00200.

Beinborn, Lisa and Zesch, Torsten and Gurevych, Iryna. 2015. Candidate evaluation strategies for improved difficulty prediction of language tests. In Joel Tetreault, Jill Burstein, and Claudia Leacock (Eds.), *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'15, pages 1–11. Denver, Colorado, USA. Association for Computational Linguistics. doi:10.3115/v1/W15-0601.

Benedetto, Luca and Cappelli, Andrea and Turrin, Roberto and Cremonesi, Paolo. 2020. R2DE: A NLP approach to estimating IRT parameters of newly generated questions. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge*, LAK'20, pages 412–421. Frankfurt, Germany. Association for Computing Machinery. doi:10.1145/3375462.3375517.

Bennane, Abdellah. 2013. Adaptive educational software by applying reinforcement learning. *Informatics in Education*, 12(1):13–27. doi:10.13140/RG.2.1.3759.2489.

Berens, Florian and Wendebourg, Katharina and Kholin, Mareike and Colling, Leona and Schmidt-Peterson, Julia and Hopp, Manuel and Heck, Tanja and Bodnar, Stephen and El Hefny, Walid and Nuxoll, Florian and Deeg, Christoph and Krey, Katja and Schrader, Josef and Schröter, Hannes and Nagengast, Benjamin and Meurers, Detmar and Trautwein, Ulrich. 2024. AI2Teach: Effectiveness of a teacher training on technology-supported teaching and learning. Registry of Efficacy and Effectiveness Studies (Registry ID: 20365.1v1).

Bhuiyan, Nadia and Baghel, Amit. 2005. An overview of continuous improvement: from the past to the present. *Management Decision*, 43(5):761–771. doi:10.1108/00251740510597761.

Bick, Eckhard. 2000. Live use of corpus data and corpus annotation tools in CALL: Some new developments in VISL. In Henrik Holmboe (Ed.), *Nordic Language Technology, Årbog for Nordisk Sprogteknologisk Forskningsprogram*, volume 2004, pages 171–185.

Bimba, Andrew Thomas and Idris, Norisma and Al-Hunaiyyan, Ahmed and Mahmud, Rohana Binti and Shuib, Nor Liyana Bt Mohd. 2017. Adaptive feedback in computer-based learning environments: A review. *Adaptive Behavior*, 25(5):217–234. doi:10.1177/1059712317727590.

Bjork, Elizabeth Ligon and Bjork, Robert A. 2011. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In Morton Ann Gernsbacher and James R. Pomerantz (Eds.), *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, pages 56–64.

Bliesener, Ulrich. 1998. Foreign language teaching in Germany. *Bildung und Wissenschaft*, pages 2–32.

Block, James H. and Burns, Robert B. 1976. Mastery learning. *Review of Research in Education*, 4:3–49. doi:10.2307/1167112.

Bloom, Benjamin S. 1968. Learning for mastery. In *Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints*, volume 1, pages 10–12.

Brekelmans, Gwen and Lavan, Nadine and Saito, Haruka and Clayards, Meghan and Wonnacott, Elizabeth. 2022. Does high variability training improve the learning of non-native phoneme contrasts over low variability training? a replication. *Journal of Memory and Language*, 126. doi:10.1016/j.jml.2022.104352.

Brophy, Jere E. 1982. How teachers influence what is taught and learned in classrooms. *Elementary School Journal*, 83(1):1–13. doi:10.1086/461287.

Brown, Helen and Gunning, Lydia and Wonnacott, Elizabeth. 2016. Do shared distributional contexts aid learning of Italian gender classes in 7-year-old children? In *Procedings of the 5th Implicit Learning Seminar - Provisional Programme*, ILS'16. Lancaster, England, UK. Lancaster University.

Brown, Helen and Smith, Kenny and Samara, Anna and Wonnacott, Elizabeth. 2022. Semantic cues in language learning: an artificial language study with adult and child learners. *Language, Cognition and Neuroscience*, 37(4):509–531. doi:10.1080/23273798.2021.1995612.

Brown, James Dean. 1989. Cloze item difficulty. *JALT CALL Journal*, 11(1):46–67.

Brown, Jonathan and Frishkoff, Gwen and Eskenazi, Maxine. 2005. Automatic question generation for vocabulary assessment. In Raymond Mooney, Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff (Eds.), *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT'05, pages 819–826. Vancouver, British Columbia, Canada. Association for Computational Linguistics. doi:10.3115/1220575.1220678.

Brown, Tom and Mann, Benjamin and Ryder, Nick and Subbiah, Melanie and Kaplan, Jared D. and Dhariwal, Prafulla and Neelakantan, Arvind and Shyam, Pranav and Sastry, Girish and Askell, Amanda and Agarwal, Sandhini and Herbert-Voss, Ariel and Krueger, Gretchen and Henighan, Tom and Child, Rewon and Ramesh, Aditya and Ziegler, Daniel and Wu, Jeffrey and Winter, Clemens and Hesse, Chris and Chen, Mark and Sigler, Eric and Litwin, Mateusz and Gray, Scott and Chess, Benjamin and Clark, Jack and Berner, Christopher and McCandlish, Sam and Radford, Alec and Sutskever, Ilya and Amodei, Dario. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pages 1877–1901. Vancouver, British Columbia, Canada. Curran Associates, Inc.

Brusilovsky, Peter L. 1992. A framework for intelligent knowledge sequencing and task sequencing. In Claude Frasson, Gilles Gauthier, and Gordon I. McCalla (Eds.), *Proceedings of the 2nd International Conference on Intelligent Tutoring Systems*, ITS'92, pages 499–506. Montréal, Quebec, Canada. Springer Nature. doi:10.1007/3-540-55606-0_59.

Brusilovsky, Peter L. 1994. The construction and application of student models in intelligent tutoring systems. *Journal of Computer and Systems Sciences International*, 32(1):70–89.

Brusilovsky, Peter L. 1999. Adaptive and intelligent technologies for Web-based eduction. *Künstliche Intelligenz*, 4(Special Issue on Intelligent Systems and Teleteaching):19–25.

Brusilovsky, Peter L. 2000. Adaptive hypermedia: From intelligent tutoring systems to Web-based education. In Gilles Gauthier, Claude Frasson, and Kurt VanLehn (Eds.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, ITS'00, pages 1–7. Montréal, Quebec, Canada. Springer Nature. doi:10.1007/3-540-45108-0_1.

Brusilovsky, Peter L. and Millán, Eva. 2007. User models for adaptive hypermedia and adaptive educational systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.), *The Adaptive Web. Lecture Notes in Computer Science*, volume 4321, pages 3–53. doi:10.1007/978-3-540-72079-9_1.

Bull, Susan and Ginon, Blandine and Boscolo, Clelia and Johnson, Matthew. 2016. Introduction of learning visualisations and metacognitive support in a persuadable open learner model. In *Proceedings of the 6th International Conference on Learning Analytics and Knowledge*, LAK'16, pages 30–39. Edinburgh, Scotland, UK. Association for Computing Machinery. doi:10.1145/2883851.2883853.

Bulut, Okan and Shin, Jinnie and Yildirim-Erbasli, Seyma N. and Gorgun, Guher and Pardos, Zachary A. 2023. An introduction to Bayesian Knowledge Tracing with pyBKT. *Psych*, 5(3):770–786. doi:10.3390/psych5030050.

Burns, Anne and Siegel, Joseph. 2018. *International Perspectives on Teaching the Four Skills in ELT: Listening, Speaking, Reading, Writing*. doi:10.1007/978-3-319-63444-9.

Burns, Hugh L. and Capps, Charles G. 1988. Foundations of intelligent tutoring systems : An introduction. In Martha C. Polson and J. Jeffrey Richardson (Eds.), *Foundations of Intelligent Tutoring Systems*, Interacting with Computers Series, pages 1–19. doi:10.4324/9780203761557.

Burstein, Jill and Shore, Jane and Sabatini, John and Moulder, Brad and Holtzman, Steven and Pedersen, Ted. 2012. The Language Muse system: Linguistically focused instructional authoring. *ETS Research Report Series*, 2012(2):i–36. doi:10.1002/j.2333-8504.2012.tb02303.x.

Burton, Graham. 2022. Grammar Syllabus. In *The TESOL Encyclopedia of English Language Teaching*, pages 1–6. doi:10.1002/9781118784235.eelt1013.

Burton, Richard R. and Brown, John Seely. 1979. An investigation of computer coaching for informal learning activities. *International Journal of Man-Machine Studies*, 11(1):5–24. doi:10.1016/S0020-7373(79)80003-6.

Bybee, Joan. 2008. Usage-based grammar and second language acquisition. In Peter Robinson and Nick C. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*, chapter 10. doi:10.4324/9780203938560-18. 2010.

Carpenter, Shana and Mueller, Frank. 2013. The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, 41(5):671–682. doi:10.3758/s13421-012-0291-4.

Carrell, Patricia L. 1984. Schema theory and ESL reading: Classroom implica-

tions and applications. *Modern Language Journal*, 68(4):332–343. doi:10.1111/j.1540-4781.1984.tb02509.x.

Casenhiser, Devin and Goldberg, Adele E. 2005. Fast mapping between a phrasal form and meaning. *Developmental Science*, 8(6):500–508. doi:10.1111/j.1467-7687.2005.00441.x.

Cen, Hao and Koedinger, Kenneth R. and Junker, Brian. 2006. Learning Factors Analysis – A general method for cognitive model evaluation and improvement. In Mitsuru Ikeda, Kevin D. Ashley, and Tak-Wai Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, volume 4053 of *ITS'06*, pages 164–175. Jhongli, Taiwan. Springer Nature. doi:10.1007/11774303_17.

Chan, Sophia and Somasundaran, Swapna and Ghosh, Debanjan and Zhao, Mengxuan. 2022. AGReE: A system for generating Automated Grammar Reading Exercises. In Wanxiang Che and Ekaterina Shutova (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP'22, pages 169–177. Abu Dhabi, UAE. Association for Computational Linguistics. doi:10.18653/v1/2022.emnlp-demos.17.

Chen, Chia-Yin and Liou, Hsien-Chin and Chang, Jason S. 2006. FAST – An automatic generation system for grammar tests. In James Curran (Ed.), *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics - Interactive Presentation Sessions*, COLING/ACL'06, pages 1–4. Sydney, Australia. Association for Computational Linguistics. doi:10.3115/1225403.1225404.

Chen, Chih-Ming. 2008. Intelligent web-based learning system with personalized learning path guidance. *Computers & Education*, 51(2):787–814. doi:10.1016/j.compedu.2007.08.004.

Chen, Tao and Zheng, Naijia and Zhao, Yue and Chandrasekaran, Muthu Kumar and Kan, Min-Yen. 2015. Interactive second language learning from news websites. In Hsin-Hsi Chen, Yuen-Hsien Tseng, Yuji Matsumoto, and Lung Hsiang Wong (Eds.), *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, NLP-TEA'15, pages 34–42. Beijing, China. Association for Computational Linguistics. doi:10.18653/v1/W15-4406.

Chen, Xiaobin and Meurers, Detmar. 2019. Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, 32(4):418–447. doi:10.1080/09588221.2018.1527358.

Chinkina, Maria and Meurers, Detmar. 2016. Linguistically aware information retrieval: Providing input enrichment for second language learners. In Joel Tetreault, Jill Burstein, Claudia Leacock, and Helen Yannakoudakis (Eds.), *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'16, pages 188–198. San Diego, California, USA. Association for Computational Linguistics. doi:10.18653/v1/W16-0521.

Choi, Inn-Chull. 2016. Efficacy of an ICALL tutoring system and process-oriented corrective feedback. *Computer Assisted Language Learning*, 29(2):334–364. doi:10.1080/09588221.2014.960941.

Chrysafiadi, Konstantina and Troussas, Christos and Virvou, Maria and Sakkopoulos, Evangelos. 2019. ICALM: An intelligent mechanism for the creation of dynamically adaptive learning material. *Sensors and Transducers*, 234(6):22–29.

Chrysafiadi, Konstantina and Virvou, Maria. 2013. Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, 40(11):4715–4729. doi:10.1016/j.eswa.2013.02.007.

Clopper, Cynthia G. and Pisoni, David B. 2004. Effects of talker variability on perceptual learning of dialects. *Language and Speech*, 47(3):207–238. doi:10.1177/00238309040470030101.

Colling, Leona and Pieronczyk, Ines and Parrisius, Cora and Holz, Heiko and Bodnar, Stephen and Nuxoll, Florian and Meurers, Detmar. 2024. Towards task-oriented ICALL: A criterion-referenced learner dashboard organising digital practice. In Oleksandra Poquet, Alejandro Ortega-Arranz, Olga Viberg, Irene Angelica Chounta, Bruce McLaren, and Jelena Jovanovic (Eds.), *Proceedings of the 16th International Conference on Computer Supported Education*, volume 1: EKM of *CSEDU'24*, pages 668–679. Angers, France. Institute for Systems and Technologies of Information, Control and Communication. doi:10.5220/0012753000003693.

Colpaert, Jozef and Sevinc, Emre. 2010. ClozeFox: Gap exercise generator with scalable intelligence for Mozilla Firefox. https://github.com/emres/clozefox. [Online; accessed 08-November-2022].

Coniam, David. 1997. A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal*, 14(2):15–33. doi:10.1558/cj.v14i2-4.15-33.

Corbalan, Gemma and Kester, Liesbeth and van Merriënboer, Jeroen J. G. 2006. Towards a personalized task selection model with shared instructional control. *Instructional Science*, 34(5):399–422. doi:10.1007/s11251-005-5774-2.

Corbett, A. and Koedinger, Kenneth R. and Hadley, W. 2001. Cognitive tutors: From the research classroom to all classrooms. In Paul S. Goodman (Ed.), *Technology Enhanced Learning: Opportunities for Change*, chapter 9, pages 235–263. doi:10.4324/9781410601933-19.

Corder, S. P. 1967. The significance of learner's errors. *International review of applied linguistics in language teaching*, 5(1-4):161–170. doi:10.1515/iral.1967.5.1-4.161.

Correia, Rui and Baptista, Jorge and Eskénazi, Maxine and Mamede, Nuno J. 2012. Automatic generation of cloze question stems. In Helena Caseli, Aline Villavicencio, António Teixeira, and Fernando Perdigão (Eds.), *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*, PROPOR'12, pages 168–178. Coimbra, Portugal. Springer Nature. doi:10.1007/978-3-642-28885-2_19.

Csíkszentmihályi, Mihály. 1991. *Flow: The Psychology of Optimal Experience*.

Cui, Peng and Sachan, Mrinmaya. 2023. Adaptive and personalized exercise generation for online language learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Long Papers*, ACL'23, pages 10184–10198. Toronto, Ontario, Canada. Association for Computational Linguistics. doi:10.48550/arxiv.2306.02457.

Dagger, Declan and Wade, Vincent and Conlan, Owen. 2005. Personalisation for

all: Making adaptive course composition easy. *Journal of Educational Technology and Society*, 8(3):9–25.

Daller, Helmut and Phelan, David. 2006. The C-test and TOEIC ® as measures of students' progress in intensive short courses in EFL. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, 13(2):101–119.

Daller, Michael and Müller, Amanda and Wang-Taylor, Yixin. 2021. The C-test as predictor of the academic success of international students. *International Journal of Bilingual Education and Bilingualism*, 24(10):1502–1511. doi:10.1080/13670050.2020.1747975.

Danenas, Paulius and Skersys, Tomas. 2022. Exploring natural language processing in model-to-model transformations. *IEEE Access*, 10:116942–116958. doi:10.1109/ACCESS.2022.3219455.

De Groot, Annette M. B. and Keijzer, Rineke. 2000. What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, 50(1):1–56. doi:10.1111/0023-8333.00110.

De Kuthy, Kordula and Meurers, Detmar. 2021. Approaches to the analysis and parameterisation of linguistic complexity of texts and tasks. https://www.uni-bamberg.de/fileadmin/germ-lingdaf2/Tagungsdokus/Tagung_Abstracts/DeKuthy.Meurers-21.pdf. [Online; accessed 06-December-2024].

Deininger, Hannah and Parrisius, Cora and Lavelle-Hill, Rosa and Meurers, Detmar and Trautwein, Ulrich and Nagengast, Benjamin and Kasneci, Gjergji. 2025. Who did what to succeed? Individual differences in which learning behaviors are linked to achievement.

Deininger, Hannah and Pieronczyk, Ines and Parrisius, Cora and Plumley, Robert D. and Meurers, Detmar and Kasneci, Gjergji and Nagengast, Benjamin and Trautwein, Ulrich and Greene, Jeffrey A. and Bernacki, Matthew L. 2024. Data preparation and analysis code for the article "Using theory-informed learning analytics to understand how homework behavior predicts achievement". https://psycharchives.org/en/item/6532792d-8f11-4078-b9ef-262b19f56b32. [Online; accessed 06-December-2024].

DeKeyser, Robert M. 1997. Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, 19(2):195–221. doi:10.1017/S0272263197002040.

DeKeyser, Robert M. 2001. Automaticity and automatization. In Peter Robinson (Ed.), *Cognition and Second Language Instruction*, Cambridge Applied Linguistics, chapter 5, pages 125–151. doi:10.1017/CBO9781139524780.007.

DeKeyser, Robert M. 2018. *Learning Language through Task Repetition*, volume 11 of *Task-Based Language Teaching*. doi:10.1075/tblt.11.01dek.

Derkow-Disselbeck, Barbara and Abbey, Susan and Woppert, Allen J. and Harger, Laurence. 2008. *English G 21: Ausgabe A – Band 3: 7. Schuljahr*. English G 21 - Ausgabe A.

Devedžic, Vladan. 2006. *Personalization issues*, Integrated Series in Information Systems, pages 175–220. doi:10.1007/978-0-387-35417-0.

Doignon, Jean-Paul and Falmagne, Jean-Claude. 1985. Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, 23(2):175–196. doi:10.1016/S0020-7373(85)80031-6.

Doroudi, Shayan. 2020. Mastery learning heuristics and their hidden models. In Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán (Eds.), *Proceedings of the 21st International Conference on Artificial Intelligence in Education*, AIED'20, pages 86–91. Ifrane, Morocco. Springer Nature. doi:10.1007/978-3-030-52240-7_16.

Doughty, Catherine and Long, Michael. 2003. Optimal psycholinguistic environments for distance foreign language learning. *Language Learning & Technology*, 7(3):50–80.

Dušková, Libuše. 1969. On sources of errors in foreign language learning. *International Review of Applied Linguistics in Language Teaching*, 7(1):11–36. doi:10.1515/iral.1969.7.1.11.

Ebbinghaus, Hermann. 1885. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie.*

Ebbinghaus, Hermann. 1913. Memory: A contribution to experimental psychology. *Annals of Neurosciences*, 20(4):155. doi:10.1037/10011-000.

Eckes, Thomas. 2011. Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling*, 53(4):414–439.

Effenberger, Tomáš. 2018. Adaptive system for learning programming. Master's thesis, Masaryk University.

Eidsvåg, Sunniva Sørhus and Austad, Margit and Plante, Elena and Asbjørnsen, Arve E. 2015. Input variability facilitates unguided subcategory learning in adults. *Journal of Speech, Language, and Hearing Research*, 58(3):826–839. doi:10.1044/2015_JSLHR-L-14-0172.

Elbers, Ed and Rojas-Drummond, Sylvia and van de Pol, Janneke. 2013. Conceptualising and grounding scaffolding in complex educational contexts. *Learning, Culture and Social Interaction*, 2(1):1–2. doi:10.1016/j.lcsi.2012.12.002.

Ellis, Nick C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2):143–188. doi:10.1017/S0272263102002024.

Ellis, Rod. 1985. A variable competence model of second language acquisition. *International review of applied linguistics in language teaching*, 23(1):47–59.

Ellis, Rod. 2003. *Task-Based Language Learning and Teaching.* Oxford Applied Linguistics.

Ellis, Rod. 2005. Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, 27(2):141–172. doi:10.1017/S0272263105050096.

Ellis, Rod. 2015. Researching acquisition sequences: Idealization and de-idealization in SLA. *Language Learning*, 65(1):181–209. doi:10.1111/lang.12089.

Elshani, Lumbardh. 2021. Personalized learning path generation based on genetic algorithms. Bachelor's thesis, University for Business and Technology in Kosovo. doi:10.33107/ubt-etd.2021.2649.

Ennouamani, Soukaina and Mahani, Zouhir. 2017. An overview of adaptive e-learning systems. In *Proceedings of the 8th International Conference on Intelligent Computing and Information Systems*, ICICIS'17, pages 342–347. Cairo, Egypt. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/INTELCIS.2017.8260060.

Enns, Kevin. 2018. High-variability training in second-language reading instruction. Master's thesis, University of Manitoba.

Erss, Maria. 2018. 'Complete freedom to choose within limits' – teachers' views of curricular autonomy, agency and control in Estonia, Finland and Germany. *The Curriculum Journal*, 29(2):238–256. doi:10.1080/09585176.2018.1445514.

Esichaikul, Vatcharaporn and Lamnoi, S. and Bechter, Clemens. 2011. Student modelling in adaptive e-learning systems. *Knowledge Management & E-Learning: An International Journal*, 3(3):342–355. doi:10.34105/j.kmel.2011.03.025.

Fakharzadeh, Mehrnoosh and Youhanaee, Manijeh and Nejadansari, Daryoush. 2014. Proceduralization, transfer of training and retention of knowledge as a result of output practice. *International Journal of Research Studies in Language Learning*, 3(2):3–16. doi:10.5861/ijrsll.2013.431.

Fang, Jiansheng and Zhao, Wei and Jia, Dongya. 2019. Exercise difficulty prediction in online education systems. In *IEEE International Conference on Data Mining Workshops*, ICDMW'19, pages 311–317. Beijing, China. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/ICDMW.2019.00053.

Farhady, Hossein and Jamali, Ferdos. 2006. Varieties of C-test as measures of general language proficiency. In Hossein Farhady (Ed.), *Twenty-Five Years of Living with Applied Linguistics: Collection of Articles*, pages 287–302.

Feng, Mingyu and Roschelle, Jeremy and Heffernan, Neil T. and Fairman, Janet and Murphy, Robert. 2014. Implementation of an intelligent tutoring system for online homework support in an efficacy trial. In Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, ITS'14, pages 561–566. Honolulu, Hawaii, USA. Springer Nature. doi:10.1007/978-3-319-07221-0_71.

Ferreira, Kledilson and Pereira Jr., Álvaro R. 2018. Verb tense classification and automatic exercise generation. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, WebMedia'18, pages 105–108. Salvador, Brazil. Association for Computing Machinery. doi:10.1145/3243082.3267451.

Finkenbinder, Erwin Oliver. 1913. The curve of forgetting. *American Journal of Psychology*, 24(1):8–32. doi:10.2307/1413271.

Fournier-Viger, Philippe and Nkambou, Roger and Nguifo, Engelbert Mephu. 2010. Building intelligent tutoring systems for ill-defined domains. In Roger Nkambou,

Jacqueline Bourdeau, and Riichiro Mizoguchi (Eds.), *Advances in Intelligent Tutoring Systems*, chapter 5, pages 81–101. doi:10.1007/978-3-642-14363-2_5.

Frehan, Pádraic. 1999. Beyond the sentence: Finding a balance between bottom-up and top-down reading approaches. *JALT CALL Journal*, 23:11–14.

Galasso, Sabrina. 2018. Automated C-test difficulty prediction: Integrating lexical, sentence, and text features in a multi-lingual perspective. Master's thesis, University of Tübingen, Tübingen.

Gao, Lingyu and Gimpel, Kevin and Jensson, Arnar. 2020. Distractor analysis and selection for multiple-choice cloze questions for second-language learners. In Jill Burstein, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch (Eds.), *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'20, pages 102–114. Seattle, Washington, USA. Association for Computational Linguistics. doi:10.18653/v1/2020.bea-1.10.

Gates, Donna Marie. 2011. How to generate cloze questions from definitions: A syntactic approach. In *Papers from the 2011 AAAI Fall Symposium - Question Generation*, volume FS-11-04 of *AAAI-FS'11*, pages 19–22. Arlington, Virginia, USA. Association for the Advancement of Artificial Intelligence (AAAI).

Gavin, Bui Hiu Yuet. 2014. Task readiness: Theoretical framework and empirical evidence from topic familiarity, strategic planning, and proficiency levels. In Peter Skehan (Ed.), *Processing Perspectives on Task Performance*, chapter 3, pages 63–94. doi:10.1075/tblt.5.03gav.

Gierl, Mark J. and Lai, Hollis and Turner, Simon R. 2012. Using automatic item generation to create multiple-choice test items. *Medical Education*, 46(8):757–765. doi:10.1111/j.1365-2923.2012.04289.x.

Goldberg, Adele E. and Casenhiser, Devin M. and Sethuraman, Nitya. 2004. Learning argument structure generalizations. *Cognitive Linguistics*, 15(3):289–316. doi:10.1515/cogl.2004.011.

Gómez, Rebecca L. 2002. Variability and detection of invariant structure. *Psychological Science*, 13(5):431–436. doi:10.1111/1467-9280.00476. PMID: 12219809.

Gooding, Sian and Tragut, Manuel. 2022. One size does not fit all: The case for personalised word complexity models. In Marine Carpuat, de Marie-Catherine Marneffe, and Ivan Vladimir Meza Ruiz (Eds.), *Findings of the Association for Computational Linguistics*, NAACL'22, pages 353–365. Seattle, Washington, USA. Association for Computational Linguistics. doi:10.18653/v1/2022.findings-naacl.27.

Goto, Takuya and Kojiri, Tomoko and Watanabe, Toyohide and Iwata, Tomoharu and Yamada, Takeshi. 2010. Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning: An International Journal*, 2(3):210–224. doi:10.34105/j.kmel.2010.02.016.

Graesser, Arthur C. and Conley, Mark W. and Olney, Andrew. 2012. Intelligent tutoring systems. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, and H. L. Swanson (Eds.), *APA Educational Psychology Handbook*, volume 3: Applications to learning and teaching, pages 451–473. doi:10.1037/13275-018.

Graham, Charles R. 2005. Blended learning systems: Definition, current trends, and future directions. In Curtis J. Bonk and Charles R. Graham (Eds.), *The Handbook of Blended Learning: Global Perspectives, Local Designs*, volume 1 of *Pfeiffer essential resources for training and HR professionals*, chapter 1, pages 3–21.

Grellet, Françoise. 1981. *Developing Reading Skills: A Practical Guide to Reading Comprehension Exercises*. Cambridge Language Teaching Library.

Grubišić, Ani and Stankov, Slavomir and Žitko, Branko. 2013. Stereotype student model for an adaptive e-learning system. *International Journal of Computer and Information Engineering*, 7(4):440–447. doi:10.5281/zenodo.1081417.

Gündüz, Nazlı. 2005. Computer assisted language learning. *Journal of Language and Linguistic Studies*, 1(2):193–214.

Guo, Qi and Kulkarni, Chinmay and Kittur, Aniket and Bigham, Jeffrey P. and Brunskill, Emma. 2016. Questimator: Generating knowledge assessments for arbitrary topics. In Gerhard Brewka (Ed.), *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 3726–3732. New York, New York, USA. Association for the Advancement of Artificial Intelligence (AAAI).

Gutl, Christian and Lankmayr, Klaus and Weinhofer, Joachim and Hofler, Margit. 2011. Enhanced Automatic Question Creator–EAQC: Concept, development and evaluation of an automatic test item creation tool to foster modern e-education. *Electronic Journal of e-Learning*, 9(1):23–38.

Ha, Le An and Yaneva, Victoria and Baldwin, Peter and Mee, Janet. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In Helen Yannakoudakis, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, and Torsten Zesch (Eds.), *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'19, pages 11–20. Florence, Italy. Association for Computational Linguistics. doi:10.18653/v1/W19-4402.

Hafidi, Mohamed and Bensebaa, Tahar. 2014. Developing adaptive and intelligent tutoring systems (AITS): A general framework and its implementations. *International Journal of Information and Communication Technology Education*, 10(4):70–85. doi:10.4018/ijicte.2014100106.

Haladyna, Thomas M. and Downing, Steven M. 1989. Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1):51–78. doi:10.1207/s15324818ame0201_4.

Haladyna, Thomas M. and Downing, Steven M. 1993. How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53(4):999–1010. doi:10.1177/0013164493053004013.

Hanus, Pamela (Ed.). 2014. *Camden Town 3. Textbook. Allgemeine Ausgabe. Gymnasien*. Camden Town.

Hanus, Pamela and Reuter, Christoph and Wauer, Sylvia. 2020. *Camden Town – Allgemeine Ausgabe 2020 Für Gymnasien: Textbook 7*.

Harger, Laurence and Niemitz-Rossant, Cecile J. 2014. *Access – Allgemeine Ausgabe / Band 3: 7. Schuljahr*. English G Access - Allgemeine Ausgabe.

Hartig, Johannes and Frey, Andreas and Nold, Gűnter and Klieme, Eckhard. 2012. An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, 72(4):665–686. doi:10.1177/0013164411430707.

Hartley, James. 1973. The effect of pre-testing on post-test performance. *Instructional Science*, 2(2):193–214. doi:10.1007/BF00139871.

Hasanov, Aziz and Laine, Teemu H. and Chung, Tae-Sun. 2019. A survey of adaptive context-aware learning environments. *Journal of Ambient Intelligence and Smart Environments*, 11(5):403–428. doi:10.3233/AIS-190534.

Hawkins, John A. and Filipović, Luna. 2012. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*, volume 1 of *English Profile Studies*.

Hawkins, William J. and Heffernan, Neil T. and Baker, Ryan S. J. D. 2014. Learning Bayesian Knowledge Tracing parameters with a knowledge heuristic and empirical probabilities. In Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, volume 8474 of *ITS'14*, pages 150–155. Honolulu, Hawaii, USA. Springer Nature. doi:10.1007/978-3-319-07221-0_18.

Heck, Tanja and Meurers, Detmar. 2022. Parametrizable exercise generation from authentic texts: Effectively targeting the language means on the curriculum. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (Eds.), *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'22, pages 154–166. Seattle, Washington, USA. Association for Computational Linguistics. doi:10.18653/v1/2022.bea-1.20.

Heift, Trude. 2001. Error-specific and individualised feedback in a Web-based language tutoring system: Do they read it? *The Journal of the European Association for Computer Assisted Language Learning*, 13(1):99–109. doi:10.1017/S095834400100091X.

Heift, Trude. 2010. Developing an Intelligent Language Tutor. *CALICO Journal*, 27(3):443–459. doi:10.11139/cj.27.3.443-459.

Heift, Trude. 2016. Web delivery of adaptive and interactive language tutoring: Revisited. *International Journal of Artificial Intelligence in Education*, 26(1):489–503. doi:10.1007/s40593-015-0061-0.

Heift, Trude and McFetridge, Paul. 1999. Exploiting the student model to emphasize language teaching pedagogy in natural language processing. In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing*, ASSESSEVALNLP'99, pages 55–61. College Park, Maryland, USA. doi:10.3115/1598834.1598845.

Heift, Trude and Nicholson, Devlan. 2001. Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education*, 12.

Heift, Trude and Rimrott, Anne. 2012. Task-related variation in computer-assisted language learning. *Modern Language Journal*, 96(4):525–543. doi:10.1111/j.1540-4781.2012.01392.x.

Hildebrand, Wilfried and Scheibner-Herzig, Gudrun. 1986. Retention and forgetting of grammatical structures in learning English as a foreign language. *Journal of Experimental Education*, 54(3):149–152. doi:10.1080/00220973.1986.10806413.

Hill, Jennifer and Simha, Rahul. 2016. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. In Joel Tetreault, Jill Burstein, Claudia Leacock, and Helen Yannakoudakis (Eds.), *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'16, pages 23–30. San Diego, California, USA. Association for Computational Linguistics. doi:10.18653/v1/W16-0503.

Hirschmann, Hagen and Lüdeling, Anke and Rehbein, Ines and Reznicek, Marc and Zeldes, Amir. 2013. Underuse of syntactic categories in Falko: A case study on modification. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier (Eds.), *Proceedings of the 1st Learner Corpus Research Conference - Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead*, LCR'11, pages 223–234. Louvain-la-Neuve, Belgium. UCL Presses Universitaires de Louvain.

Hnida, Meriem and Idrissi, Mohammed Khalidi and Bennani, Samir. 2018. Overview of CALEP: A competency based learning path generation system. In *Proceedings of the 17th International Conference on Information Technology Based Higher Education and Training*, ITHET'18, pages 1–6. Olhao, Portugal. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/ITHET.2018.8424772.

Holzknecht, Franz and McCray, Gareth and Eberharter, Kathrin and Kremmel, Benjamin and Zehentner, Matthias and Spiby, Richard and Dunlea, Jamie. 2021. The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing*, 38(1):41–61. doi:10.1177/0265532220917316.

Hoshino, Ayako and Nakagawa, Hiroshi. 2007. A cloze test authoring system and its automation. In Howard Leung, Frederick Li, Rynson Lau, and Qing Li (Eds.), *Proceedings of the 6th International Conference on Web-based Learning - Advances in Web Based Learning*, ICWL'07, pages 252–263. Edinburgh, Scotland, UK. Springer, Springer Nature. doi:10.1007/978-3-540-78139-4_23.

Hoshino, Yuko. 2013. Relationship between types of distractor and difficulty of multiple-choice vocabulary tests in sentential context. *Language Testing in Asia*, 3(1):16–29. doi:10.1186/2229-0443-3-16.

Housen, Alex and Kuiken, Folkert and Vedder, Ineke. 2012. Complexity, accuracy and fluency: Definitions, measurement and research. In Nina Spada and Laura Gurzynski-Weiss (Eds.), *Language Learning & Language Teaching*, chapter 1, pages 1–20. doi:10.1075/lllt.32.01hou.

Hsieh, Tung-Cheng and Lee, Ming-Che and Su, Chien-Yuan. 2013. Designing and implementing a personalized remedial learning system for enhancing the programming learning. *Journal of Educational Technology and Society*, 16(4):32–46.

Hu, Mingjia and Nosofsky, Robert. 2024. High-variability training does not enhance generalization in the prototype-distortion paradigm. *Memory & Cognition*, 52(5):1017–1032. doi:10.3758/s13421-023-01516-1.

Huang, Zhenya and Liu, Qi and Chen, Enhong and Zhao, Hongke and Gao, Mingyong and Wei, Si and Su, Yu and Hu, Guoping. 2017. Question difficulty prediction for READING problems in standard tests. In Satinder Singh

and Shaul Markovitch (Eds.), *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, AAAI'17, pages 1352–1359. San Francisco, California, USA. Association for the Advancement of Artificial Intelligence (AAAI). doi:10.1609/aaai.v31i1.10740.

Hulstijn, Jan H. 2012. Incidental learning in second language acquisition. In Carol A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*, volume 5, pages 2632–2640. doi:10.1002/9781405198431.wbeal0530.

Ipek, Hulya. 2009. Comparing and contrasting first and second language acquisition: Implications for language teachers. *English Language Teaching*, 2(2):155–163. doi:10.5539/elt.v2n2p155.

Irfani, Bambang. 2017. Syllabus design for English courses. *English Education: Jurnal Tadris Bahasa Inggris*, 6(1):21–41.

James, Mark Andrew. 2018. Teaching for transfer of second language learning. *Language Teaching*, 51(3):330–348. doi:10.1017/S0261444818000137.

Jansen, Louise M. 2000. Second language acquisition: From theory to data. *Second Language Research*, 16(1):27–43. doi:10.1191/026765800671563747.

Jensen, Isabel Nadine and Slabakova, Roumyana and Westergaard, Marit and Lundquist, Björn. 2020. The Bottleneck Hypothesis in L2 acquisition: L1 Norwegian learners' knowledge of syntax and morphology in L2 English. *Second Language Research*, 36(1):3–29. doi:10.1177/0267658318825067.

Jiang, Shu and Lee, John Sie Yuen. 2017. Distractor generation for Chinese fill-in-the-blank items. In Joel Tetreault, Jill Burstein, Claudia Leacock, and Helen Yannakoudakis (Eds.), *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'17, pages 143–148. Copenhagen, Denmark. Association for Computational Linguistics. doi:10.18653/v1/W17-5015.

Jones, Carolyn and Karabulut-Klöppelt, Nilgül and Marks, Jon and Wooder, Alison and Baer-Engel, Jennifer and Kaminski, Cornelia and Köhler-Davidson, Elise. 2020. *Green line 3 G9*.

Kaipa, Ramesh and Robb, Michael and Jones, Richard. 2017. The effectiveness of constant, variable, random, and blocked practice in speech-motor learning. *Journal of Motor Learning and Development*, 5(1):103–125. doi:10.1123/jmld.2015-0044.

Kamimoto, Tadamitsu. 1993. Tailoring the test to fit the students : Improvement of the C-test through classical item analysis. *Language Laboratory*, 30:47–61. doi:10.24539/llaj.30.0_47.

Karamanis, Nikiforos and Mitkov, Ruslan et al. 2006. Generating multiple-choice test items from medical text: A pilot study. In Nathalie Colineau, Cécile Paris, Stephen Wan, and Robert Dale (Eds.), *Proceedings of the 4th International Natural Language Generation Conference*, INLG'06, pages 111–113. Sydney, Australia. Association for Computational Linguistics. doi:10.3115/1706269.1706291.

Karampiperis, Pythagoras and Sampson, Demetrios. 2005. Adaptive learning resources sequencing in educational hypermedia systems. *Educational Technology & Society*, 8(4):128–147. doi:10.1177/0895904804270777.

Karasimos, Athanasios. 2022. The battle of language learning apps: A cross-platform overview. *Research Papers in Language Teaching and Learning*, 12(1):150–168.

Katinskaia, Anisia and Hou, Jue and Vu, Anh-duc and Yangarber, Roman. 2023. Linguistic constructs represent the domain model in intelligent language tutoring. In Danilo Croce and Luca Soldaini (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics - System Demonstrations*, EACL'23, pages 136–144. Dubrovnik, Croatia. Association for Computational Linguistics. doi:10.18653/v1/2023.eacl-demo.16.

Katinskaia, Anisia and Nouri, Javad and Yangarber, Roman. 2018. Revita: A language-learning platform at the intersection of ITS and CALL. In Calzolari Nicoletta (Ed.), *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC'18. Miyazaki, Japan. European Language Resources Association (ELRA).

Katz, Sandra and Albacete, Patricia and Chounta, Irene-Angelica and Jordan, Pamela and Mclaren, Bruce and Zapata-Rivera, Diego. 2021. Linking dialogue with student modelling to create an adaptive tutoring system for conceptual physics. *International Journal of Artificial Intelligence in Education*, 31(3):397–445. doi:10.1007/s40593-020-00226-y.

Keller, Stanley Uros. 2021. Automatic generation of word problems for academic education via natural language processing (NLP). *Computing Research Repository*. doi:10.48550/arXiv.2109.13123.

Kerr, Robert and Booth, Bernard. 1978. Specific and varied practice of motor skill. *Perceptual and Motor Skills*, 46(2):395–401. doi:10.1177/003151257804600201.

Kerres, Michael and Buntins, Katja and Buchner, Josef and Drachsler, Hendrik and Zawacki-Richter, Olaf. 2023. Lernpfade in adaptiven und künstlich-intelligenten Lernprogrammen. Eine kritische Analyse aus mediendidaktischer Sicht. In de Claudia Witt, Christina Gloerfeld, and Silke Elisabeth Wrede (Eds.), *Künstliche Intelligenz in der Bildung*, pages 109–131. doi:10.1007/978-3-658-40079-8_6.

Kertz-Welzel, Alexandra. 2004. Didaktik of music: a German concept and its comparison to American music pedagogy. *International Journal of Music Education*, 22(3):277–286. doi:10.1177/0255761404049806.

Kim, Taewon and Wright, David L. and Feng, Wuwei. 2021. Commentary: Variability of practice, information processing, and decision making – How much do we know? *Frontiers in Psychology*, 12. doi:10.3389/fpsyg.2021.685749.

Kirchner, Wayne K. 1958. Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, 55(4):352–358. doi:10.1037/h0043688.

Klein-Braley, Christine. 1996. Towards a theory of C-test processing. In Rüdiger R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*, volume 3, pages 263–296.

Knoop, Susanne and Wilske, Sabrina. 2013. WordGap – Automatic generation of gap-filling vocabulary exercises for mobile learning. In Elena Volodina, Lars Borin, and Hrafn Loftsson (Eds.), *Proceedings of the 2nd Workshop on NLP for*

Computer-Assisted Language Learning, number 086 in NLP4CALL'13, pages 39–47. Linköping, Sweden. Linköping University Press.

Kobayashi, Miyoko. 2002. Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *Modern Language Journal*, 86(4):571–586. doi:10.1111/1540-4781.00162.

Koedinger, Kenneth R. and Brunskill, Emma and Baker, Ryan S. J. D. and McLaughlin, Elizabeth and Stamper, John. 2013. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):37–41. doi:10.1609/aimag.v34i3.2484.

Kohnke, Lucas and Moorhouse, Benjamin Luke and Zou, Di. 2023. ChatGPT for language teaching and learning. *RELC Journal: A Journal of Language Teaching and Research*, 54(2):537–550. doi:10.1177/00336882231162868.

Koraishi, Osama. 2023. Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. *Language Education & Technology*, 3(1):55–72.

Koutsojannis, Constantinos and Beligiannis, Grigorios and Hatzilygeroudis, Ioannis and Papavlasopoulos, Constantinos and Prentzas, Jim. 2007. Using a hybrid AI approach for exercise difficulty level adaptation. *International Journal of Continuing Engineering Education and Life-Long Learning*, 17(4-5):256–272. doi:10.1504/IJCEELL.2007.015042.

Krashen, Stephen D. 1982. *Second language acquisition theory*, pages 9–56.

Kubinger, Klaus D. and Gottschall, Christian H. 2007. Item difficulty of multiple choice tests dependant on different item response formats – An experiment in fundamental research on psychological assessment. *Psychology Science*, 49(4):361–374.

Kumar, Girish and Banchs, Rafael and D'Haro, Luis. 2015. RevUP: Automatic gap-fill question generation from educational texts. In Joel Tetreault, Jill Burstein, and Claudia Leacock (Eds.), *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'15, pages 154–161. Denver, Colorado, USA. Association for Computational Linguistics. doi:10.3115/v1/W15-0618.

Kunichika, Hidenobu and Urushima, Minoru and Hirashima, Tsukasa and Takeuchi, Akira. 2002. A computational method of complexity of questions on contents of English sentences and its evaluation. In *Proceedings of the International Conference on Computers in Education*, volume 1 of *ICCE'02*, pages 97–101. Auckland, New Zealand. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/CIE.2002.1185873.

Kurdi, Ghader and Leo, Jared and Parsia, Bijan and Sattler, Uli and Al-Emari, Salam. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204. doi:10.1007/s40593-019-00186-y.

Kwasnicka, Halina and Szul, Dorota and Markowska-Kaczmar, Urszula and Myszkowski, Pawel B. 2008. Learning Assistant - Personalizing learning paths in e-learning environments. In *Proceedings of the 7th Computer Information*

Systems and Industrial Management Applications, CISIM'08, pages 308–314. Ostrava, Czech Republic. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/CISIM.2008.51.

Labaj, Martin and Bieliková, Mária. 2014. Utilization of exercise difficulty rating by students for recommendation. In Yiwei Cao, Terje Väljataga, Jeff K. T. Tang, Howard Leung, and Mart Laanpere (Eds.), *Proceedings of the ICWL 2014 International Workshops - New Horizons in Web Based Learning*, ICWL'14, pages 13–22. Tallinn, Estonia. Springer Nature. doi:10.1007/978-3-319-13296-9_2.

Larsen-Freeman, Diane and Celce-Murcia, Marianne and Frodesen, Jan and White, Benjamin and Williams, Howard Alan. 2016. *The Grammar Book: Form, Meaning, and Use for English Language Teachers*.

LEAD Graduate School & Research Network and Eberhard Karls Universität Tübingen. Wie Künstliche Intelligenz dabei helfen soll, das Lernen individueller zu gestalten. https://lead.schule/aktuelle-studien/wie-kuenstliche-intelligenz-dabei-helfen-soll-das-lernen-individueller-zu-gestalten. [Online; accessed 27-February-2025].

Lee, Ji-Ung and Schwan, Erik and Meyer, Christian. 2019. Manipulating the difficulty of C-tests. In Anna Korhonen, David Traum, and Lluís Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL'19, pages 360–370. Florence, Italy. Association for Computational Linguistics. doi:10.18653/v1/P19-1035.

Lee, John Sie Yuen and Seneff, Stephanie. 2007. Automatic generation of cloze items for prepositions. In *Procedings of the 8th Annual Conference of the International Speech Communication Association*, ISCA'07. Antwerp, Belgium. International Speech Communication Association. doi:10.21437/Interspeech.2007-592.

Lee, John Sie Yuen and Sturgeon, Donald and Luo, Mengqi. 2016. A CALL system for learning preposition usage. In Katrin Erk and Noah A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics - Long Papers*, ACL'16, pages 984–993. Berlin, Germany. Association for Computational Linguistics. doi:10.18653/v1/P16-1093.

Lee, Siok H. and Muncie, James. 2006. From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task. *Tesol Quarterly*, 40(2):295–320. doi:10.2307/40264524.

Lei, Lei. 2008. Validation of the C-test amongst Chinese ESL learners. *Journal of Asia TEFL*, 5(2):117–140.

Leibniz-Institut für Wissensmedien. 2019. Jahresbericht 2019. https://www.iwm-tuebingen.de/@@/downloads/jahresberichte/IWM_Jahresbericht_2019.pdf. [Online; accessed 06-December-2024].

Li, Shaofeng and Ellis, Rod and Zhu, Yan. 2016. Task-based versus task-supported language instruction: An experimental study. *Annual Review of Applied Linguistics*, 36:205–229. doi:10.1017/S0267190515000069.

Liang, Chen and Yang, Xiao and Wham, Drew and Pursel, Bart and Passonneaur, Rebecca and Giles, C. Lee. 2017. Distractor generation with generative adversarial nets for automatically creating fill-in-the-blank questions. In Ós-

car Corcho, Krzysztof Janowicz, Giuseppe Rizzo, Ilaria Tiddi, and Daniel Garijo (Eds.), *Proceedings of the 9th Knowledge Capture Conference*, K-CAP'17, pages 1–4. Austin, Texas, USA. Association for Computing Machinery. doi:10.1145/3148011.3154463.

Lightbrown, Patsy Martin. 2007. Transfer appropriate processing as a model for classroom second language acquisition. In ZhaoHong Han (Ed.), *Understanding Second Language Process*, chapter 3, pages 27–44. doi:10.21832/9781847690159-005.

Lightbrown, Patsy Martin. 2019. Perfecting practice. *Modern Language Journal*, 103(3):703–712. doi:10.1111/modl.12588.

Likourezos, Vicki and Kalyuga, Slava and Sweller, John. 2019. The variability effect: When instructional variability is advantageous. *Educational Psychology Review*, 31(2):479–497. doi:10.1007/s10648-019-09462-8.

Lim, Cher Ping and Zhao, Yong and Tondeur, Jo and Chai, Ching Sing and Tsai, Chin-Chung. 2013. Bridging the gap: Technology trends and tse of technology in schools. *Journal of Educational Technology and Society*, 16(2):59–68.

Lim, Lyn and Bannert, Maria and van der Graaf, Joep and Singh, Shaveen and Fan, Yizhou and Surendrannair, Surya and Rakovic, Mladen and Molenaar, Inge and Moore, Johanna and Gašević, Dragan. 2023. Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. *Computers in Human Behavior*, 139:107547. doi:10.1016/j.chb.2022.107547.

Liu, Bo and Wei, Ying and Zhang, Yu and Yang, Qiang. 2017a. Deep neural networks for high dimension, low sample size data. In Carles Sierra (Ed.), *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2287–2293. Melbourne, Australia. Association for the Advancement of Artificial Intelligence (AAAI). doi:10.24963/ijcai.2017/318.

Liu, Chao-Lin and Wang, Chun-Hung and Gao, Zhao Ming and Huang, Shang-Ming. 2005. Applications of lexical information for algorithmically composing multiple-choice cloze items. In Jill Burstein and Claudia Leacock (Eds.), *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, BEA'05, pages 1–8. Ann Arbor, Michigan, USA. Association for Computational Linguistics. doi:10.3115/1609829.1609830.

Liu, Ming and Rus, Vasile and Liu, Li. 2017b. Automatic Chinese multiple choice question generation using mixed similarity strategy. *IEEE Transactions on Learning Technologies*, 11(2):193–202. doi:10.1109/TLT.2017.2679009.

Liu, Qi and Huang, Zhenya and Yin, Yu and Chen, Enhong and Xiong, Hui and Su, Yu and Hu, Guoping. 2021. EKT: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115. doi:10.1109/TKDE.2019.2924374.

López, Sergio-Ramón Rossetti and Ramírez, Ma-Teresa García and Rodríguez, Isaac-Shamir Rojas. 2021. Evaluation of the implementation of a learning object developed with H5P technology. *Vivat Academia*, 24(154):1–23. doi:10.15178/va.2021.154.e1224.

Lorge, Irving and Kruglov, Lorraine. 1953. The improvement of estimates of

test difficulty. *Educational and Psychological Measurement*, 13(1):34–46. doi:10.1177/001316445301300104.

Loukina, Anastassia and Yoon, Su-Youn and Sakano, Jennifer and Wei, Youhua and Sheehan, Kathy. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In Yuji Matsumoto and Rashmi Prasad (Eds.), *Proceedings of the 26th International Conference on Computational Linguistics - Technical Papers*, COLING'16, pages 3245–3253. Osaka, Japan. Coling 2016 Organizing Committee.

Lukas, Jos. 1997. Modellierung von Fehlkonzepten in einer algebraischen Wissensstruktur. *Kognitionswissenschaft*, 6(4):196–204. doi:10.1007/BF03354921.

Lv, Qianxi and Liang, Junying. 2019. Is consecutive interpreting easier than simultaneous interpreting? – A corpus-based study of lexical simplification in interpretation. *Perspectives*, 27(1):91–106. doi:10.1080/0907676X.2018.1498531.

Lyster, Roy and Sato, Masatoshi. 2013. Skill acquisition theory and the role of practice in L2 development. In del María Pilar García Mayo, María Juncal Gutiérrez Mangado, and María Martínez-Adrián (Eds.), *Contemporary Approaches to Second Language Acquisition*, chapter 4, pages 71–92. doi:10.1075/aals.9.07ch4.

Mac, Thi Thoa and Copot, Cosmin and Tran, Duc Trung and De Keyser, Robin. 2016. Heuristic approaches in robot path planning: A survey. *Robotics and Autonomous Systems*, 86(C):13–28. doi:10.1016/j.robot.2016.08.001.

MacWhinney, Brian. 1987. The Competition Model. In Brian MacWhinney (Ed.), *Mechanisms of Language Acquisition*, chapter 8, pages 249–308. doi:10.4324/9781315798721-15.

MacWhinney, Brian. 1997. Second language acquisition and the Competition Model. In de Annette M. B. Groot and Judith F. Kroll (Eds.), *Tutorials in Bilingualism*, chapter 4, pages 113–142. doi:10.4324/9781315806051.

Madlener, Karin. 2015. *Optimizing variation or: Can less be more?*, chapter 6. doi:10.1515/9783110405538-009.

Madlener, Karin. 2016. Input optimization: Effects of type and token frequency manipulations in instructed second language learning. In Heike Behrens and Stefan Pfänder (Eds.), *Experience Counts: Frequency Effects in Language*, volume 54, pages 133–174. doi:10.1515/9783110346916-007.

Mak, Barley and Xiao, Yangyu. 2018. The CUHK LPATE training courses: Writing, speaking and classroom language. In David Coniam and Peter Falvey (Eds.), *High-Stakes Testing: The Impact of the LPATE on English Language Teachers in Hong Kong*, pages 207–235. doi:10.1007/978-981-10-6358-9_11.

Malafeev, Alexey. 2015. Exercise Maker: Automatic language exercise generation. In V. P. Selegey (Ed.), *Papers from the Annual International Conference "Dialogue" - Computational Linguistics and Intellectual Technologies*, volume 14 of *Dialogue'15*, pages 441–452. Moscow, Russia.

Mansouri, Fethi and Duffy, Loretta. 2005. The pedagogic effectiveness of developmental readiness in ESL grammar instruction. *Australian Review of Applied Linguistics*, 28(1):81–99. doi:10.1075/aral.28.1.06man.

Maritxalar, Montse and Dhonnchadha, Elaine and Foster, Jennifer and Ward, Monica. 2011. Quizzes on Tap: Exporting a test generation system from one less-resourced language to another. In Joseph Vetulani, Zygmuntand Mariani (Ed.), *Proceedings of the 9th Language and Technology Conference - Human Language Technology. Challenges for Computer Science and Linguistics*, volume 8387 of *LTC'19*, pages 502–514. Poznan, Poland. Springer Nature. doi:10.1007/978-3-319-08958-4_41.

Martin, Katherine I. and Ellis, Nick C. 2012. The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, 34(3):379–413. doi:10.1017/S0272263112000125.

Mashal, Nira and Metzuyanim-Gorelick, Shlomit. 2019. New information on the effects of transcranial direct current stimulation on n-back task performance. *Experimental Brain Research*, 237(5):1315–1324. doi:10.1007/s00221-019-05500-7.

Mayo, Michael and Mitrovic, Antonija. 2001. Optimising ITS behaviour with Bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education*, 12:124–153.

McCarthy, Arya D. and Yancey, Kevin P. and LaFlair, Geoffrey T. and Egbert, Jesse and Liao, Manqian and Settles, Burr. 2021. Jump-starting item parameters for adaptive language tests. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP'21, pages 883–899. Punta Cana, Dominican Republic. Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.67.

McKendree, Jean. 1990. Effective feedback content for tutoring complex skills. *Human-computer Interaction*, 5(4):381–413. doi:10.1207/s15327051hci0504_2.

Medawela, Rajapakse Mudiyanselage Sumudu Himesha Bandara and Ratnayake, Dugganna Ralalage Dilini Lalanthi and Abeyasinghe, Wijeyapala and Jayasinghe, Ruwan and Marambe, Kosala. 2018. Effectiveness of "fill in the blanks" over multiple choice questions in assessing final year dental undergraduates. *Educación Médica*, 19(2):72–76. doi:10.1016/j.edumed.2017.03.010.

Melvin, Bernice S. and Stout, David F. 1987. Motivating language learners through authentic materials. In Wilga M. Rivers (Ed.), *Interactive Language Teaching*, Cambridge Language Teaching Library, chapter 4, pages 44–57.

van Merriënboer, Jeroen J. G. 1997. *Training Complex Cognitive Skills: A Four-Component Instructional Design Model for Technical Training*.

Mettadewi, Bella and Sitohang, Mangangar and Sutrisno, Bejo. 2023. Exploring factors that cause the students demotivation in learning English grammar and English writing STIBA IEC students during COVID-19 pandemic. *Journal of English Language and Literature*, 8(2):287–300. doi:10.37110/jell.v8i02.193.

Meurers, Detmar. 2012. Natural language processing and language learning. In Carol A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*, pages 1–15. doi:10.1002/9781405198431.wbeal0858.

Meurers, Detmar and De Kuthy, Kordula and Nuxoll, Florian and Rudzewitz, Björn and Ziai, Ramon. 2019. Scaling up intervention studies to investigate real-life

foreign language learning in school. *Annual Review of Applied Linguistics*, 39:161–188. doi:10.1017/S0267190519000126.

Meurers, Detmar and Ziai, Ramon and Amaral, Luiz A. and Boyd, Adriane and Dimitrov, Aleksandar and Metcalf, Vanessa and Ott, Niels. 2010. Enhancing authentic Web pages for language learners. In Joel Tetreault, Jill Burstein, and Claudia Leacock (Eds.), *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'10, pages 10–18. Los Angeles, California, USA. Association for Computational Linguistics.

Millán, Eva and Agosta, John Mark and Pérez de la Cruz, J. L. 2001. Bayesian student modeling and the problem of parameter specification. *British Journal of Educational Technology*, 32(2):171–181. doi:10.1111/1467-8535.00188.

Millán, Eva and Loboda, Tomasz Dominik and Pérez-de-la-Cruz, Jose Luis. 2010. Bayesian networks for student model engineering. *Computers & Education*, 55(4):1663–1683. doi:10.1016/j.compedu.2010.07.010.

Mindt, Dieter. 2014. Corpora and the teaching of English in Germany. In Anne Wichmann and Steven Fligelstone (Eds.), *Teaching and language corpora*, chapter 3, pages 40–50. doi:10.4324/9781315842677-4.

Ministerium für Kultus, Jugend und Sport Baden-Württemberg. 2002. Werbung, Wettbewerbe und Erhebungen in Schulen. https://www.landesrecht-bw.de/bsbw/document/VVBW-VVBW000002696. [Online; Reference number: 6499.10/417].

Mitkov, Ruslan and Le An, Ha and Karamanis, Nikiforos. 2006. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2):177–194. doi:10.1017/S1351324906004177.

Mitkov, Ruslan and Varga, Andrea and Rello, Luz et al. 2009. Semantic similarity of distractors in multiple-choice tests: Extrinsic evaluation. In Roberto Basili and Marco Pennacchiotti (Eds.), *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS'09, pages 49–56. Athens, Greece. Association for Computational Linguistics. doi:10.3115/1705415.1705422.

Mitrovic, Antonija and Martin, Brent and Mayo, Michael. 2002. Using evaluation to shape ITS design: Results and experiences with SQL-Tutor. *User Modeling and User-adapted Interaction*, 12(2):243–279. doi:10.1023/A:1015022619307.

Moeyaert, Mariola and Wauters, Kelly and Desmet, Piet and Van den Noortgate, Wim. 2016. When easy becomes boring and difficult becomes frustrating: Disentangling the effects of item difficulty level and person proficiency on learning and motivation. *Systems*, 4(1):14–31. doi:10.3390/systems4010014.

Monteira, Sabela F. and Jiménez-Aleixandre, María Pilar and Siry, Christina. 2022. Scaffolding children's production of representations along the three years of ECE: A longitudinal study. *Research in Science Education*, 52(1):127–158. doi:10.1007/s11165-020-09931-z.

Moser, Josef Robert and Gütl, Christian and Liu, Wei. 2012. Refined distractor generation with LSA and stylometry for automated multiple choice question generation. In Dongmo Thielscher, Michaeland Zhang (Ed.), *Proceedings of the 25th International Australasian Joint Conference on Advances in Artificial Intelligence*, AI'12, pages 95–106. Sydney, Australia. Springer Nature. doi:10.1007/978-3-642-35101-3_9.

Mostow, Jack and Jang, Hyeju. 2012. Generating diagnostic multiple choice comprehension cloze questions. In Joel Tetreault, Jill Burstein, and Claudia Leacock (Eds.), *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'12, pages 136–146. Montréal, Quebec, Canada. Association for Computational Linguistics.

Muhammad, Alva and Zhou, Qingguo and Beydoun, Ghassan and Xu, Dongming and Shen, Jun. 2016. Learning path adaptation in online learning systems. In *Proceedings of the 20th IEEE International Conference on Computer Supported Cooperative Work in Design*, CSCWD'16, pages 421–426. Nanchang, China. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/CSCWD.2016.7566026.

Müller-Hartmann, Andreas and Schocker-von Ditfurth, Marita. 2013. Research on the use of technology in task-based language teaching. In Michael Thomas and Hayo Reinders (Eds.), *Task-Based Language Learning and Teaching with Technology*, Second Language Teaching and Curriculum Center Honolulu, Hawaii: Technical report, chapter 2, pages 17–40.

Munby, John. 1981. *Language skills selection*, Communicative Syllabus Design, chapter 7.

Murray, Charles R. and VanLehn, Kurt and Mostow, Jack. 2004. Looking ahead to select tutorial actions: A decision-theoretic approach. *International Journal of Artificial Intelligence in Education*, 14(3-4):235–278.

Murray, Tom and Arroyo, Ivon. 2002. Toward measuring and maintaining the Zone of Proximal Development in adaptive instructional systems. In Stefano A. Cerri, Guy Gouardères, and Fàbio Paraguaçu (Eds.), *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, ITS'02, pages 749–758. Biarritz, France and San Sebastian, Spain. Springer Nature. doi:10.1007/3-540-47987-2_75.

Nagata, Noriko. 2009. Robo-Sensei's NLP-based error detection and feedback generation. *CALICO Journal*, 26(3):562–579. doi:10.1558/cj.v26i3.562-579.

Nagatani, Koki and Zhang, Qian and Sato, Masahiro and Chen, Yan-Ying and Chen, Francine and Ohkuma, Tomoko. 2019. Augmenting knowledge tracing by considering forgetting behavior. In Ling Liu and Ryen White (Eds.), *The World Wide Web Conference*, WWW'19, pages 3101–3107. San Francisco, California, USA. Association for Computing Machinery. doi:10.1145/3308558.3313565.

Nakata, Tatsuya and Suzuki, Yuichi. 2019. Mixing grammar exercises facilitates long-term retention: Effects of blocking, interleaving, and increasing practice. *Modern Language Journal*, 103(3):629–647. doi:10.1111/modl.12581.

Nation, Paul. 2007. The four strands. *Innovation in Language Learning and Teaching*, 1(1):2–13. doi:10.2167/illt039.0.

Newlin, Philip and Moss, Jarrod. 2020. Proceduralization and working memory in association learning. In Stephanie Denison, Michael Mack, Yang Xu, and Blair C. Armstrong (Eds.), *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines*, CogSci'20. Online. Cognitive Science Society.

Ní Chiaráin, Neasa and Ní Chasaide, Ailbhe. 2020. The potential of text-to-speech synthesis in computer-assisted language learning. In Alberto Andujar (Ed.), *Re-*

cent Tools for Computer- and Mobile-Assisted Foreign Language Learning, chapter 7, pages 149–169. doi:10.4018/978-1-7998-1097-1.ch007.

Nikolova, Ivelina. 2009. New Issues and solutions in computer-aided design of MCTI and distractor selection for Bulgarian. In Elena Paskaleva, Stelios Piperidis, Milena Slavcheva, and Cristina Vertan (Eds.), *Proceedings of the Workshop Multilingual resources, technologies and evaluation for central and Eastern European languages*, RANLP'09, pages 40–46. Borovets, Bulgaria. Association for Computational Linguistics.

Nikouee, Majid. 2021. *Grammar practice and communicative language teaching: Groundwork for an investigation into the concept of transfer-appropriateness*. Ph.D. thesis, University of Alberta. doi:10.7939/r3-mdt5-ew23.

Nye, Benjamin and Graesser, Arthur and Hu, Xiangen. 2014. AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4):427–469. doi:10.1007/s40593-014-0029-5.

Oh, Eunjung Grace and Reeves, Thomas C. 2013. Generational differences and the integration of technology in learning, instruction, and performance. In J. Michael Spector, M. David Merrill, Jan Elen, and M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology*, pages 819–828. doi:10.1007/978-1-4614-3185-5_66.

Ohlsson, Stellan. 1994. Constraint-based student modeling. In Jim E. Greer and Gordon I. McCalla (Eds.), *Proceedings of the Workshop on Student Modelling: The Key to Individualized Knowledge-Based Instruction*, pages 167–189. Sainte-Adèle, Quebec, Canada. Springer Nature. doi:10.1007/978-3-662-03037-0_7.

Oller, John W. 1972. Scoring methods and difficulty levels for cloze tests of proficiency in English as a Second Language. *Modern Language Journal*, 56(3):151–158. doi:10.2307/324037.

Opitz, Bertram and Ferdinand, Nicola and Mecklinger, Axel. 2011. Timing matters: The impact of immediate and delayed feedback on artificial language learning. *Frontiers in Human Neuroscience*, 5. doi:10.3389/fnhum.2011.00008.

Oxford, Rebecca L. 2016. *The soul of L2 learning strategies: Self-regulation, agency, autonomy, and associated factors in the strategic self-regulation (S2R) model*, chapter 2. doi:10.4324/9781315719146-14.

Pallotti, Gabriele. 2007. An operational definition of the emergence criterion. *Applied Linguistics*, 28(3):361–382. doi:10.1093/applin/amm018.

Pallotti, Gabriele. 2019. An approach to assessing the linguistic difficulty of tasks. *Journal of the European Second Language Association*, 3(1):58–70. doi:10.22599/jesla.61.

Pandarova, Irina and Schmidt, Torben and Hartig, Johannes and Boubekki, Ahcene and Jones, Roger and Brefeld, Ulf. 2019. Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education*, 29(3):342–367. doi:10.1007/s40593-019-00180-4.

Papasalouros, Andreas and Kanaris, Konstantinos and Kotis, Konstantinos I. 2008. Automatic generation of multiple choice questions from domain ontologies. In

Proceedings of the IADIS International Conference on e-Learning, volume 1 of *MCCSIS'08*. Amsterdam, The Netherlands.

Park, Jung Yeon and Cornillie, Frederik and van der Maas, Han L. J. and Van den Noortgate, Wim. 2019. A multidimensional IRT approach for dynamically monitoring ability growth in computerized practice environments. *Frontiers in Psychology*, 10. doi:10.3389/fpsyg.2019.00620.

Park, Ok-choon and Lee, Jung. 2004. Adaptive instructional systems. In David Jonassen and Marcy Driscoll (Eds.), *Handbook of Research for Educational Communications and Technology*, chapter 35, pages 651–684. doi:10.4324/9781410609519.

Parkinson, Meghan M. and Dinsmore, Daniel L. 2019. Understanding the developmental trajectory of second language acquisition and foreign language teaching and learning using the Model of Domain Learning. *System*, 86. doi:10.1016/j.system.2019.102125.

Parrisius, Cora and Pieronczyk, Ines and Blume, Carolyn and Wendebourg, Katharina and Pili-Moss, Diana and Assmann, Mirjam and Beilharz, Sabine and Bodnar, Stephen and Colling, Leona and Holz, Heiko et al. 2022a. Using an intelligent tutoring system within a task-based learning approach in English as a foreign language classes to foster motivation and learning outcome (Interact4School): Preregistration of the study design.

Parrisius, Cora and Wendebourg, Katharina and Rieger, Sven and Loll, Ines and Pili-Moss, Diana and Colling, Leona and Blume, Carolyn and Pieronczyk, Ines and Holz, Heiko and Bodnar, Stephen et al. 2022b. Effective features of feedback in an intelligent tutoring system – A randomized controlled field trial (pre-registration).

Patil, Rajkumar and Palve, Sachin Bhaskar and Vell, Kamesh and Boratne, Abhijit Vinod. 2016. Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. *International Journal of Community Medicine and Public Health*, 3(6):1612–1616. doi:10.18203/2394-6040.ijcmph20161638.

Pavlik, Philip I. and Cen, Hao and Koedinger, Kenneth R. 2009. Performance Factors Analysis – A new alternative to knowledge tracing. In Vania Dimitrova, Riichiro Mizoguchi, du Benedict Boulay, and Art Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, volume 200 of *AIED'09*, pages 531–538. Brighton, England, UK. IOS Press. doi:10.3233/978-1-60750-028-5-531.

Peacock, Matthew. 1997. The effect of authentic materials on the motivation of EFL learners. *English Language Teaching*, 51(2):144–156. doi:10.1093/elt/51.2.144.

Pelánek, Radek. 2017. Bayesian Knowledge Tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-adapted Interaction*, 27(3):313–350. doi:10.1007/s11257-017-9193-2.

Pelánek, Radek. 2018. Conceptual issues in mastery criteria: Differentiating uncertainty and degrees of knowledge. In Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and du Benedict Boulay (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence in Education*, AIED'18,

pages 450–461. London, England, UK. Springer Nature. doi:10.1007/978-3-319-93843-1_33.

Pelánek, Radek. 2019. Managing items and knowledge components: Domain modeling in practice. *Educational Technology Research and Development*, 68(1):529–550. doi:10.1007/s11423-019-09716-w.

Pelánek, Radek. 2022. Adaptive, intelligent, and personalized: Navigating the terminological maze behind educational technology. *International Journal of Artificial Intelligence in Education*, 32(1):151–173. doi:10.1007/s40593-021-00251-5.

Pelánek, Radek and Effenberger, Tomáš and Čechák, Jaroslav. 2021. Complexity and difficulty of items in learning systems. *International Journal of Artificial Intelligence in Education*, 32(1):196–232. doi:10.1007/s40593-021-00252-4.

Pelánek, Radek and Řihák, Jiří. 2017. Experimental analysis of mastery learning criteria. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP'17, pages 156–163. Bratislava, Slovakia. Association for Computing Machinery. doi:10.1145/3079628.3079667.

Pelánek, Radek and Řihák, Jiří. 2018. Analysis and design of mastery learning criteria. *New Review of Hypermedia and Multimedia*, 24(3):133–159. doi:10.1080/13614568.2018.1476596.

Pelegrina, Santiago and Lechuga, María and García-Madruga, Juan Antonio and Elosua, Maria and Macizo, Pedro and Carreiras, Manuel and Fuentes, Luis and Bajo, Maria. 2015. Normative data on the n-back task for children and young adolescents. *Frontiers in Psychology*, 6. doi:10.3389/fpsyg.2015.01544.

Peña, Alejandro and Kayashima, Michiko and Mizoguchi, Riichiro and Dominguez, Rafael. 2011. Improving students' meta-cognitive skills within intelligent educational systems: A review. In Dylan D. Schmorrow and Cali M. Fidopiastis (Eds.), *Proceedings of the 6th International Conference on Foundations of Augmented Cognition - Directing the Future of Adaptive Systems*, FAC'11, pages 442–451. Orlando, Florida, USA. Springer Nature. doi:10.1007/978-3-642-21852-1_51.

Pérez, Naiara and Cuadros, Montse. 2017. Multilingual CALL framework for automatic language exercise generation from free text. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics - Software Demonstrations*, EACL'17, pages 49–52. Valencia, Spain. Association for Computational Linguistics. doi:10.18653/v1/E17-3013.

Perez-Beltrachini, Laura and Gardent, Claire and Kruszewski, German. 2012. Generating grammar exercises. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'12, pages 147–156. Montréal, Quebec, Canada. Association for Computational Linguistics.

Perry, Lynn K. and Samuelson, Larissa K. and Malloy, Lisa M. and Schiffer, Ryan N. 2010. Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychological Science*, 21(12):1894–1902. doi:10.1177/0956797610389189. PMID: 21106892.

Peterson, Mark. 2013. Task-based language teaching in network-based CALL: An analysis of research on learner interaction in synchronous CMC. In Michael Thomas and Hayo Reinders (Eds.), *Task-Based Language Learning and Teaching*

with Technology, Second Language Teaching and Curriculum Center Honolulu, Hawaii: Technical report, pages 41–62.

Phillips, Andrea and Pane, John F. and Reumann-Moore, Rebecca and Shenbanjo, Oluwatosin. 2020. Implementing an adaptive intelligent tutoring system as an instructional supplement. *Educational Technology Research and Development*, 68(3):1409–1437. doi:10.1007/s11423-020-09745-w.

Phobun, Pipatsarun and Vicheanpanya, Jiracha. 2010. Adaptive intelligent tutoring systems for e-learning systems. *Procedia Social and Behavioral Sciences*, 2(2):4064–4069. doi:10.1016/j.sbspro.2010.03.641.

Pienemann, Manfred. 1998. *Language Processing and Second Language Development: Processability Theory*, volume 15 of *Studies in bilingualism*. doi:10.1075/sibil.15.

Pienemann, Manfred. 2015. An outline of Processability Theory and its relationship to other approaches to SLA. *Language Learning*, 65(1):123–151. doi:10.1111/lang.12095.

Pienemann, Manfred and Johnston, Malcolm and Brindley, Geoff. 1988. Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition*, 10(2):217–243. doi:10.1017/s0272263100007324.

Pilán, Ildikó and Volodina, Elena. 2014. Reusing Swedish FrameNet for training semantic roles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC'14, pages 1359–1363. Reykjavik, Iceland. European Language Resources Association (ELRA).

Pilán, Ildikó and Volodina, Elena and Borin, Lars. 2017. Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues*, 57(3):67–91. doi:10.48550/arXiv.1706.03530.

Pino, Juan Miguel and Eskénazi, Maxine. 2009. Semi-automatic generation of cloze question distractors effect of students' L1. In *ISCA International Workshop on Speech and Language Technology in Education*, SLaTE'09, pages 65–68. Wroxall, England, UK. International Speech Communication Association. doi:10.21437/SLaTE.2009-27.

Pino, Juan Miguel and Heilman, Michael and Eskenazi, Maxine. 2008. A selection strategy to improve cloze question quality. In Vincent Aleven, Kevin Ashley, Collin Lynch, and Niels Pinkwart (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, ITS'08, pages 22–32. Montréal, Quebec, Canada.

Plante, Elena and Patterson, Dianne and Gómez, Rebecca and Almryde, Kyle R. and White, Milo G. and Asbjørnsen, Arve E. 2015. The nature of the language input affects brain activation during learning from a natural language. *Journal of Neurolinguistics*, 36:17–34. doi:10.1016/j.jneuroling.2015.04.005.

Plass, Jan L. and Pawar, Shashank. 2020. Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, 52(3):275–300. doi:10.1080/15391523.2020.1719943.

Pozo, Juan-Ignacio and Pérez Echeverría, María-Puy and Cabellos, Beatriz and Sánchez, Daniel L. 2021. Teaching and learning in times of COVID-19: Uses of digital technologies during school lockdowns. *Frontiers in Psychology*, 12. doi:10.3389/fpsyg.2021.656776.

Presser, Shawn. 2020. Homemade BookCorpus. https://github.com/soskek/bookcorpus/issues/27. [Online; accessed 08-November-2022].

Pufahl, Ingrid and Rhodes, Nancy C. and Christian, Donna. 2001. What can we learn from foreign language teaching in other countries. *ERIC Digest*.

Quan, Pei and Shi, Yong and Niu, Lingfeng and Liu, Ying and Zhang, Tianlin. 2018. Automatic Chinese multiple-choice question generation for human resource performance appraisal. *Procedia Computer Science*, 139:165–172. doi:10.1016/j.procs.2018.10.235.

Quixal, Martí and Rudzewitz, Björn and Bear, Elizabeth and Meurers, Detmar. 2021. Automatic annotation of curricular language targets to enrich activity models and support both pedagogy and adaptive systems. In David Alfter, Elena Volodina, Ildikó Pilan, Johannes Graën, and Lars Borin (Eds.), *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, NLP4CALL'21, pages 15–27. Online. LiU Electronic Press.

Rasheed, Fareeha and Wahid, Abdul. 2019. Sequence generation for learning: A transformation from past to future. *The International Journal of Information and Learning Technology*, 36(5):434–452. doi:10.1108/IJILT-01-2019-0014.

Rasoli, Mohammad Kabir. 2021. Validation of C-test among Afghan students of English as a foreign language. *International Journal of Language Testing*, 11(2):109–121.

Rathod, Manav and Tu, Tony and Stasaski, Katherine. 2022. Educational multi-question generation for reading comprehension. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (Eds.), *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'22, pages 216–223. Seattle, Washington, USA. Association for Computational Linguistics. doi:10.18653/v1/2022.bea-1.26.

Ravi, Gautham Adithya and Sosnovsky, Sergey. 2013. Exercise difficulty calibration based on student log mining. In F. Mšdritscher, V. Luengo, E. Lai-Chong Law, and U. Hoppe (Eds.), *Proceedings of the Workshop on Data Analysis and Interpretation for Learning Environments*, DAILE'13. Villard-de-Lans, France. doi:10.13140/2.1.3878.4647.

Raviv, Limor and Lupyan, Gary and Green, Shawn C. 2022. How variability shapes learning and generalization. *Trends in Cognitive Sciences*, 26(6):462–483. doi:10.1016/j.tics.2022.03.007.

Razzaq, Leena and Heffernan, Neil T. 2006. Scaffolding vs. hints in the Assistment System. In Mitsuru Ikeda, Kevin D. Ashley, and Tak-Wai Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, ITS'06, pages 635–644. Jhongli, Taiwan. Springer Nature. doi:10.1007/11774303_63.

Révész, Andrea and Kourtali, Nektaria-Efstathia and Mazgutova, Diana. 2017. Ef-

fects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning*, 67(1):208–241. doi:10.1111/lang.12205.

Reynolds, Robert and Schaf, Eduard and Meurers, Detmar. 2014. A VIEW of Russian: Visual input enhancement and adaptive feedback. In Elena Volodina, Lars Borin, and Ildikó Pilán (Eds.), *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*, NLP4CALL'14, pages 98–112. Uppsala, Sweden. LiU Electronic Press.

Rich, Elaine. 1979. User modeling via stereotypes. *Cognitive Science*, 3(4):329–354. doi:10.1016/S0364-0213(79)80012-9.

Richey, J. Elizabeth and Nokes-Malach, Timothy J. 2013. How much is too much? learning and motivation effects of adding instructional explanations to worked examples. *Learning and Instruction*, 25:104–124. doi:10.1016/j.learninstruc.2012.11.006.

Ridgeway, Karl and Mozer, Michael C. and Bowles, Anita R. 2017. Forgetting of foreign-language skills: A corpus-based analysis of online tutoring software. *Cognitive Science*, 41(4):924–949. doi:10.1111/cogs.12385.

Robinson, Peter. 2021. The Cognition Hypothesis, the Triadic Componential Framework and the SSARC model: An instructional design theory of pedagogic task sequencing. In Mohammad Javad Ahmadian and Michael H. Long (Eds.), *The Cambridge Handbook of Task-Based Language Teaching*, Cambridge Handbooks in Language and Linguistics, chapter 6, pages 205–225. doi:10.1017/9781108868327.013.

Robson, Robby and Barr, Avron. 2013. Lowering the barrier to adoption of intelligent tutoring systems through standardization. In Robert Sottilare, Heather Holden, Arthur Graesser, and Xiangen Hu (Eds.), *Design Recommendations for Intelligent Tutoring Systems*, volume 1 – Learner Modeling, chapter 2, pages 7–13.

Roediger, Henry and Marsh, Elizabeth. 2005. The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):1155–1159. doi:10.1037/0278-7393.31.5.1155.

Rosell-Aguilar, Fernando. 2007. Top of the pods – In search of a podcasting "podagogy" for language learning. *Computer Assisted Language Learning*, 20(5):471–492. doi:10.1080/09588220701746047.

Rouhani, Mahmood. 2008. Another look at the C-test: A validation study with Iranian EFL learners. *The Asian EFL Journal*, 10(1):154–180.

Rudzewitz, Björn and Ziai, Ramon and De Kuthy, Kordula and Meurers, Detmar. 2017. Developing a web-based workbook for English supporting the interaction of students and teachers. In Elena Volodina, Gintarė Grigonytė, Ildikó Pilán, Kristina Nilsson Björkenstam, and Lars Borin (Eds.), *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, NLP4CALL&LA'16, pages 36–46. Gothenburg, Sweden. LiU Electronic Press.

Rudzewitz, Björn and Ziai, Ramon and De Kuthy, Kordula and Möller, Verena and Nuxoll, Florian and Meurers, Detmar. 2018. Generating feedback for English foreign language exercises. In Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis (Eds.), *Proceedings of the 13th Work-*

shop on Innovative Use of NLP for Building Educational Applications, BEA'18, pages 127–136. New Orleans, Louisiana, USA. Association for Computational Linguistics. doi:10.18653/v1/W18-0513.

Ruiz, Simón and Rebuschat, Patrick and Meurers, Detmar. 2023. Supporting individualized practice through intelligent CALL. In Yuichi Suzuki (Ed.), *Practice and Automatization in Second Language Research*, chapter 5, pages 119–143. doi:10.4324/9781003414643-7.

Rus, Vasile and Niraula, Nobal B. and Banjade, Rajendra. 2015. DeepTutor: An effective, online intelligent tutoring system that promotes deep learning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, AAAI'15, pages 4294–4295. Austin, Texas, USA. Association for the Advancement of Artificial Intelligence (AAAI). doi:10.1609/aaai.v29i1.9269.

Sabbah, Sabah. 2018. English language syllabuses: Definition, types, design, and selection. *Arab World English Journal*, 9(2):127–142. doi:10.2139/ssrn.3201914.

Sakaguchi, Keisuke and Arase, Yuki and Komachi, Mamoru. 2013. Discriminative approach to fill-in-the-blank quiz generation for language learners. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics - Short Papers*, ACL'13, pages 238–242. Sofia, Bulgaria. Association for Computational Linguistics.

Sarapuu, Ingrid and Alas, Ene. 2016. Developing a C-test to measure language ability as an alternative to a skills-based test. *Eesti Rakenduslingvistika Ühingu Aastaraamat*, 12(1):237–252. doi:10.5128/ERYa12.14.

Saville, Nick. 2023. The future of language learning in the era of ChatGPT. *Journal of e-Learning and Knowledge Society*, 19(3):6–10. doi:10.20368/1971-8829/1135875.

Schecker, Horst and Parchmann, Ilka. 2007. *Standards and competence models: The German situation*, pages 147–164.

Scherer, George A. C. 1957. The forgetting rate in learning German. *German Quarterly*, 30(4):275–277. doi:10.2307/401773.

Scheutz, Matthias and Heilman, Michael and Wenger, Aaron and Ryan, Colleen. 2005. MALT – A multi-lingual adaptive language tutor. In Chee-Kit Looi, Gord McCalla, Bert Bredeweg, and Joost Breuker (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, AIED'05, pages 920–922. Amsterdam, Netherlands. IOS Press.

Schleepen, Tamara M. J. and Jonkman, Lisa M. 2009. The development of non-spatial working memory capacity during childhood and adolescence and the role of interference control: An n-back task study. *Developmental Neuropsychology*, 35(1):37–56. doi:10.1080/87565640903325733.

Schmidt, Richard A. 1975. A schema theory of discrete motor skill learning. *Psychological Review*, 82(4):225–260. doi:10.1037/h0076770.

Schmidt, Richard W. 1990. The role of consciousness in second language learning. *Applied Linguistics*, 11(2):129–158. doi:10.1093/applin/11.2.129.

Schorn, Julia M. and Knowlton, Barbara J. 2021. Interleaved practice benefits implicit sequence learning and transfer. *Memory & Cognition*, 49(7):1436–1452. doi:10.3758/s13421-021-01168-z.

Schulz, Johannes Maximilian. 2024. *Multi-word-constructions and linguistic development in early foreign language classrooms: the role of input variability*. Ph.D. thesis, University of Oxford.

Schulze Bernd, Nadine and Chounta, Irene-Angelica. 2024. Beyond the grey area: Exploring the effectiveness of scaffolding as a learning measure. In Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt (Eds.), *Proceedings of the 25th International Conference on Artificial Intelligence in Education*, AIED'24, pages 365–378. Recife, Brazil. Springer Nature. doi:10.1007/978-3-031-64302-6_26.

Schwarz, Daniel and Oleggini, Lorenzo and Steiner, Christina M. and Stoecker, Martin. 2012. The 80Days game. In Michael D. Kickmeier-Rust and Dietrich Albert (Eds.), *An Alien's Guide to Multi-Adaptive Educational Computer Games*, pages 153–186.

Settles, Burr and T. LaFlair, Geoffrey and Hagiwara, Masato. 2020. Machine learning-driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263. doi:10.1162/tacl_a_00310.

Shabani, Karim and Mohammad, Khatib and Ebadi, Saman. 2010. Vygotsky's Zone of Proximal Development: Instructional implications and teachers' professional development. *English Language Teaching*, 3:237–248. doi:10.5539/elt.v3n4p237.

Shea, Charles. 2002. Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review*, 9(2):185–211. doi:10.3758/BF03196276.

Shei, Chi-Chiang. 2001. FollowYou!: An automatic language lesson generation system. *Computer Assisted Language Learning*, 14(2):129–144. doi:10.1076/call.14.2.129.5777.

Shen, Shuanghong and Liu, Qi and Huang, Zhenya and Zheng, Yonghe and Yin, Minghao and Wang, Minjuan and Chen, Enhong. 2024. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*, 17:1898–1919. doi:10.1109/tlt.2024.3383325.

Shin, Jinnie and Bulut, Okan and Gierl, Mark J. 2020. The effect of the most-attractive-distractor location on multiple-choice item difficulty. *Journal of Experimental Education*, 88(4):643–659. doi:10.1080/00220973.2019.1629577.

Shute, Valerie. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189. doi:10.3102/0034654307313795.

Sigott, Günther. 1995. The C-test: Some factors of difficulty. *Aaa-arbeiten Aus Anglistik Und Amerikanistik*, 20(1):43–53.

Singh, Ravi and Saleem, Muhammad and Pradhan, Prabodha and Heffernan, Cristina and Heffernan, Neil T. and Razzaq, Leena and Dailey, Matthew D. and O'Connor, Cristine and Mulcahy, Courtney. 2011. Feedback during Web-based homework: The role of hints. In Gautam Biswas, Susan Bull, Judy Kay, and Antonija Mitrovic (Eds.), *Proceedings of the 15th International Conference*

on Artificial Intelligence in Education, AIED'11, pages 328–336. Auckland, New Zealand. Springer Nature. doi:10.1007/978-3-642-21869-9_43.

Slavuj, Vanja and Kovačić, Božidar and Jugo, Igor. 2015. Intelligent tutoring systems for language learning. In *Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics*, MIPRO'15, pages 814–819. Opatija, Croatia. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/MIPRO.2015.7160383.

Slavuj, Vanja and Nacinovic Prskalo, Lucia and Brkic Bakaric, Marija. 2022. Automatic generation of language exercises based on a universal methodology: An analysis of possibilities. *Bulletin of the Transilvania University of Brasov. Series IV: Philology and Cultural Studies*, 14 (63)(2):29–48. doi:10.31926/but.pcs.2021.63.14.2.3.

Smith, Simon and Avinesh, P. V. S. and Kilgarriff, Adam. 2010. Gap-fill tests for language learners: Corpus-driven item generation. In *Proceedings of the 8th International Conference on Natural Language Processing*, ICON'10, pages 1–6. Kharagpur, India. Macmillan Publishers.

Smith, Simon and Kilgarriff, Adam and Wen-liang, Gong and Sommers, Scott and Guang-zhong, Wu. 2009. Automatic multiple choice gap-fill generation for English proficiency testing. In *Proceedings of the LTTC Conference: A new look at language teaching and testing : English as subject and vehicle*, LTTC'09. Taipei, Taiwan.

Song, Qi and Luo, Wenjie. 2023. SFBKT: A synthetically forgetting behavior method for knowledge tracing. *Applied Sciences*, 13(13). doi:10.3390/app13137704.

Spada, Nina and Tomita, Yasuyo. 2010. Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, 60(2):263–308. doi:10.1111/j.1467-9922.2010.00562.x.

Stafford, Catherine A. and Bowden, Harriet Wood and Sanz, Cristina. 2012. Optimizing language instruction: Matters of explicitness, practice, and cue learning. *Language Learning*, 62(3):741–768. doi:10.1111/j.1467-9922.2011.00648.x.

Stasaski, Katherine and Hearst, Marti A. 2017. Multiple choice question generation utilizing an ontology. In Joel Tetreault, Jill Burstein, Claudia Leacock, and Helen Yannakoudakis (Eds.), *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'17, pages 303–312. Copenhagen, Denmark. Association for Computational Linguistics. doi:10.18653/v1/W17-5034.

Stratton, James M. 2022. Intentional and incidental vocabulary learning: The role of historical linguistics in the second language classroom. *Modern Language Journal*, 106(4):837–857. doi:10.1111/modl.12805.

Sumita, Eiichiro and Sugaya, Fumiaki and Yamamoto, Seiichi. 2005. Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. In Jill Burstein and Claudia Leacock (Eds.), *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, BEA'05, pages 61–68. Ann Arbor, Michigan, USA. Association for Computational Linguistics. doi:10.3115/1609829.1609839.

Susanti, Yuni and Iida, Ryu and Tokunaga, Takenobu. 2015. Automatic generation of English vocabulary tests. In Markus Helfert, Maria Teresa Restivo, Susan Zvacek, and James Uhomoibhi (Eds.), *Proceedings of the 7th International Conference on Computer Supported Education*, volume 1 of *CSEDU'15*, pages 77–87. Lisbon Portugal. Science and Technology Publications, (SCITEPRESS). doi:10.5220/0005437200770087.

Susanti, Yuni and Tokunaga, Takenobu and Nishikawa, Hitoshi and Obari, Hiroyuki. 2018. Automatic distractor generation for multiple-choice English vocabulary questions. *Research and Practice in Technology Enhanced Learning*, 13(1). doi:10.1186/s41039-018-0082-z.

Susanti, Yunik and Tokunaga, Takenobu and Nishikawa, Hitoshi and Obari, Hiroyuki. 2017. Controlling item difficulty for automatic vocabulary question generation. *Research and Practice in Technology Enhanced Learning*, 12(1):25–40. doi:10.1186/s41039-017-0065-5.

Suzuki, Yuichi. 2022. Automatization and practice. In Aline Godfroid and Holger Hopp (Eds.), *The Routledge Handbook of Second Language Acquisition and Psycholinguistics*, pages 308–321. doi:10.4324/9781003018872-29.

Suzuki, Yuichi. 2023. Introduction: Practice and automatization in a second language. In Yuichi Suzuki (Ed.), *Practice and Automatization in Second Language Research*, pages 1–36. doi:10.4324/9781003414643-1.

Suzuki, Yuichi and Nakata, Tatsuya and DeKeyser, Robert M. 2019. The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *Modern Language Journal*, 103(3):713–720. doi:10.1111/modl.12585.

Suzuki, Yuichi and Yokosawa, Satoko and Aline, David. 2022. The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax. *Language Teaching Research*, 26(4):671–695. doi:10.1177/1362168820913985.

Svetashova, Yulia. 2015. C-test item difficulty prediction: Exploring the linguistic characteristics of C-tests using machine learning. Master's thesis, Eberhard Karls Universität Tübingen.

Swanson, David and Holtzman, Kathleen and Allbee, Krista and Clauser, Brian. 2006. Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Academic Medicine*, 81(10):52–55. doi:10.1097/01.ACM.0000236518.87708.9d.

Tabari, Mahmoud Abdi and Cho, Minyoung. 2022. Task sequencing and L2 writing development: Exploring the SSARC model of pedagogic task sequencing. *Language Teaching Research*, pages 1–29. doi:10.1177/13621688221090922.

Tafazoli, Dara and Gómez Parra, María Elena and Huertas Abril, Cristina. 2019. Intelligent language tutoring system: Integrating intelligent computer-assisted language learning into language education. *International Journal of Information and Communication Technology Education*, 15(3):60–74. doi:10.4018/IJICTE.2019070105.

Tarone, Elaine. 2018. Interlanguage. In Carol A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*, pages 1–7. doi:10.1002/9781405198431.wbeal0561.pub2.

Tham, Duong My and Nhi, Tran Phuong. 2021. A corpus-based study on reporting verbs used in TESOL research articles by native and non-native writers. *VNU Journal of Foreign Studies*, 37(3):135–148. doi:10.25073/2525-2445/vnufs.4729.

Theis, Sven and Wackermann, Rainer. 2020. Leitfaden zur Erstellung von Lernaufgaben: Naturwissenschaftliche Fächer. Technical report, Qualitäts- und UnterstützungsAgentur – Landesinstitut für Schule.

Thomas, Michael and Reinders, Hayo. 2013. Deconstructing tasks and technology. In Michael Thomas and Hayo Reinders (Eds.), *Task-Based Language Learning and Teaching with Technology*, Second Language Teaching and Curriculum Center Honolulu, Hawaii: Technical report, pages 1–6.

Toole, Janine and Heift, Trude. 2001. Generating learning content for an intelligent language tutoring system. In *Proceedings of NLP-CALL Workshop at the 10th International Conference on Artificial Intelligence in Education*, AIED'01, pages 1–8. San Antonio, Texas, USA.

Torre, Ilaria and Mercurio, Marco and Torsani, Simone. 2011. Design of adaptive micro-content in Second Language Acquisition. *Journal of E-Learning and Knowledge Society*, 7(3):109–119. doi:10.20368/1971-8829/555.

Troussas, Christos and Chrysafiadi, Konstantina and Virvou, Maria. 2021. Personalized tutoring through a stereotype student model incorporating a hybrid learning style instrument. *Education and Information Technologies*, 26(2):1–13. doi:10.1007/s10639-020-10366-2.

Tversky, Amos. 1964. On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology*, 1(2):386–391. doi:10.1016/0022-2496(64)90010-0.

Twigg, Carol A. 2003. Improving learning and reducing costs: New models for online learning. *Educause review*, 38(5):28–38.

Twomey, Katherine E. and Ranson, Samantha L. and Horst, Jessica S. 2014. That's more like it: Multiple exemplars facilitate word learning. *Infant and Child Development*, 23(2):105–122. doi:10.1002/icd.1824.

Ur, Penny. 2018. PPP: Presentation–Practice–Production. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching*, pages 1–6. doi:10.1002/9781118784235.eelt0092.

Vanderhyde, James. 2019. Scaffolding assignments: How much is just enough? *Journal of Computing Sciences in Colleges*, 34(3):55–63.

Vanlehn, Kurt. 2006. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3):227–265.

Varanoglulari, Feryal and Lopez, Claude and Gansrigler, Eva and Pessanha, Liliane and Williams, Melanie. 2008. Secondary EFL courses. *English Language Teaching*, 62(4):401–419. doi:10.1093/elt/ccn044.

Viviani, Eva and Ramscar, Michael and Wonnacott, Elizabeth. 2022. Go above and beyond: Does input variability affect children's ability to learn spatial adpositions in a novel language? *PCI Registered Reports*. doi:10.17605/OSF.IO/37DXR. [In Principle Acceptance].

Volodina, Elena and Pilán, Ildikó and Borin, Lars and Tiedemann, Therese Lindström. 2014. A flexible language learning platform based on language resources and Web services. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC'14, pages 3973–3978. Reykjavik, Iceland. European Language Resources Association (ELRA).

Vygotsky, Lew Semjonowitsch. 1978. *Mind in Society: Development of Higher Psychological Processes*. doi:10.2307/j.ctvjf9vz4.

Wahlheim, Christopher N. and Finn, Bridgid and Jacoby, Larry L. 2012. Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & Cognition*, 40:703–716. doi:10.3758/s13421-011-0180-2.

Walz, Joel. 1989. Context and contextualized language practice in foreign language teaching. *Modern Language Journal*, 73(2):160–168. doi:10.2307/326571.

Wang, Chun and Nydick, Steven W. 2020. On longitudinal item response theory models: A didactic. *Journal of Educational and Behavioral Statistics*, 45(3):339–368. doi:10.3102/1076998619882026.

Wang, Shan and Liu, Xiaojun and Zhou, Jie. 2022. Readability is decreasing in language and linguistics. *Scientometrics*, 127(8):4697–4729. doi:10.1007/s11192-022-04427-1.

Wang, Ya-huei and Tseng, Ming-Hseng and Liao, Hung-Chang. 2009. Data mining for adaptive learning sequence in English language instruction. *Expert Systems with Applications*, 36(4):7681–7686. doi:10.1016/j.eswa.2008.09.008.

Wauters, Kelly and Desmet, Piet and Van den Noortgate, Wim. 2012. Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58(4):1183–1193. doi:10.1016/j.compedu.2011.11.020.

Webb, Stuart. 2009. The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal: A Journal of Language Teaching and Research*, 40(3):360–376. doi:10.1177/0033688209343854.

Welbl, Johannes and Liu, Nelson F. and Gardner, Matt. 2017. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (Eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, W-NUT'17, pages 94–106. Copenhagen, Denmark. Association for Computational Linguistics. doi:10.18653/v1/W17-4413.

Wetzel, Nicole and Widmann, Andreas and Schröger, Erich. 2009. The cognitive control of distraction by novelty in children aged 7-8 and adults. *Psychophysiology*, 46(3):607–616. doi:10.1111/j.1469-8986.2009.00789.x.

Whitmer, Daphne and Johnson, Cheryl and Marraffino, Matthew. 2022. Examining two adaptive sequencing approaches for flashcard learning: The tradeoff between training efficiency and tong-term retention. In Robert A. Sottilare and Jessica Schwarz (Eds.), *Proceedings of the 4th International Conference on Adaptive Instructional Systems*, AIS'22, pages 126–139. Online. Springer Nature. doi:10.1007/978-3-031-05887-5_10.

Wiener, Seth and Chan, Marjorie K. M. and Ito, Kiwako. 2020. Do explicit instruction and high variability phonetic training improve nonnative speakers' Mandarin tone productions? *Modern Language Journal*, 104(1):152–168. doi:10.1111/modl.12619.

Wilson, Eve. 1994. A user-adaptive interface for computer assisted language learning. In Thomas Ottmann and Ivan Tomek (Eds.), *Proceedings of the World Conference on Educational Multimedia and Hypermedia*, ED-MEDIA'94, pages 571–576. Vancouver, British Columbia, Canada. Association for the Advancement of Computing in Education.

Winiharti, Menik. 2012. The difference between modal verbs in deontic and epistemic modality. *Humaniora*, 3(2):532. doi:10.21512/humaniora.v3i2.3396.

Wong, Wynne and Van Patten, Bill. 2003. The evidence is IN: Drills are OUT. *Foreign Language Annals*, 36(3):403–423. doi:10.1111/j.1944-9720.2003.tb02123.x.

Wonnacott, Elizabeth and Boyd, Jeremy K. and Thomson, Jennifer and Goldberg, Adele E. 2012. Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language*, 66(3):458–478. doi:10.1016/j.jml.2011.11.004.

Xiong, Yao and Meurers, Detmar and Quixal, Martí and Gawrilow, Caterina and Rudzewitz, Björn and Heck, Tanja. 2024. Supporting mastery learning by adapting exercise sequences in English as a foreign language to working memory and learner performance.

Xue, Kang and Yaneva, Victoria and Runyon, Christopher and Baldwin, Peter. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In Jill Burstein, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch (Eds.), *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'20, pages 193–197. Seattle, Washington, USA (Online). Association for Computational Linguistics. doi:10.18653/v1/2020.bea-1.20.

Yamada, Masaru. 2019. Language learners and non-professional translators as users. In Minako O'Hagan (Ed.), *The Routledge Handbook of Translation and Technology*, chapter 11, pages 183–199. doi:10.4324/9781315311258-11.

Yazdani, Masoud. 1986. Intelligent tutoring systems: An overview. *Expert Systems*, 3(3):154–163. doi:10.1111/j.1468-0394.1986.tb00488.x.

Yeung, Chak Yan and Lee, John Sie Yuen and T'sou, Benjamin Ka-Yin. 2019. Difficulty-aware distractor generation for gap-fill items. In Meladel Mistica, Massimo Piccardi, and Andrew MacKinlay (Eds.), *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, ALTA'19, pages 167–172. Sydney, Australia. Australasian Language Technology Association.

Yildirim, Ilker and Degen, Judith and Tanenhaus, Michael and Jaeger, Florian. 2013. Linguistic variability and adaptation in quantifier meanings. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, volume 35 of *CogSci'13*, pages 3835–3840. Berlin, Germany.

Yilmaz, Maide and Özdem Erturk, Zeynep. 2017. A contrastive corpus-based analy-

sis of the use of reporting verbs by native and non-native ELT researchers. *Novitas-ROYAL*, 11(2):112–127.

Yin, Chengjiu and Hirokawa, Sachio and Flanagan, Brendan and Suzuki, Takahiko and Tabata, Yoshiyuki. 2012. Mistake discovery and generation of exercises automaticity in context. In *Proceedings of the 2012 IIAI International Conference on Advanced Applied Informatics*, IIAIAAI'12, pages 163–167. Fukuoka, Japan. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/IIAI-AAI.2012.41.

Yudelson, Michael V. and Koedinger, Kenneth R. and Gordon, Geoffrey J. 2013. Individualized Bayesian Knowledge Tracing odels. In H. Chad Lane, Kalina Yacef, Jack Mostow, and Philip Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education*, AIED'13, pages 171–180. Memphis, Tennessee, USA. Springer Nature. doi:10.1007/978-3-642-39112-5_18.

Zaidi, Ahmed and Caines, Andrew and Moore, Russell and Buttery, Paula and Rice, Andrew. 2020. Adaptive forgetting curves for spaced repetition language learning. In Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán (Eds.), *Proceedings of the 21st International Conference on Artificial Intelligence in Education*, AIED'20, pages 358–363. Ifrane, Morocco. Springer Nature. doi:10.48550/arXiv.2004.11327.

Zeng, Liren and Lin, Ling. 2011. An interactive vocabulary learning system based on word frequency lists and Ebbinghaus' curve of forgetting. In *Proceedings of the 2011 Workshop on Digital Media and Digital Content Management*, DMDCM'11, pages 313–317. Hangzhou, China. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/DMDCM.2011.71.

Zesch, Torsten and Melamud, Oren. 2014. Automatic generation of challenging distractors using context-sensitive inference rules. In Joel Tetreault, Jill Burstein, and Claudia Leacock (Eds.), *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications*, BEA'14, pages 143–148. Baltimore, Maryland, USA. Association for Computational Linguistics. doi:10.3115/v1/W14-1817.

Zheng, Lanqin. 2016. The effectiveness of self-regulated learning scaffolds on academic performance in computer-based learning environments: A meta-analysis. *Asia Pacific Education Review*, 17(2):187–202. doi:10.1007/s12564-016-9426-9.

Zhu, Dandan. 2020. Programming of English word review planning time based on Ebinhaus forgetting curve. In *Proceedings of the International Conference on Intelligent Transportation, Big Data & Smart City*, ICITBS'20, pages 901–904. Vientiane, Laos. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/ICITBS49701.2020.00199.

Žitko, Branko and Stankov, Slavomir and Rosić, Marko and Grubišić, Ani. 2009. Dynamic test generation over ontology-based knowledge representation in authoring shell. *Expert Systems with Applications*, 36(4):8185–8196. doi:10.1016/j.eswa.2008.10.028.

Zúñiga, Eulices Córdoba. 2016. Implementing task-based language teaching to integrate language skills in an EFL program at a Colombian university. *Profile: Issues in Teachers' Professional Development*, 18:13–27. doi:10.15446/profile.v18n2.49754.

Glossary

actionable element	Part of an exercise for which the learner needs to provide an answer and receives a scoring point if the provided answer is correct.
answer hypothesis	Correct and incorrect answers to an exercise that learners might provide.
context-dependent exercise difficulty	Measure of external factors that influence performance on an exercise irrespective of the learner.
core annotations	Linguistic annotations that must be present in an exercise item in order for the item to practice a linguistic construct.
discriminative annotations	Linguistic annotations that demarcate different variations of the same linguistic construct.
environment annotations	Linguistic annotations that must be present in another item of the same exercise in order for the exercise item to practice a linguistic construct.
errors	Systematic incorrect productions (Corder, 1967).

general exercise difficulty	Measure of static exercise characteristics that influence performance independently of the learner.
learner-dependent exercise difficulty	Measure of learner characteristics that influence performance on an exercise.
learning target	Pedagogically motivated high-level language means that constitutes a grammar or vocabulary unit of a curriculum.
linguistic construct	Unit at the lowest, and only linguistic, level of the domain model defined through linguistic annotations.
macro-adaptivity	Item selection strategy to adjust practice items to the learner model with the aim of individualizing practice (Pelánek, 2017).
meso-adaptivity	Dynamic adjustment of an exercise in response to learner errors with the aim of making the exercise easier.
meta-exercise	Abstract exercise definition that accumulates (1) all items of the same exercise type and which practice the same KC and (2) information common to all these items.
micro-adaptivity	Scaffoldig support within an exercise in response to learner errors with the aim of assisting learners in finding the correct solution (Pelánek, 2017).
mistakes	Momentary and random incorrect productions (Corder, 1967).
property	Unit at the lowest pedagogical level of the domain model that defines a characteristic common to multiple linguistic constructs of the linguistic level underneath.

realization of a learning target	Pedagogically motivated unit at the second highest level of the domain model that cannot be substituted by any other unit of this level in order to achieve a functional goal.

Appendix A

A.1 Approvals for the Variability Study

A.1.1 Governmental Approval

In Germany, any scientific study conducted in schools must be approved by the government. In the federal state Baden-Württemberg, studies that encompass multiple schools within the same school district, as is the case for our variability study, need to be approved by the regional council of that district (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2002). We here provide the governmental approval granted for our study.

Baden-Württemberg
REGIERUNGSPRÄSIDIUM TÜBINGEN
SCHULE UND BILDUNG

Regierungspräsidium Tübingen · Postfach 26 66 · 72016 Tübingen

Tübingen 03.07.2024
Name Aberle, Claudia
Durchwahl [redacted]
Geschäftszeichen [redacted]
(Bitte bei Antwort angeben)

Leibniz-Institut f. Wissensmedien
Prof. Dr. Detmar Meurers
Schleichstraße 6
72076 Tübingen

Genehmigung der wissenschaftlichen Erhebung "Transferierbarkeit und Automatisierung von sprachlichem Wissen durch variables Üben"

Ihr Antrag vom 01.07.2024

Sehr geehrte Herr Prof. Dr. Detmar Meurers,

das Regierungspräsidium Tübingen dankt Ihnen für Ihren Antrag und genehmigt die o. g. geplante wissenschaftliche Erhebung.

Die Genehmigung erfolgt gemäß der Verwaltungsvorschrift „Werbung, Wettbewerbe und Erhebungen in Schulen“ vom 21. September 2002 (K. u. U. 2002, Seite 309) - zuletzt geändert durch Verwaltungsvorschrift vom 28.10.2005 - anhand der von Ihnen vorgelegten Unterlagen.

Folgende Maßgaben sind zu beachten:

Genehmigungsbereich:

- Die Genehmigung gilt nur für die Befragung von Schüler-/innen und Lehrkräften (Klassenstufe 7, Allgemeinbildende Gymnasien) **im Regierungsbezirk Tübingen**.

- Bei beabsichtigter Teilnahme von Schulen **außerhalb** des Regierungsbezirks Tübingen **verliert diese Genehmigung an Gültigkeit** und es bedarf einer entsprechenden Genehmigung durch das Kultusministerium Baden-Württemberg.

Freiwilligkeit der Teilnahme, Einverständniserklärung, Organisatorisches:

- **Informationsschreiben an alle Beteiligten** (Schulen, Lehrkräfte, Schüler/-innen, Eltern / Erziehungsberechtigte):

  In allen Informationsschreiben muss auf folgende Punkte deutlich hingewiesen werden:
  - Mit der Genehmigung durch das Regierungspräsidium Tübingen ist **keine wissenschaftliche Qualitätskontrolle** verbunden, sondern die Prüfung erfolgte nur anhand der Vorgaben der Verwaltungsvorschrift „Werbung, Wettbewerbe und Erhebungen in Schulen“. Die Zustimmung des Regierungspräsidiums Tübingen bezieht sich auf die Studie insgesamt und nicht konkret auf jede einzelne Frage im Fragebogen.
  - Mit der Genehmigung durch das Regierungspräsidium Tübingen ist **keine Aufforderung zur Teilnahme und/oder zur Beantwortung aller Fragen** verbunden.
  - Die Teilnahme aller Beteiligten beruht auf **Freiwilligkeit**.
  - Eine **Nichtteilnahme bzw. der jederzeitige Abbruch** muss möglich sein und ist mit **keinerlei Nachteilen oder Konsequenzen** für die Beteiligten verbunden.

  In den Informationsschreiben ist auch über die Art der Untersuchung zu informieren, darüber wer die Daten erhebt, wo und wie lange diese gespeichert und von wem und wie sie ausgewertet werden (in anonymisierter, keine Rückschlüsse auf Einzelne zulassender Form).

- **Einverständniserklärung der Eltern bzw. der volljährigen Schülerinnen und Schüler (sofern als Zielgruppe der Erhebung relevant):**
  Wenn volljährige Schüler/-innen nicht an der Erhebung teilnehmen möchten bzw. bei minderjährigen Schüler/-innen ihre Eltern/Erziehungsberechtigte dies nicht wünschen, müssen sie dies nicht eigens zum Ausdruck bringen. Sie haben daher keine Erklärung auszufüllen und auch kein „Nein“-Kästchen o. ä. anzukreuzen.

Schweigen gilt hier nicht als Zustimmung. Deshalb können nur Kinder/ Jugendliche an der Erhebung teilnehmen, die selbst (bei Volljährigkeit) oder deren Eltern/ Erziehungsberechtigte ausdrücklich zugestimmt haben (z. B. durch Ankreuzen eines „Ja“-Kästchens o. ä.).

- **Unzulässigkeit von Incentives:**
  Kein Beteiligter darf für seine Teilnahme an einer wissenschaftlichen Erhebung eine finanzielle Aufwandsentschädigung erhalten (Gutscheine, Geschenke oder Geld). Damit soll verhindert werden, dass sich die Beteiligten auf Grund eines finanziellen Anreizes für die Teilnahme entscheiden.

- **Erhebung während der Unterrichtszeit (falls erforderlich):**
  Die Erhebung kann während der Unterrichtszeit durchgeführt werden, wenn die organisatorischen Rahmenbedingungen vor Ort gegeben sind und die Erreichung der Unterrichtsziele gewährleitet ist. Es liegt im Ermessen der Schulen, ob die Erhebung nur außerhalb des Unterrichts oder auch während der Unterrichtszeit durchgeführt werden kann. Die Entscheidung trifft die jeweilige Schulleitung.

Datenschutz:

- Die Genehmigung durch das Regierungspräsidium Tübingen umfasst **keine** abschließende wissenschaftliche Qualitätskontrolle. Die Prüfung erfolgte vorrangig nach rechtlichen, insbesondere datenschutzrechtlichen Kriterien.

- Aus der Darstellung des Untersuchungsergebnisses dürfen keine Rückschlüsse auf Einzelpersonen möglich sein. Es dürfen zudem keine personen-, klassen- oder schulbezogenen Ergebnisse veröffentlicht werden.

- Es bestehen keine Bedenken, wenn die Datenschutz- und Datensicherheitsregelungen eingehalten werden, insbesondere wenn

  - über die Art der Untersuchung informiert wird,
  - aus den Anschreiben klar hervorgeht, wer die Daten erhebt, wo und wie lange diese gespeichert und von wem und wie sie ausgewertet werden,
  - die Anonymisierung (= keine Rückschlüsse auf Einzelne) und die Zugriffsbeschränkungen entsprechend umgesetzt werden

  - die Listen entsprechend sicher gelagert und gegen Zugriff bzw. Zugang durch Dritte gesichert werden und
  - die personenbezogenen Daten nur solange gespeichert werden, wie es für den Erhebungszweck erforderlich ist.

- Die bei der Befragung gewonnenen Daten dürfen nur für den Zweck, der Grundlage der Genehmigung war, verwendet werden.

Das Regierungspräsidium Tübingen wünscht Ihnen einen erfolgreichen Verlauf Ihres Projektes.
Da auch auf unserer Seite Interesse an den Ergebnissen Ihrer Untersuchung besteht, möchten wir Sie bitten, uns diese nach Abschluss der Untersuchung zur Verfügung zu stellen.

Bitte richten Sie künftige Anfragen bezüglich wissenschaftlicher Untersuchungen **mittels Formular** direkt an die Unterzeichnerin ([redacted]) und planen Sie mindestens sechs Wochen Bearbeitungszeit ein. Da vor der Genehmigung eine hausinterne Abstimmung notwendig ist, kann nur mit genügend Vorlauf eine rechtzeitige Bearbeitung Ihres Antrags gewährleistet werden.

Mit freundlichen Grüßen

Claudia Aberle

Informationen zum Schutz personenbezogener Daten finden Sie auf unserer Internetseite unter
Datenschutzerklärung zur Verwaltungstätigkeit der Regierungspräsidien
Auf Wunsch werden diese Informationen in Papierform versandt.

A.1.2 Ethics Approval

In order to ensure that our variability study complied with current ethical standards, we sought approval for the study with the Ethics committee of the Leibniz-Institut für Wissensmedien in Tübingen (Leibniz-Institut für Wissensmedien, 2019). We here provide the ethics approval granted for our study.

Leibniz-Institut für Wissensmedien

Stiftung Medien in der Bildung (SbR)
Leibniz-Institut für Wissensmedien

Direktorin: Prof. Dr. Ulrike Cress
Schleichstraße 6
72076 Tübingen

Ethikkommission

Prof. Dr. Sonja Utz
Tel +49 7071 979[redacted]
Fax +49 7071 979[redacted]
[redacted]
www.iwm-tuebingen.de

Leibniz-Institut für Wissensmedien / Schleichstraße 6 / 72076 Tübingen

Prof. Dr. Walt Detmar Meurers
Stiftung Medien in der Bildung (SbR)
Leibniz-Institut für Wissensmedien
Schleichstraße 6
72076 Tübingen

Tübingen, 11.07.2024

Confirmation

As chairwoman of the Ethics Committee of the Leibniz-Institut für Wissensmedien I herewith confirm that the research project "Transferierbarkeit und Automatisierung von sprachlichem Wissen durch variables Üben", as submitted by Tanja Heck and Prof. Dr. Walt Detmar Meurers (Leibniz-Institut für Wissensmedien) for evaluation, was classified as ethically defensible.

Conditions:
Please refer to the expert opinions for conditions.

The file is stored under LEK 2024/027.

We wish all involved success for their research.

[redacted]

(Prof. Dr. Sonja Utz)

Vorstand:
Prof. Dr. Ulrike Cress
Prof. Dr. Stephan Schwan
Dr. Robert Polgar
IBAN [redacted]
BIC SOLADES1TUB
USt-IdNr. [redacted]

A.2 Study Material for the Variability Study

A.2.1 Pre-test Exercises

Lies den Satz und stelle so viele Fragen wie möglich, die der Satz beantworten könnte.[1]

1 The rock music annoys the neighbours in the house at night because it is very loud.

2 Use the chat and ask your friends if you don't know the answer.

3 Carlos washes his dad's car every weekend in the driveway with a sponge.

4 Use the blue balls, they belong to the school.

A.2.2 Explanations

Fragen werden für die meisten Verben mit *do + Subjekt + Inifintiv* gebildet.[2]
Fragen mit Modalverben werden mit *Verb + Subjekt* gebildet.[3]
Zu den Modalverben zählen unter anderem: *can, must, would, will, could, shall, may, might.*[4]
Fragen, in denen das Fragewort das Subjekt ersetzt, werden mit *Fragewort + Verb* gebildet.[5]

[1] English translation: *Read the sentence and ask as many questions as possible that the sentence could answer.*

[2] English translation: *Most questions are formed with* do + subject + infinitive.

[3] English translation: *Questions with modal verbs are formed with* verb + subject.

[4] English translation: *Modal verbs include, among others:* can, must, would, will, could, shall, may, might.

[5] English translation: *Questions in which the question word substitutes the subject are formed with* question word + verb.

A.2.3 Interim Test Exercises

Lies den Satz und stelle so viele Fragen wie möglich, die der Satz beantworten könnte.[6]

1 The sun burns the tourists at the beach during the day because it is very hot.

2 Open the book and check the solution page if you think that your answer is correct.

3 Ricky visits his mum's best friend in France every year by train.

4 Watch the movie about the baby animals, it's Sally's favourite.

[6]English translation: *Read the sentence and ask as many questions as possible that the sentence could answer.*

A.2.4 Practice Exercises

Vervollständige den Satz mit dem Verb und dem Subjekt in den Klammern, um eine Frage in der Gegenwart zu bilden. Frage nach dem unterstrichenen Teil der Antwort.[7]

	Control group:		**Intervention group:**	
1	whom + do	*Whom does Peter invite* (invite, Peter) to the cinema on Saturday? - Peter invites <u>his girlfriend</u> to the cinema on Saturday.	whom + do	*Whom does Peter invite* (invite, Peter) to the cinema on Saturday? - Peter invites <u>his girlfriend</u> to the cinema on Saturday.
2	obj. what + modal *can* (epistemic)	*What can you reach* (can reach, you) with a ladder? - I can reach <u>the book</u> with a ladder.	how + *be able to* (epistemic)	*How are you able to reach* (able to reach, you) the book? - I'm able to reach the book <u>with a ladder</u>.
3	whom + do	*Whom do Sarah's parents call* (call, Sarah's parents) at school? - Sarah's parents call <u>the teacher</u> at school.	subj. who + do	*Who calls the teacher* (call, the teacher) at school? - <u>Sarah's parents</u> call the teacher at school.
4	obj. what + do	*What do many animals eat* (eat, many animals)? - Many animals eat <u>meat</u>.	obj. what + do	*What do many animals eat* (eat, many animals)? - Many animals eat <u>meat</u>.
5	whom + *have to* (epistemic)	*Whom does Peter have to like* (have to like, Peter) a lot? - Peter has to like <u>Sarah</u> a lot because he gives her lots of presents.	subj. who + modal *must* (epistemic)	*Who must like Sarah* (must like, Sarah) a lot? - <u>David</u> must like Sarah a lot because he gives her lots of presents.

[7]English translation: *Complete the sentence with the verb and the subject in parentheses in order to form a question in the present tense. Ask about the underlined part of the answer.*

6	obj. what + *have to* (deontic)	*What does Kate have to clean* (have to clean, Kate) because it's dirty? - Kate has to clean her room because it's dirty.	why + modal *must* (deontic)	*Why must Kate clean* (must clean, Kate) her room? - Kate must clean her room because it's dirty.
7	obj. what + *have to* (deontic)	*What does Alex have to practice* (have to practice, Alex), Maths or English? - Alex has to practice English.	which + *have to* (deontic)	*Which subject does Alex have to practice* (have to practice, Alex, subject), Maths or English? - Alex has to practice English.
8	obj. what + *be*	*What are bad students* (be, bad students) like? - Bad students are lazy.	obj. what + *be*	*What are bad students* (be, bad students) like? - Bad students are lazy.
9	obj. what + modal *can* (deontic)	*What can Paul borrow* (can borrow, Paul) during the test? - Paul can borrow Harry's pencil during the test.	when + modal *may* (deontic)	*When may Paul borrow* (may borrow, Paul) Harry's pencil? - Paul may borrow Harry's pencil during the test.
10	whom + *have to* (deontic)	To *whom does the teacher have to speak* (have to speak, the teacher) after class? - The teacher has to speak to the class president after class.	whom + *need to* (deontic)	To *whom does the teacher need to speak* (need to speak, the teacher) after class? - The teacher needs to speak to the class president after class.
11	whom + *be*	To *whom is Sandra* (be, Sandra) always nice? - Sandra is always nice to her grandmother.	whom + *be*	To *whom is Sandra* (be, Sandra) always nice? - Sandra is always nice to her grandmother.

12	whom + modal *can* (deontic)	*Whom can Charlie help* (can help, Charlie) with the homework? - Charlie can help his best friend with the homework.	whom + modal *can* (deontic)	*Whom can Charlie help* (can help, Charlie) with the homework? - Charlie can help his best friend with the homework.
13	obj. what + do	*What does the cook cut* (cut, the cook) with a knife? - The cook cuts the vegetables with a knife.	how + do	*How does the cook cut* (cut, the cook) the vegetables? - The cook cuts the vegetables with a knife.
14	obj. what + *be*	*What is Rome* (be, Rome) like in summer? - Rome is very hot in summer.	how + *be*	*How is Rome* (be, Rome) in summer? - Rome is very hot in summer.
15	obj. what + modal *can* (epistemic)	*What can the weather be* (can be, the weather) like on Sunday if the clouds disappear? - The weather can still be good on Sunday if the clouds disappear.	how + modal *may* (epistemic)	*How may the weather be* (may be, the weather) on Sunday if the clouds disappear? - The weather may still be good on Sunday if the clouds disappear.
16	obj. what + modal *can* (deontic)	*What can the teacher explain* (can explain, the teacher) by showing pictures? - The teacher can explain the topic by showing pictures.	how + modal *should* (deontic)	*How should the teacher explain* (should explain, the teacher) the topic? - The teacher should explain the topic by showing pictures.
17	obj. what + do	*What do Abby and Sophie play* (play, Abby and Sophie) at the concert? - Abby and Sophie play the piano at the concert.	subj. who + do	*Who plays the piano* (play, the piano) at the concert? - Abby and Sophie play the piano at the concert.

18	obj. what + *have to* (epistemic)	*What does Adrian's dad have to earn* (have to earn, Adrian's dad) ? - Adrian's dad has to earn a lot of money. They fly to New York every year.	subj. who + *have to* (epistemic)	*Who has to earn a lot of money* (have to earn, a lot of money) ? - Adrian's dad has to earn a lot of money. They fly to New York every year.
19	obj. what + *be*	*What is Tina* (be, Tina) to Hannah? - Tina is a good friend to Hannah.	subj. who + *be*	*Who is a good friend* (be, a good friend) to Hannah? - Tina is a good friend to Hannah.
20	obj. what + *have to* (deontic)	*What does Eric's mother have to buy* (have to buy, Eric's mother) at the supermarket? - Eric's mother has to buy vegetables at the supermarket.	obj. what + *need to* (deontic)	*What does Eric's mother need to buy* (need to buy, Eric's mother) at the supermarket? - Eric's mother needs to buy vegetables at the supermarket.
21	obj. what + modal *can* (epistemic)	*What can be* (can be) inside the box? - There can be books inside the box.	obj. what + modal *can* (epistemic)	*What can be* (can be) inside the box? - There can be books inside the box.
22	whom + do	*Whom does Sam drive* (drive, Sam) to school? - Sam drives his sister to school because he's older.	why + do	*Why does Sam drive* (drive, Sam) his sister to school? - Sam drives his sister to school because he's older.
23	obj. what + *be*	*What is the door* (be, the door) always because the key is lost? - The door is always open because the key is lost.	why + *be*	*Why is the door* (be, the door) always open? - The door is always open because the key is lost.

24	obj. what + modal *can* (epistemic)	*What can Daniel reach* (can reach, Daniel) because he's very tall? - Daniel can reach the branch because he's very tall.	why + *be able to* (epistemic)	*Why is Daniel able to reach* (be able to reach, Daniel) the branch? - Daniel is able to reach the branch because he's very tall.
25	obj. what + do	*What does the class plan* (plan, the class) after the Easter holidays? - The class plans the school trip after the Easter holidays.	when + do	*When does the class plan* (plan, the class) the school trip? - The class plans the school trip after the Easter holidays.
26	obj. what + *have to* (epistemic)	*What does the baby have to be* (be, the baby) at night? - The baby has to be lonely at night.	when + *have to* (epistemic)	*When does the baby have to be* (be, the baby) lonely? - The baby has to be lonely at night.
27	obj. what + *be*	*What is Sabrina* (be, Sabrina) always afraid of? - Sabrina is always afraid of spiders.	when + *be*	*When is Sabrina* (be, Sabrina) afraid of spiders? - Sabrina is always afraid of spiders.
28	obj. what + do	*What does Andrew like* (like, Andrew) better than the white t-shirt? - Andrew likes the black t-shirt best.	which + do	*Which t-shirt does Andrew like* (like, Andrew, t-shirt) better? - Andrew likes the black t-shirt better than the white one.
29	obj. what + *be*	*What is* (be) the vegetarian sandwich? - The vegetarian sandwich is most delicious.	which + *be*	*Which sandwich is* (be, sandwich) most delicious? - The vegetarian sandwich is most delicious.
30	obj. what + modal *can* (epistemic)	*What can the short dress be* (can be, the short dress) in warm weather? - The short dress can be better in warm weather.	which + modal *should* (epistemic)	*Which dress should be* (should be, dress) better in warm weather? - The short dress should be better in warm weather.

A.2.5 Post-test Exercises

A.2.5.1 Half-open Exercises

Lies den Satz und stelle so viele Fragen wie möglich, die der Satz beantworten könnte.[8]

1 The big fire keeps the campers warm outside at night because it produces a lot of heat.

2 Use you phones and call me if you get lost.

3 Pamela plants her grandmother's flowers in the garden in spring with a small shovel.

4 Go into the blue room, that's Tom's.

A.2.5.2 Fill-in-the-Blanks Exercises

`Linguistic constructs` **practiced by none of the study conditions**:

1	obj. what + modal *shall*	*What shall the students use* (shall use, the students) on the school trip? - The students shall use their own sleeping bags on the trip.
2	obj. what + modal *could* (deontic)	*What could Evan drink* (could drink, Evan) because he's 18 now? - Evan could drink whiskey because he's 18 now.
3	obj. what + modal *could* (epistemic)	*What could the football team win* (could win, the football team) next week? - The football team could win a game next week. They practice every day.

[8] English translation: *Read the sentence and ask as many questions as possible that the sentence could answer.*

4	obj. what + modal *might*	*What might Mary sing* (might sing, Mary) at the school concert? - Mary might sing her favourite song at the school concert.
5	obj. what + *be allowed to*	*What is Alice allowed to bring* (be allowed to bring, Alice) to her friend's party? Alice is allowed to bring her dog to her friend's party.
6	whose + modal *can* (circumstantial)	*Whose sister can play* (can play, sister) the piano? - Joe's sister can play the piano.
7	where + modal *can* (deontic)	*Where can the headmaster park* (can park, the headmaster) his car? - The headmaster can park his car in front of the school.
8	subj. what + modal *can* (epistemic)	*What can cross the road* (can cross, the road) in the countryside sometimes? Animals can sometimes cross the road in the countryside.
9	where + do	*Where does Emma find* (find, Emma) beautiful flowers? Emma finds beautiful flowers in the garden.
17	subj. what + do	*What makes many people* (make, many people) happy? - Ice cream makes many people happy.
11	whose + do	*Whose friend plays* (play, friend) in the basketball match? - Jack's friend plays in the basketball match.
`Linguistic constructs` **practiced by the intervention condition**:		
12	how + *be able to* (epistemic)	*How are we able to cross* (be able to cross, we) the river? - We're able to cross the river by boat.

13	which + modal *should* (deontic)	*Which cake should Pete's mother make* (should make, Pete's mother, cake) for the school event, the one with chocolate cream or the one with sugar icing? - Pete's mother should make the cake with chocolate cream for the school event.
14	subj. who + do	*Who helps the old lady* (help, the old lady) to cross the road? - Simon and Alex help the old lady to cross the road.
15	why + modal *must* (deontic)	*Why must Ethan practice* (must practice, Ethan) his presentation? - Ethan must practice his presentation because he's nervous.
16	when + modal *may* (denotic)	*When may Sophia use* (may use, Sophia) Cindy's book? Sophia may use Cindy's book during lunch break.
17	when + modal *may* (epistemic)	When may northern lights be (may be, northern lights) visible in the sky? - Northern lights may be visible in the sky in winter.
18	how + *need to* (deontic)	*How does Sophie need to lift* (need to lift, Sophie) the box? - Sophie needs to lift the box with a rope.
19	subj. who + modal *must* (epistemic)	*Who must be responsible* (must be, responsible) for the bake sale? - Sally's mom must be responsible for the bake sale. She does all the planning.

20	why + modal *should* (epistemic)	*Why should the class test be* (should be, the class test) easy for Max? - The class test should be easy for Max because he practiced a lot for it.
21	which + do	*Which restaurant has* (have, restaurant) cheaper food? - The Chinese restaurant has cheaper food than the Greek restaurant.
22	how + *be*	*How is the food* (be, the food) at the party? - The food at the party is very unhealthy.
`Linguistic constructs` **practiced by both study conditions**:		
23	whom + *be*	For *whom is the present* (be, the present)? - The present is for Kelly.
24	obj. what + do	*What does David play* (play, David) in the school band? - David plays the drums in the school band.
25	obj. what + modal *can* (epistemic)	*What can traffic jams block* (can block, traffic jams) during summer vacations? - Traffic jams can block the roads during summer vacations.
26	whom + have to (epistemic)	To *whom does the car have to belong* (belong, the car) ? - The car has to belong to Siena's dad.
27	whom + have to (deontic)	*Whom does Andrew have to take* (have to take, Andrew) to the concert next month? Andrew has to take his sister to the concert next month.
28	subj. what + do	*What do most people do* (do, most people) on a Sunday morning? Most people sleep in on a Sunday morning.

29	obj. what + be	*What is the water* (be, the water) like in the Caribbean? - The water is deep blue in the Caribbean.
30	whom + modal *can* (deontic)	*Whom can Sarah visit* (can visit, Sarah) during the holidays? - Sarah can visit her grandparents during the holidays.

A.3 Detailed Evaluation Results

A.3.1 Number of Exercises Practicing a `Property`

The number of exercises practicing each property was determined in the context of evaluating whether exercises can be linked automatically to a `realization`. `Properties` practiced by an exercise were identified based on automatically determined linguistic annotations of the exercises.

`Property`	**n exercises**
Positive statements in the simple past with regular verbs	63
Positive statements in the simple past with irregular verbs	4
Positive statements in the simple past with modals	0
Negative statements in the simple past with regular verbs	21
Negative statements in the simple past with irregular verbs	6
Negative statements in the simple past with modals	0
Mixed statements in the simple past with regular verbs	60
Mixed statements in the simple past with irregular verbs	15
Mixed statements in the simple past with modals	0
Affirmative conditional type 1	75
Negative conditional type 1	114
Affirmative conditional type 2	75
Negative conditional type 2	100
Affirmative mixed conditionals	45
Negative mixed conditionals	136
Positive Yes/No questions in the simple present with ‘be’	26
Positive Yes/No questions in the simple present with do-support	33
Positive Yes/No questions in the simple present with modals	35
Negative Yes/No questions in the simple present with ‘be’	10
Negative Yes/No questions in the simple present with do-support	8
Negative Yes/No questions in the simple present with modals	8

Continued on next page

Table A.1: Number of exercises per `property`

Continued from previous page

`property`	**n exercises**
Positive Yes/No questions in the simple past with 'be'	15
Positive Yes/No questions in the simple past with do-support	35
Positive Yes/No questions in the simple past with modals	34
Negative Yes/No questions in the simple past with 'be'	8
Negative Yes/No questions in the simple past with do-support	0
Negative Yes/No questions in the simple past with modals	8
Positive wh-questions in the simple present with 'be'	25
Positive wh-questions in the simple present with do-support	21
Positive subject wh-questions in the simple present	83
Positive wh-questions in the simple present with modals	31
Negative wh-questions in the simple present with 'be'	0
Negative wh-questions in the simple present with do-support	0
Negative wh-questions in the simple present with modals	0
Positive wh-questions in the simple past with 'be'	10
Positive wh-questions in the simple past with do-support	8
Positive subject wh-questions in the simple past	79
Positive wh-questions in the simple past with modals	27
Negative wh-questions in the simple past with 'be'	0
Negative wh-questions in the simple past with do-support	0
Negative wh-questions in the simple past with modals	0
Mixed questions in the simple present with 'be'	2
Mixed questions in the simple present with do-support	6
Mixed questions in the simple present with modals	2
Mixed questions in the simple past with 'be'	3
Mixed questions in the simple past with do-support	6
Mixed questions in the simple past with modals	4
Positive statements in the present perfect with regular verbs	63
Positive statements in the present perfect with irregular verbs	8
Negative statements in the present perfect with regular verbs	21
Negative statements in the present perfect with irregular verbs	6

Continued on next page

Table A.1: Number of exercises per `property`

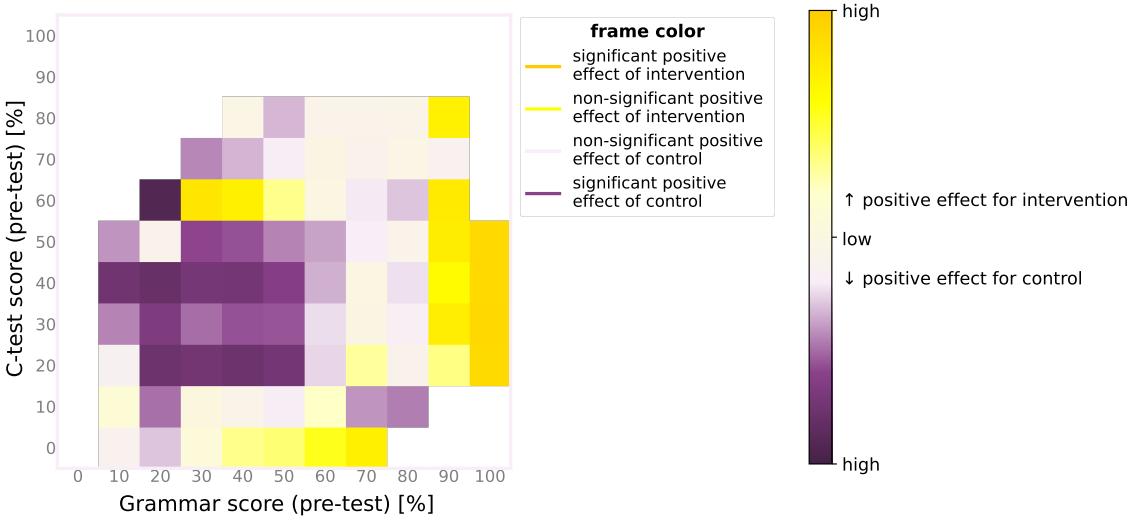
Continued from previous page

property	n exercises
Mixed statements in the present perfect with regular verbs	64
Mixed statements in the present perfect with irregular verbs	8
Simple past vs. present perfect with regular verbs	23
Simple past vs. present perfect with irregular verbs	7
Simple past vs. present perfect with modals	0

Table A.1: Number of exercises per `property`.

A.3.2 Significance of the Meso-adaptivity Treatment per `Learning Target`

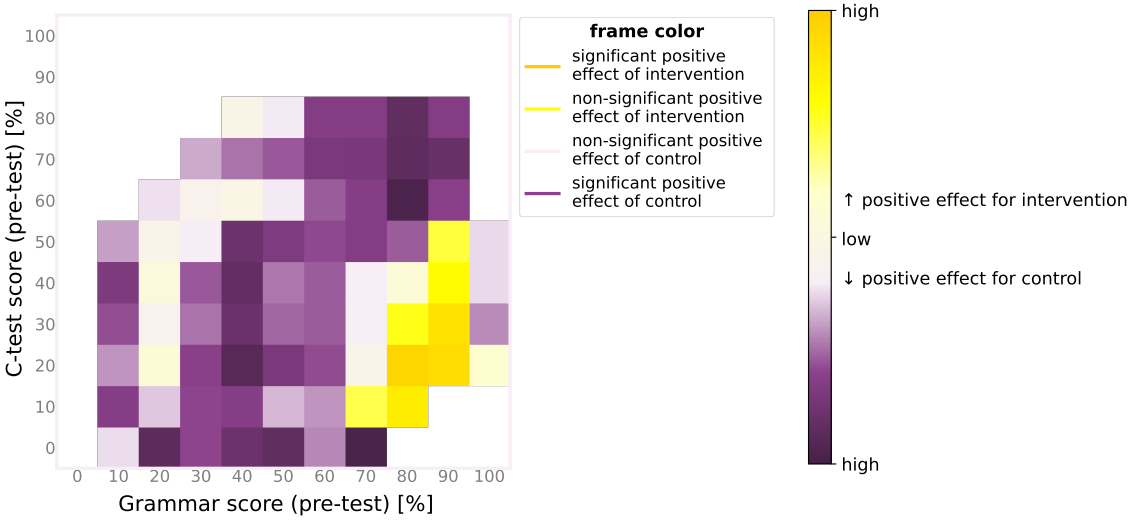
The heatmaps visualize significance values of the effect of the meso-adaptivity treatment. each cell represents the effect for learners within ± 20% of the indicated C-test and grammar test scores. Different colors indicate whether the treatment of meso-adaptivity was beneficial (yellow) or not (purple). Increasing richness of the colors represent increasing significance of the effect. The color of the frame surrounding each heatmap indicates the effect for the entire learner sample with the treatment threshold indicated above the plot.



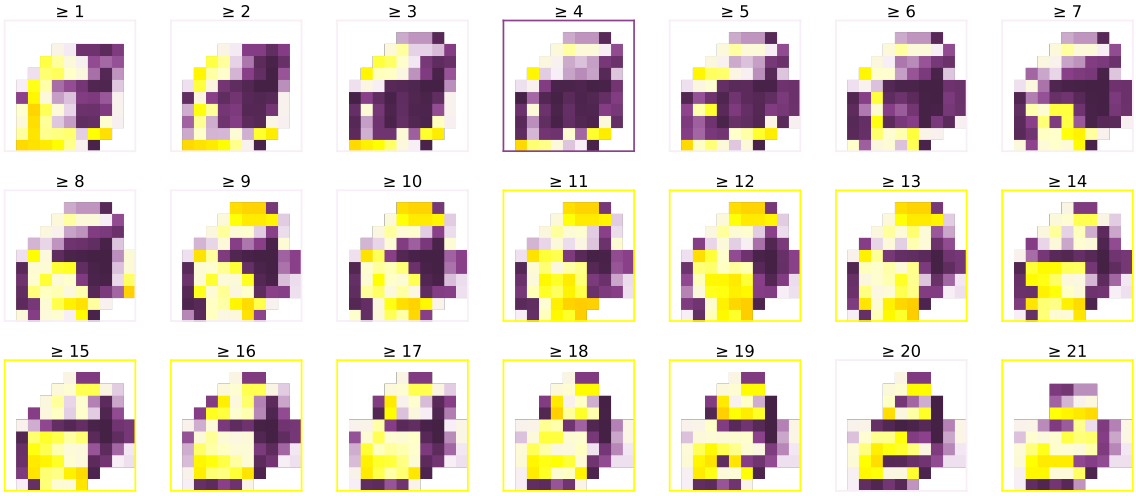
(a) As randomized.

(b) As treated (per number of scaffolded exercises).

Figure A.1: Significance of the treatment for learner subgroups on *Tenses*.

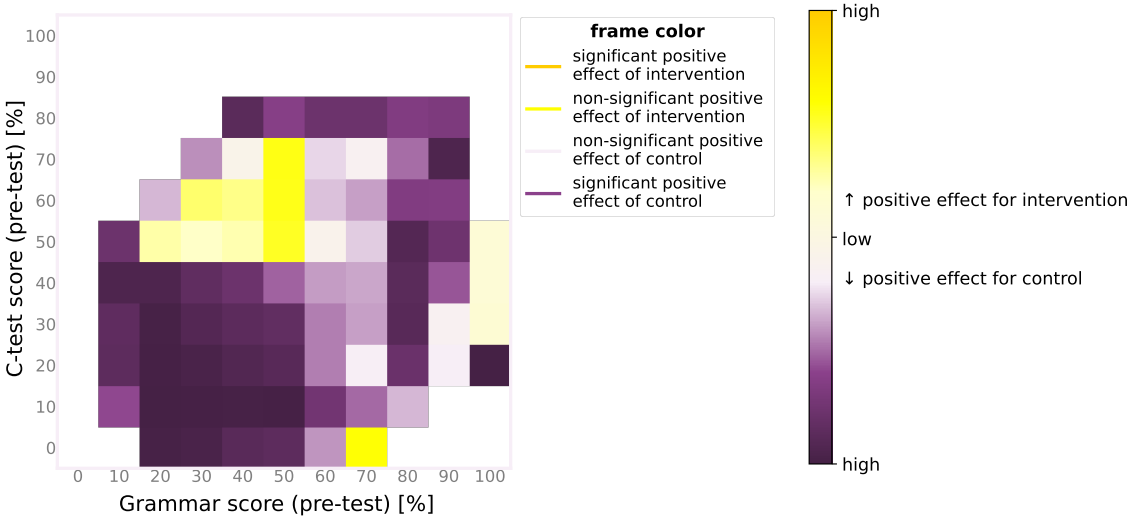


(a) As randomized.

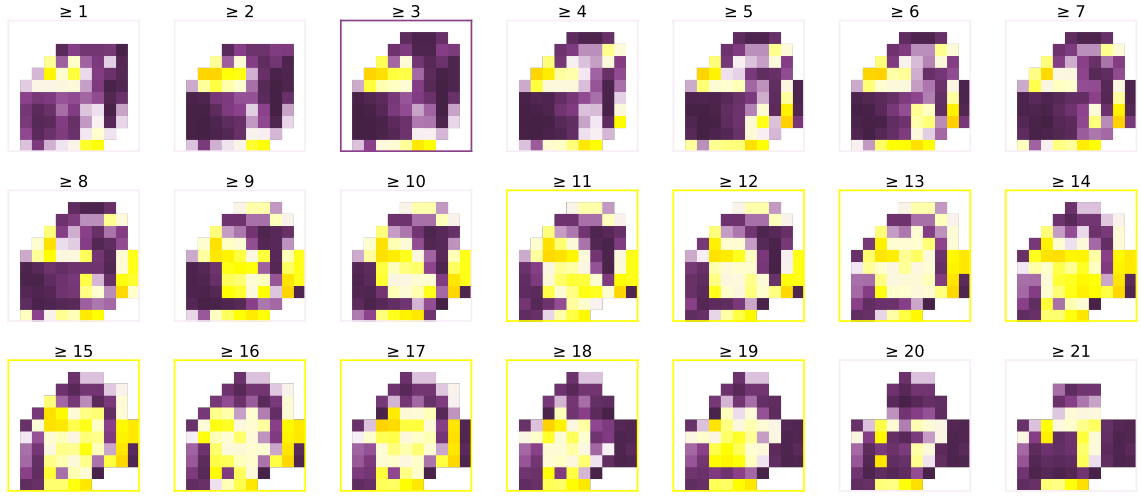


(b) As treated (per number of scaffolded exercises).

Figure A.2: Significance of the treatment for learner subgroups on *Questions*.



(a) As randomized.



(b) As treated (per number of scaffolded exercises).

Figure A.3: Significance of the treatment for learner subgroups on *Conditionals*.