Christina Schneegass

EMBEDDING MOBILE LEARNING INTO EVERYDAY LIFE SETTINGS

Embedding Mobile Learning into Everyday Life Settings

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von

Christina Schneegass
M.Sc. Angewandte Kognitions- und Medienwissenschaften

München, den 13.08.2021

Erstgutachter:	Prof. Dr. Heinrich Hußmann
Zweitgutachter:	Prof. Dr. Gerhard Fischer
Drittgutachter:	Dr. Tilman Dingler

Tag der mündlichen Prüfung: 15. Oktober 2021

Abstract

The increasing ubiquity of smartphones has changed the way we interact with information and acquire new knowledge. The prevalence of personal mobile devices in our everyday lives creates new opportunities for learning that exceed the narrow boundaries of a school's classroom and provide the foundations for lifelong learning. Learning can now happen whenever and wherever we are; whether on the sofa at home, on the bus during our commute, or on a break at work. However, the flexibility offered by mobile learning also creates its challenges. Being able to learn anytime and anywhere does not necessarily result in learning uptake. Without the school environment's controlled schedule and teacher guidance, the learners must actively initiate learning activities, keep up repetition schedules, and cope with learning in interruption-prone everyday environments. Both interruptions and infrequent repetition can harm the learning process and long-term memory retention. We argue that current mobile learning applications insufficiently support users in coping with these challenges.

In this thesis, we explore how we can utilize the ubiquity of mobile devices to ensure frequent engagement with the content, focusing primarily on language learning and supporting users in dealing with learning breaks and interruptions. Following a user-centered design approach, we first analyzed mobile learning behavior in everyday settings. Based on our findings, we proposed concepts and designs, developed research prototypes, and evaluated them in laboratory and field evaluations with a specific focus on user experience.

To better understand users' learning behavior with mobile devices, we first characterized their interaction with mobile learning apps through a detailed survey and a diary study. Both methods confirmed the enormous diversity in usage situations and preferences. We observed that learning often happens unplanned, infrequently, among the company of friends or family, or while simultaneously performing secondary tasks such as watching TV or eating. The studies further uncovered a significant prevalence of interruptions in everyday settings that affected users' learning behavior, often leading to suspension and termination of the learning activities. We derived design implications to support learning in diverse situations, particularly aimed at mitigating the adverse effects of multitasking and interruptions. The proposed strategies should help designers and developers create mobile learning applications that adapt to the opportunities and challenges of learning in everyday mobile settings.

We explored four main challenges, emphasizing that (1) we need to consider that *Learning in Everyday Settings is Diverse and Interruption-prone*, (2) learning performance is affected by *Irregular and Infrequent Practice Behavior*, (3) we need to move *From Static to Personalized Learning*, and (4) that *Interruptions and Long Learning Breaks can Negatively Affect Performance*. To tackle these challenges, we propose to embed learning into everyday smartphone interactions, which could foster frequent engagement with – and implicitly personalize – learning content (according to users' interests

and skills). Further, we investigate how memory cues could be applied to support task resumption after interruptions in mobile learning.

To confirm that our idea of embedding learning into everyday interactions can increase exposure, we developed an application integrating learning tasks into the smartphone authentication process. Since unlocking the smartphone is a frequently performed action without any other purpose, our subjects appreciated the idea of utilizing this process to perform quick and simple learning interactions. Evidence from a comparative user study showed that embedding learning tasks into the unlocking mechanism led to significantly more interactions with the learning content without impairing the learning quality. We further explored a method for embedding language comprehension assessment into users' digital reading and listening activities. By applying physiological measurements as implicit input, we reliably detected unknown words during laboratory evaluations. Identifying such knowledge gaps could be used for the provision of in-situ support and to inform the generation of personalized language learning content tailored to users' interests and proficiency levels.

To investigate memory cueing as a concept to support task resumption after interruptions, we complemented a theoretical literature analysis of existing applications with two research probes implementing and evaluating promising design concepts. We showed that displaying memory cues when the user resumes the learning activity after an interruption improves their subjective user experience. A subsequent study presented an outlook on the generalizability of memory cues beyond the narrow use case of language learning. We observed that the helpfulness of memory cues for reflecting on prior learning is highly dependent on the design of the cues, particularly the granularity of the presented information. We consider interactive cues for specific memory reactivation (e.g., through multiple-choice questions) a promising scaffolding concept for connecting individual micro-learning sessions when learning in everyday settings.

The tools and applications described in this thesis are a starting point for designing applications that support learning in everyday settings. We broaden the understanding of learning behavior and highlight the impact of interruptions in our busy everyday lives. While this thesis focuses mainly on language learning, the concepts and methods have the potential to be generalized to other domains, such as STEM learning. We reflect on the limitations of the presented concepts and outline future research perspectives that utilize the ubiquity of mobile devices to design mobile learning interactions for everyday settings.

Zusammenfassung

Die Allgegenwärtigkeit von Smartphones verändert die Art und Weise wie wir mit Informationen umgehen und Wissen erwerben. Die weite Verbreitung von mobilen Endgeräten in unserem täglichen Leben führt zu neuen Möglichkeiten des Lernens, welche über die engen Grenzen eines Klassenraumes hinausreichen und das Fundament für lebenslanges Lernen schaffen. Lernen kann nun zu jeder Zeit und an jedem Ort stattfinden: auf dem Sofa Zuhause, im Bus während des Pendelns oder in der Pause auf der Arbeit. Die Flexibilität des mobilen Lernens geht jedoch zeitgleich mit Herausforderungen einher. Ohne den kontrollierten Ablaufplan und die Unterstützung der Lehrpersonen im schulischen Umfeld sind die Lernenden selbst dafür verantwortlich, aktiv Lernsitzungen zu initiieren, Wiederholungszyklen einzuhalten und Lektionen in unterbrechungsanfälligen Alltagssituationen zu meistern. Sowohl Unterbrechungen als auch unregelmäßige Wiederholung von Inhalten können den Lernprozess behindern und der Langzeitspeicherung der Informationen schaden. Wir behaupten, dass aktuelle mobile Lernanwendungen die Nutzer*innen nur unzureichend in diesen Herausforderungen unterstützen.

In dieser Arbeit erforschen wir, wie wir uns die Allgegenwärtigkeit mobiler Endgeräte zunutze machen können, um zu erreichen, dass Nutzer*innen regelmäßig mit den Lerninhalten interagieren. Wir fokussieren uns darauf, sie im Umgang mit Unterbrechungen und Lernpausen zu unterstützen. In einem nutzerzentrierten Designprozess analysieren wir zunächst das Lernverhalten auf mobilen Endgeräten in alltäglichen Situationen. Basierend auf den Erkenntnissen schlagen wir Konzepte und Designs vor, entwickeln Forschungsprototypen und werten diese in Labor- und Feldstudien mit Fokus auf *User Experience* (wörtl. “Nutzererfahrung”) aus.

Um das Lernverhalten von Nutzer*innen mit mobilen Endgeräten besser zu verstehen, versuchen wir zuerst die Interaktionen mit mobilen Lernanwendungen durch eine detaillierte Umfrage und eine Tagebuchstudie zu charakterisieren. Beide Methoden bestätigen eine enorme Vielfalt von Nutzungssituationen und -präferenzen. Wir beobachten, dass Lernen oft ungeplant, unregelmäßig, im Beisein von Freunden oder Familie, oder während der Ausübung anderer Tätigkeiten, beispielsweise Fernsehen oder Essen, stattfindet. Die Studien decken zudem Unterbrechungen in Alltagssituationen auf, welche das Lernverhalten der Nutzer*innen beeinflussen und oft zum Aussetzen oder Beenden der Lernaktivität führen. Wir leiten Implikationen ab, um Lernen in vielfältigen Situationen zu unterstützen und besonders die negativen Einflüsse von Multitasking und Unterbrechungen abzuschwächen. Die vorgeschlagenen Strategien sollen Designer*innen und Entwickler*innen helfen, mobile Lernanwendungen zu erstellen, welche sich den Möglichkeiten und Herausforderungen von Lernen in Alltagssituationen anpassen. Wir haben vier zentrale Herausforderungen identifiziert: (1) *Lernen in Alltagssituationen ist divers und anfällig für Unterbrechungen*; (2) *Die Lerneffizienz wird durch unregelmäßiges Wiederholungsverhalten beeinflusst*; (3) *Wir müssen von statischem zu personalisiertem Lernen übergehen*; (4) *Unterbrechungen und lange Lern-*

pausen können dem Lernen schaden. Um diese Herausforderungen anzugehen, schlagen wir vor, Lernen in alltägliche Smartphoneinteraktionen einzubetten. Dies führt zu einer vermehrten Beschäftigung mit Lerninhalten und könnte zu einer impliziten Personalisierung von diesen anhand der Interessen und Fähigkeiten der Nutzer*innen beitragen. Zudem untersuchen wir, wie *Memory Cues* (wörtl. "Gedächtnishinweise") genutzt werden können, um das Fortsetzen von Aufgaben nach Unterbrechungen im mobilen Lernen zu erleichtern.

Um zu zeigen, dass unsere Idee des Einbettens von Lernaufgaben in alltägliche Interaktionen wirklich die Beschäftigung mit diesen erhöht, haben wir eine Anwendung entwickelt, welche Lernaufgaben in den Entsperrprozess von Smartphones integriert. Da die Authentifizierung auf dem Mobilgerät eine häufig durchgeführte Aktion ist, welche keinen weiteren Mehrwert bietet, begrüßten unsere Studienteilnehmenden die Idee, den Prozess für die Durchführung kurzer und einfacher Lerninteraktionen zu nutzen. Ergebnisse aus einer vergleichenden Nutzerstudie haben gezeigt, dass die Einbettung von Aufgaben in den Entsperrprozess zu signifikant mehr Interaktionen mit den Lerninhalten führt, ohne dass die Lernqualität beeinträchtigt wird. Wir haben außerdem eine Methode untersucht, welche die Messung von Sprachverständnis in die digitalen Lese- und Höraktivitäten der Nutzer*innen einbettet. Mittels physiologischer Messungen als implizite Eingabe können wir in Laborstudien zuverlässig unbekannte Wörter erkennen. Die Aufdeckung solcher Wissenslücken kann genutzt werden, um in-situ Unterstützung bereitzustellen und um personalisierte Lerninhalte zu generieren, welche auf die Interessen und das Wissensniveau der Nutzer*innen zugeschnitten sind.

Um Memory Cues als Konzept für die Unterstützung der Aufgabenfortsetzung nach Unterbrechungen zu untersuchen, haben wir eine theoretische Literaturanalyse von bestehenden Anwendungen um zwei Forschungsarbeiten erweitert, welche vielversprechende Designkonzepte umsetzen und evaluieren. Wir haben gezeigt, dass die Präsentation von Memory Cues die subjektive User Experience verbessert, wenn der Nutzer die Lernaktivität nach einer Unterbrechung fortsetzt. Eine Folgestudie stellt einen Ausblick auf die Generalisierbarkeit von Memory Cues dar, welcher über den Tellerrand des Anwendungsfalls Sprachenlernen hinausschaut. Wir haben beobachtet, dass der Nutzen von Memory Cues für das Reflektieren über gelernte Inhalte stark von dem Design der Cues abhängt, insbesondere von der Granularität der präsentierten Informationen. Wir schätzen interaktive Cues zur spezifischen Gedächtnisaktivierung (z.B. durch Mehrfachauswahlfragen) als einen vielversprechenden Unterstützungsansatz ein, welcher individuelle Mikrolerneinheiten im Alltag verknüpfen könnte.

Die Werkzeuge und Anwendungen, die in dieser Arbeit beschrieben werden, sind ein Startpunkt für das Design von Anwendungen, welche das Lernen in Alltagssituationen unterstützen. Wir erweitern das Verständnis, welches wir von Lernverhalten im geschäftigen Alltagsleben haben und heben den Einfluss von Unterbrechungen in diesem hervor. Während sich diese Arbeit hauptsächlich auf das Lernen von Sprachen fokussiert, haben die vorgestellten Konzepte und Methoden das Potential auf andere Bereiche übertragen zu werden, beispielsweise das Lernen von MINT Themen. Wir

reflektieren über die Grenzen der präsentierten Konzepte und skizzieren Perspektiven für zukünftige Forschungsarbeiten, welche sich die Allgegenwärtigkeit von mobilen Endgeräten zur Gestaltung von Lernanwendungen für den Alltag zunutze machen.

Acknowledgments

During the time of my PhD, I was lucky to be surrounded by a great number of wonderful people who made this journey a great pleasure.

First and foremost I want to thank my supervisor **Heinrich** Hußmann - you not only gave me the freedom to find my own path in the academic field (and make my own mistakes) but have been a kind and understanding colleague throughout the way. Further, I want to thank the members of my committee - **Gerhard** Fischer and **Tilman** Dingler - for your much appreciated feedback, time, curiosity, and encouragement.

Thank you also to the current and former professors of the Media Informatics Group **Andreas** Butz, **Albrecht** Schmidt, **Sven** Mayer, and **Florian** Alt, who never hesitated to provide feedback and made this group such a wonderful place to work.

I had the honor to work with an amazing team of people who shaped this thesis and my life with their discussions and encouragement. A big thank you goes to **Sarah** Aragon Bartsch, for being my PhD sister and roomie, and for walking this path with me during the last four years as friends. Thanks to the LMU girls - to **Sarah** Völkel, my fellow non-bavarian, for your contagious dedication to work while simultaneously trying to give me a 'social' life beyond all the work; to **Nađa** Terzimehić, for being such an infectiously joyful person making even Monday mornings fun; to **Mariam** Hassib, for your big and kind heart; to **Linda** Hirsch, for coffee breaks, hiking trip, badminton matches, and bearing with me through the "dark" times of the literature analysis; to **Fiona** Draxler, my bookchapter companion, for sharing the struggle and always providing calmness and positivity - without the friendship of you girls the past four years would not have been half the fun they were.

Further thanks goes to **Romina** Poguntke, my homie and twin of humor, for sharing many many laughs; to **Pascal** Knierim, for the lovely and funny chats over coffee in the early morning when all other offices were still empty; to **Thomas** Kosch, for your unlimited enthusiasm about my ideas and your help to turn them into actual research; to **Matthias** Hoppe, for taking such good care of Pino and for all the chats and welcome distraction during stressful days; to **Daniel** Buschek, for providing constructive feedback on literally any topic and sharing your continuous optimism; to **Malin** Eiband, for supervising the first student thesis with me and being the positive person you are.

To every current and former MIMUC members and friends, thanks you Beat, Florian B., Ceenu, Kai, Jingyi, Amy, Sylvia, Matthias S., Thomas W., Dennis, David, Yomar, Carl, Changkun, Gesa, Michael B., Sarah P., Lukas, Ken, Michael C., Florian L., Heiko, Jakob, Sebastian, Florian L., Daniel U., Alexander, Tonja, Lewis, Robin, Jesse, Francesco, Steeven, Luke, Hanna, Renate, Axel, Tobias, Ville, Mo, Maria, Christian, Bastian, Yomna, Radiah, Yasmeen, Simon, Jonas, and Sarah F.

Further, I want to thank the external GermanHCI crew for all the amazing and fun events - **Donald**, **Thijs**, **Tom**, **Teresa**, **Johannes**, **Sebastian** and all the wonderful people who make up this wonderful community.

To the amazing people who keep the Media Informatics team running - **Franziska**, **Christa**, **Anja**, and **Doris**. A very special thanks goes to **Rainer**, for the always open door, be it for complains about technical problems or complains about everyday life. You will always have a special place in Pino's and my heart.

Thanks to all my excellent students for making it such great collaborations, especially Andrea, Felix, Jonas, Mariam, Marius, Miriam, Sophia, Teodora, Viktoriia, and Vincent, whose work found their way into this thesis.

To the ones closest to me, my dearest friends. For bearing with me being away from home and supporting me over all those years. Thank you **Leonie**, **Melanie**, **Juliane** & **Artur**, **Anika**, and **Carla** for always being there and for making life so much brighter - I would not be the person I am today if it wasn't for you.

Der größte Dank gilt jedoch meiner Familie, die mir stets mit voller Unterstützung beiseite stand. Insbesondere möchte ich **Ursel** Schneegaß danken, für den besten Kuchen und die besten Ratschläge; meinen Eltern **Petra** und **Gerd** Schneegaß, dafür, mir das Gefühl gegeben zu haben nie, wirklich weg gewesen zu sein. Thanks to my brother **Stefan**, for always lecturing me about academic work and life, for being my mentor, role model, biggest critic, and biggest supporter.

Last but definitely not least, my thanks goes to **Evan**, for your patience, your support, and for making me feel at home wherever we go.

Statement of Collaboration

The research presented in this thesis was carried out over the past four years, between February 2017 and August 2021, in the Media Informatics group at LMU Munich. Chapters 1, 2, 9, and 10 present original work exclusively written for this thesis. Chapters 3 to 8 are based on peer-reviewed publications.

The work presented in this thesis is based on collaborations with my supervisor Heinrich Hußmann, colleagues, and supervised bachelor and master students who influenced this thesis and the results presented in it. In my role as supervisor of the included bachelor and master theses, I determined the research scope and goal, provided feedback and made the final decision in all stages of the work (concept, prototype, implementation, study design, hypotheses, evaluation, and result analysis). I will use the scientific plural throughout this dissertation to acknowledge my collaborators' contributions. The following statement will outline the details of my contribution.

Chapter 3: Mobile Learning Session Characterization The studies presented in this chapter are based on a publication at MUM 2018 [305]. I developed the original research idea and iterated on it in discussions with Nađa Terzimehić. I designed the user study setup and led the supervision of the bachelor thesis student Mariam Nettah, who executed the study. I analysed the gathered data and led the follow-up focus group. I was leading author of the resulting publication, Stefan Schneegass provided feedback on it.

- Schneegass, C., Terzimehić, N., Nettah, M., and Schneegass, S. (2018). Informing the Design of User-adaptive Mobile Language Learning Applications. In Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia (MUM'18). ACM, New York, NY, USA, 233–238.

Chapter 4: Exploring Interruptions during Mobile Learning The research idea was developed in close discussion with Fiona Draxler and resulted in a paper accepted for publication at Mensch und Computer 2021 [96]. Jonas Safranek implemented the data logging application and performed the user study as part of his master thesis supervised equally by Fiona Draxler and myself. I pre-tested the implemented application at several stages and co-designed the study setup. I led the analysis of the occurring interruptions, while Fiona Draxler's focus was the quantification of mobile learning behaviour (only marginally included in this thesis). Fiona Draxler and I contributed equally to the resulting publication. Heinrich Hußmann and Jonas Safranek provided feedback on an earlier version of the publication.

- Draxler, F.*, Schneegass, C.*, Safranek, J., and Hußmann, H. (2021). Why did you stop?- Investigating Origins and Effects of Interruptions during Mobile Language Learning. Accepted for publication at Mensch und Computer (MuC'21),September 5–8, 2021, Ingolstadt, Germany. ACM, New York, NY, USA, 15 pages.

---

* Both authors contributed equally to this research

Chapter 5: Embedding Vocabulary Acquisition into Smartphone Authentication

This chapter is based on a paper accepted for publication at Mensch und Computer 2021 [304]. I developed the research idea and refined it in discussion with Daniel Buschek and Malin Eiband. Teodora Mitrevska designed and evaluated the prototypes (Sections 5.2 and 5.3) as part of her bachelor thesis. Sophia Sigethy implemented the three follow-up learning applications and conducted the comparative user study (Section 5.4) as part of her bachelor thesis and practical project. I was the primary supervisor of Sophia and Teodora, revised and extended the analysis of the data collected by both students. I authored the publication, Daniel Buschek and Malin Eiband provided feedback on individual parts of the manuscript.

- Schneegass, C., Sigethy, S., Eiband, M. and Buschek, D. (2021). Comparing Concepts for Embedding Second-Language Vocabulary Acquisition into Everyday Smartphone Interactions. Accepted for Publication in Mensch und Computer 2021 (MuC '21), September 5–8, 2021, Ingolstadt, Germany. ACM, New York, NY, USA, 15 pages.

This paper received an *Honourable Mention Award* at the Mensch und Computer conference and was invited to be published as extended version as

- Schneegass, C., Sigethy, S., Mitrevska, T., Eiband, M. and Buschek, D. (2021). UnlockLearning - Investigating the Integration of Vocabulary Learning Tasks into the Smartphone Authentication Process. In submission to the i-com Journal of Interactive Media

Chapter 6: Embedding Comprehension Assessment into Digital Reading and Listening

The chapter is based on two papers published at INTERACT 2018 [303] and CHI 2020 [302]. I came up with the original research idea and designed the experimental setup. I performed the first user study and the majority of the data analysis supported by Thomas Kosch. For the CHI 2020 publication, I was the main supervisor of the bachelor thesis students Andrea Baumann and Marius Rusu who implemented the

study setup. I pre-tested their implementation, provided feedback, and performed the majority of the data analysis, with some help of Thomas Kosch. I was leading author of both resulting publications. Albrecht Schmidt, Heinrich Hußmann, and Mariam Hassib provided feedback on individual parts of the CHI 2020 publications.

- Schneegass, C., Kosch, T., Schmidt, A., and Hußmann, H. (2019). Investigating the Potential of EEG for Implicit Detection of Unknown Words for Foreign Language Learning. In IFIP Conference on Human-Computer Interaction (INTERACT'19) (pp. 293-313). Springer, Cham.
- Schneegass, C., Kosch, T., Baumann, A., Rusu, M., Hassib, M., and Hußmann, H. (2020). BrainCoDe: Electroencephalography-based Comprehension Detection during Reading and Listening. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). ACM, New York, NY, USA, 1–13.

Chapter 7: Design Space for Task Resumption Cues

The literature analysis and design space presented in this chapter (cf. Section 7.2) were part of a book chapter published in 2021 [300] together with the results of two focus groups (Section 7.3). The focus group results were further published as work-in-progress at MobileHCI 2019 [95] as an excerpt of the book chapter. I developed the research idea in discussion with Fiona Draxler and we contributed equally to the literature analysis and focus group. I led the resulting book chapter publications.

- Schneegass, C. and Draxler, F. (2021). Designing Task Resumption Cues for Interruptions in Mobile Learning Scenarios. In Technology-Augmented Perception and Cognition (pp. 125-181), Dingler T. & Niforatos, E. (Eds). Springer, Cham.
    - *Adjunct Work-in-Progress Publication*:
      Draxler, F., Schneegass, C., and Niforatos, E. (2019). Designing for Task Resumption Support in Mobile Learning. In Adjunct Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'19) (pp. 1-6).

Chapter 8: Memory Cues in Mobile Learning Applications

The research idea was developed together with Fiona Draxler as result of Chapter 8. Vincent Füseschi and Viktoriia Konevych each implemented one learning application and performed the respective user study as part of their master and bachelor thesis.

I closely supervised both students together with Fiona Draxler. Fiona further contributed to parts of the analysis of the first study in the lab. I analyzed the data collected in the second study in the wild and wrote most parts of the resulting manuscript, which is published as follows [301].

- Schneegass, C., Füseschi, V., Konevych, V. and Draxler, F.,. Investigating the Use of Task Resumption Cues to Support Learning in Interruption-Prone Environments. In Multimodal Technol. Interact. 2022, 6, 2.

Chapter 9: Outlook - Memory Cues Beyond Language Learning

I came up with the original research idea and concept for implementation in discussion with Fiona Draxler. I was the primary supervisor of the bachelor thesis student Miriam Halsner and co-designed the learning application and the user study with her. I analyzed the data gathered in the user study. The content of this chapter is not published but builds up upon the above-mentioned research project performed in collaboration with Fiona Draxler and Miriam Halsner.

Table of Contents

I

Introduction and Foundations

1
Introduction

Learning is a lifelong task that exceeds the boundaries of a school's classroom. Even after we finish our formal institutional education, we continue to expand our knowledge with new information every day and learn throughout our lifetime. One example of a skill that is improved on throughout peoples' lives is speaking a second or additional language. Sometimes we start learning it at school and can acquire enough knowledge to engage in fluent conversations. However, it is a lifelong task in the sense that the amount of vocabulary, expressions, or dialects, is endless and is improved on throughout a person's life. Some of this knowledge can be acquired unconsciously by observation or imitation in our daily lives [182]. Nevertheless, learning the basic rules and grammar concepts requires conscious investment of effort and persistence over a certain amount of time [119].

Around one-third of Europe's 25 to 64-year-old population are bilingual (35.2%), 21 percent trilingual, and 8.4 percent speak four or more languages[1]. Speaking multiple languages can not only help us be attractive on the job market, ease a stay abroad, or improve our basic reading and communication skills [104, 211], but it is also associated with many health benefits. Prior studies have shown that being proficient in two or more languages can be beneficial in all stages of life. Young children show better cognitive performance when being brought up bilingually than monolingual [31], and older people show delayed signs of cognitive decline and dementia onset [32, 188]. Therefore, learning a second language at any point during life is a valuable investment.

As our relationship with technology is continuously changing, we are required to rethink how we design lifelong learning practices [109] such as language learning. With smartphones gaining increased sensory and computing capabilities, they have replaced personal computers for many everyday tasks. Mobile devices have not only become ubiquitous in our society but have also become more embedded in our everyday life. On average, we spend multiple hours a day interacting with our smartphones, unlocking them more than 25 times daily [128, 142, 223]. We use them to communicate with friends, read books and newspapers, and listen to audiobooks or podcasts. Naturally, the ubiquity of mobile device usage also leads to mobile language learning (MLL) becoming more embedded into our everyday lives. Especially for teaching small learning chunks such as vocabulary, the opportunity of frequent engagement and high repetition count throughout the day has shown significant benefits for learning [74, 86].

As Mobile Language Learning (MLL) apps target a broad spectrum of users, these apps as of yet insufficiently support individual learning preferences. For example, while

[1] Eurostat Statistic Explained: `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Foreign_language_skills_statistics`, last accessed January 3, 2022

self-driven learners are likely to blossom in the flexibility these ubiquitous learning approaches offer, other learners struggle. In particular, keeping up the perseverance of engaging with learning content frequently can be difficult in informal learning environments without a fixed curriculum. Thus, going beyond the one-size-fits-all approach and advancing personalized learning using novel technology has been declared one of the 'Grand Challenges for Engineering' by the National Academy of Engineering[2] and is part of the 'Seven HCI Grand Challenges' derived by Stephanidis et al. [319].

Due to the continuously changing role of technology, this thesis starts by exploring how people use MLL applications in their everyday lives. We gather and analyze peoples' subjective characterizations of common usage situations and complement our insights with log and experience sampling data of MLL app usage in the wild. In particular, learning on the go makes users susceptible to interruptions [209], which can have severe adverse effects on memory encoding and learning [22, 57, 352]. Therefore, we aim to better understand the prevalence of interruptions and their effects to provide personalized support when learning in informal everyday settings.

Since frequent repetition of content is essential for long-term retention [168, 294], we explore two approaches to engage users more frequently in learning exercises. Firstly, by embedding vocabulary tasks into frequently occurring interactions with our mobile devices, such as the authentication process [128, 142, 223], we aim to foster frequent and continuous repetition of the content in an unobtrusive manner. Secondly, we investigate how we can utilize users' interaction with foreign-language media content to create personalized learning content. Learning with media content has been shown to support effective learning [134, 360], and when the learning objectives are adapted to individual interests, it further improves users' motivation for interacting with it [263]. Beyond personalization to subjective interests, novel sensing technologies provide the opportunity to assess the learners' language proficiency (e.g., [18, 170]). This thesis investigates Electroencephalography (EEG) as an implicit measure of word-based comprehension for the generation of language learning content personalized to the users' individual knowledge gaps. By embedding learning implicitly into these everyday interactions, we aim to seamlessly support users' second-language acquisition.

As we established the prevalence of interruptions and infrequent learning in everyday settings, we further investigate strategies to help bridge the gap between individual learning sessions. We aim to counteract everyday life interruptions, whether they are short distractions or longer breaks between sessions, by providing *Task Resumption Support*. Since interruptions can rarely be anticipated, we investigate the application of memory cues to help guide users back to their learning context after an interruption or break occurred. We perform a structured in-depth literature analysis on task resumption support in other domains and derive design opportunities and requirements. We implement memory cues in an MLL app and show their positive effect on users'

[2] NAE Grand Challenges for Engineering: `http://www.engineeringchallenges.org/challenges/learning.aspx`, last accessed January 3, 2022

learning experience in two user studies (lab-based and in-the-wild). We outline design implications for implementing task resumption memory cues in MLL apps that mitigate the adverse effect of task switches between learning and interruptions.

1.1 Vision: Embedded Mobile Language Learning

In an era where technology takes over more and more of our inherent skills such as mental calculus, navigation, or memorizing, it becomes increasingly important to view technology as a tool to augment human abilities and not replace them. In 1995, Douglas Adams proposed in his book "The Hitchhiker's Guide to the Galaxy" [2] the idea of the *Babel fish*. This small yellow fish can be placed in anyone's ear and feeds off unconscious brainwave energy. Carrying a Babel fish enables a person to understand every language spoken around them as if it were their native language. Increasingly smart, cheap, and miniaturized technology will soon have the means of realizing this vision, as it coalesces with the users themselves (such as through smart watches, glasses, or garments).

While the vision of instant translation at first seems to render speaking a second-language for communication purposes unnecessary, it is only a compromise and not a solution. To begin with, the technology behind automated translation is still years of work from functioning seamlessly. As of yet, tools such as Google Translate are still subject to severe inaccuracies[3] and biases[4]. However, even if the technology would provide significantly better translations, the important distinction lies in the use of tools vs. the development of innate human skills. In this introduction, we outlined the manifold benefits of speaking a second language on our cognitive abilities and physical health. Language and communication are far more complex than selecting and grouping words into sentences. Language transfers emotions; it is personal and creates and affects our relationships with other people. It is constantly evolving with all its varieties and can visualize subtle differences. Speaking multiple languages, therefore, remains a quality that cannot be easily replaced by technology.

Therefore, this thesis aims to assist the user in acquiring an inherent skill instead of taking over the task. Rather than translating foreign-language content, we envision technology to support us in learning and remembering the language we aim to acquire.

To make this skill acquisition easier, we argue that the ubiquity of mobile devices and foreign-language content in our everyday lives presents an opportunity for embedded

[3] ABC News "Government coronavirus messages left 'nonsensical' after being translated into other languages" by Stephanie Dalzell (2020) – https://www.abc.net.au/news/2020-11-19/government-used-google-translate-for-nonsensical-covid-19-tweet/12897200, last accessed January 3, 2022

[4] Algorithwatch "Female historians and male nurses do not exist, Google Translate tells its European users" by Nicolas Kayser-Brill (2019) – https://algorithmwatch.org/en/google-translate-gender-bias/, last accessed January 3, 2022

language learning support. "Embedding" refers to the seamless integration into everyday settings to create an unobtrusive interaction that does not disrupt or distract from the primary task or goal (cf. [181]). This form of embedded interaction refers to multiple stages of the language learning process, such as *knowledge assessment*, *knowledge acquisition*, and *knowledge repetition and retention*.

One way we envision for embedding learning interactions into everyday settings is through smart glasses. When such glasses include a front-facing camera and microphone, they can detect encounters the user has with a target language throughout the day, either written or spoken. By processing the language content (i.e., optical character recognition, natural language processing) and comparing the content to the users' knowledge base (i.e., proficiency-aware systems), we can provide in-situ support and implicitly assess the users' knowledge. For example, a personalized vocabulary learning list could be generated from unknown words during a foreign-language movie a person watches or advertisements on billboards a person sees, tailored to interests and proficiency.

Knowing what to teach a learner is only the beginning of the actual learning process. Especially when the content unit's complexity exceeds mere vocabulary acquisition, learners' must invest time and consciously engage with the learning tasks. We envisage learning applications to support users in stopping and resuming such lessons at any time to fit them more seamlessly into their daily lives. An interruption occurred in the middle of the task? No problem. Whenever the users find their way back to the learning application, the app reminds them of what they were doing before to help resume the task as if no time has passed.

To make sure the learner does not lose track of the learning application, we further envision the tasks to be seamlessly embedded into the users' lives. Learning a language usually works best when surrounded by it, i.e., living in a country where it is the native language. Mobile technology has the opportunity to create the impression of a language being ubiquitous in the users' lives. By presenting foreign content in non-critical situations (e.g., while brushing the teeth in the morning or waiting for the water to boil in the kettle), we can significantly increase learners' exposure to the target language.

1.2 Research Challenges and Contributions

This thesis aims to contribute to the overall goal of improving language learning with mobile devices in everyday settings. As one of the Seven HCI Grand Challenges recognizes, "new technologies have the potential to support new and emerging learning styles, as they have recently evolved and been influenced by the pervasiveness of technology in the everyday life of the new generations" (Stephanidis et al. [319], p.1232). The following scenario paints a picture of an everyday use case when learning with a

mobile learning app ("NormalLearningApp"), which is similar to the MLL apps currently available on the market:

SCENARIO

Anna is a curious 20-year-old girl who just came home after booking a summer vacation to Portugal with her friend Lea. Anna is excited about the trip eight weeks from now, especially about engaging with the local food culture. She wants to be able to communicate with locals, and learn about their produce and new recipes in a fluent conversation without having to use translation apps. In preparation, Anna decides the download the MLL application *NormalLearningApp* to her smartphone. She wants to learn basic Portuguese for communication that she can improve on once she arrives in Portugal.

She sits down on the couch and immediately starts the app at a beginner's level - '*Mulher*' - the woman. *NormalLearningApp* congratulates Anna on her first correct translation. Anna is excited and hopes she will learn Portuguese like this very quickly. She spends another two hours answering multiple-choice questions on basic phrases, animals, and clothing items.

The days go by and Anna is unusually busy with work so she interrupts her practice for several days. She picks up her phone in her lunch break and continues the *NormalLearningApp* lesson she left off the other night. Translate the word '*Cão*'. Maybe '*Shirt*'? Anna wonders. Incorrect, try again. '*Pants*'? Also no. '*Dog*' would have been the correct solution. "Oh right", Anna remembers that her last lesson was not about clothes but about animals. As she already has problems concentrating she decides to not start a new lesson but rather repeat the prior lectures to consolidate her knowledge. She is disappointed with herself and feels like she has already forgotten half of what she learned and is not making progress. "I'm gonna practice more frequently from now on", she promises to herself.

As the next day is a Saturday, Anna repeats the animal lesson again for 15 minutes while eating her breakfast at the dining table. She gets bored by the monotonous questions and starts to watch a cooking show with Portuguese subtitles on Netflix instead. '*Estamos preparando pastel de nata polvilhado com canela.*' [We are preparing egg custard tarts dusted with cinnamon.] says the host. Amazed by the delicious looking tarts, Anna dreams about how she will order them once she arrives in Porto. She wonders why *NormalLearningApp* is not teaching her the vocabulary for conversations about food but is showing her the words for animals she is not interested in.

> The weeks go by and Anna learns with *NormalLearningApp* one or two times a week. Finally, Anna arrives in Portugal, sitting in a small café on Porto's Rua de São João Novo - "*Uma café e one of the custard tarts, por favor*". While Anna can formulate some basic sentences and order a coffee, she does not feel well prepared for the vacation in regards to her personal interests, food and cooking, and is disappointed that she hadn't practiced more frequently.

The story of Anna in particular illustrates four challenges central in today's mobile learning applications, which we will now outline in more detail. Drawing on a diverse background of established theories and literature from cognitive psychology to the learning sciences, this thesis contributes to all four challenges. We apply methods of User-centered Design (UCD) and Human-Computer Interaction (HCI) in general to contribute on an empirical, survey, and artifact level [348].

Challenge 1: Learning in Everyday Settings is Diverse and Interruption-prone

The last two decades' technological progress enables us to learn anytime and anywhere with the help of mobile devices. Yet, we still know very little about how learning applications are actually used on a daily basis. The ability to learn wherever we go comes with sheer unlimited flexibility but also with risks inherent to uncontrolled mobile environments, such as multitasking (e.g., walking or watching TV) and interruptions (e.g., notifications or conversations) [145, 209, 256]. When taking learning outside the classroom, the everyday environment and also our learning device can distract and interrupt us, which negatively impacts the learning performance [22, 171, 186]. Getting a better grasp of when and how users engage in learning activities and understanding the role of interruptions is a core challenge of learning in everyday settings. We translate this challenge into two research questions:

RQ1a: How do people use mobile learning applications in everyday settings?

RQ1b: How do interruptions affect mobile learning in everyday settings?

CONTRIBUTION 1

Characterizing Mobile Learning and Interruptions in Everyday Settings - *Expanding our knowledge of people's mobile learning application usage behavior and investigating the prevalence and effects of interruptions in everyday environments.*

Based on a detailed online survey, we report on peoples' most common usage situations for mobile learning applications. Besides contextual factors such as environment and company, we inquire about the learners' habits of

planning learning activities, frequency of learning, and the occurrence of multitasking. We derive five clusters from users' described situations using a grounded theory approach and outline challenges of learning in everyday settings. The survey results highlight the enormous diversity in how people use MLL applications and reinforce the expectation that mobile learning takes place anywhere and anytime in peoples' everyday lives. We further deployed an application collecting in-situ reports of contextual information about learning situations using the Experience Sampling Method (ESM) and unveil a prevalence of interruptions during mobile learning in everyday settings. The ESM reports show that interruptions are ubiquitous (276 interruptions in 327 learning sessions), and while 20 percent of those interruptions require users' immediate attention, the remaining 80 percent were labeled as only moderately urgent or not urgent at all. We contribute to a better understanding of interruptions and users' tendency to suspend mobile learning activities for time-critical distractions. We discuss implications for designing technological interventions.

Challenge 2: Irregular and Infrequent Practice Behaviour

Micro-learning applications for vocabulary-based language learning have already proven to increase vocabulary recall [53, 100, 331]. Yet, the flexibility of learning with mobile devices and the absence of curricula or supervision require constant initiative from the learner to engage with the learning content. However, adhering to a regular repetition schedule and frequent exposure to content is crucial for consolidating knowledge [19, 294, 322]. Conversely, irregular learning can lead to insufficient consolidation of previously learned content [294], requiring learners to acquire knowledge multiple times to reach a state of effortless recall, just like Anna from the example outlined before who had to repeat previously learned content after a long break between two learning sessions. The challenge is identifying opportunities for systems to nudge learners to frequently engage with second-language content while remaining unobtrusive and keeping up a positive learning experience. One possibility is to exploit frequent smartphone interactions as a window of opportunity to present language content. A frequent interaction that in itself does not contribute to a certain goal is mobile authentication, which we perform on average more than 25 times a day [128, 142, 223]. Presenting tasks such as nutrition tracking on the lockscreen [167] or even connecting actions immediately to the unlock action such as journaling [362] has been positively evaluated. Thus, we see great potential for transferring this approach to learning. We pose the following research question:

RQ2: How can the integration of learning tasks into the smartphone authentication process foster frequent engagement?

CONTRIBUTION 2

Exploring Opportunities for Embedded Mobile Language Learning - *Embedding language learning tasks into everyday actions to foster frequent engagement with content relevant to the user.*

This contribution is synthesized from the results of two research projects: (1) We applied a UCD process to explore designs to embed learning tasks into the smartphone authentication process to foster frequent engagement. A usability evaluation with different designs for commonly available authentication mechanisms revealed that users consider the embedded learning idea attractive when kept simple. To investigate users' reactions when using such an application on a daily basis, we performed a three-week comparative evaluation of three mobile learning applications representing different levels of embeddedness. (2) A complementary within-subject field study contrasted (a) a common self-initiated stand-alone app with (b) an app presenting a task in a continuously visible notification, and (c) a novel app-initiated learning approach that embedded learning in the smartphone authentication process. We evaluated subjective and objective metrics revealing the difficulty of balancing frequent content exposure and disturbance. While participants favored the common stand-alone app for long learning streaks and complex learning content such as grammar knowledge, our prototype has the potential to increase content exposure significantly. We discuss its application beyond the scope of language learning.

Challenge 3: From Static to Personalized Learning

Current MLL applications group their learning content into small lessons according to certain goals. Some teach tenses or declination; others extend the vocabulary knowledge around certain domains [135]. Especially in the early stages of learning a language, the users do not require a broad vocabulary knowledge (e.g., concerning translations of animals or vehicles) but could specialize on topics of interest or relevance (e.g., clothing, food, nuclear physics). Personalization according to goals and interests are a central requirement for the adoption of lifelong learning technologies [108]. Current market applications allow only limited personalization as they would have to include content on all kinds of topics. With their limited scope, those apps risk endangering users' motivation in interacting with them. Meanwhile, users' consumption of foreign media content such as TV shows, news articles, or podcasts steadily increases. Since users select from these according to their interests, using media creates the opportunity to generate personalized language learning materials from external content. Looking back at our example of Anna, she expressed her desire to learn about a certain interest, i.e., local food. Therefore, the vocabulary used in her favorite cooking show fits her interest better than the standardized learning content of the learning app she uses.

Beyond the personalization according to interests, analyzing users' interaction with digital content can reveal insights into their comprehension of the presented learning content. Implicit evaluations through physiological sensors have already been used to assess general language proficiency levels during reading [28, 170]. Yet, we see even greater potential in this method to personalize content regarding users' interests and knowledge. More specifically, we aim to answer the following research question:

RQ3: How can we utilize users' everyday reading or listening activities to generate personalized language learning content?

CONTRIBUTION 3

Implicit Personalization of Learning Content According to Interest and Comprehension Levels - *Using EEG as a method for continuous and implicit assessment of second-language reading and listening comprehension is the first step towards learning personalization along with users' interest and proficiency.*

In two laboratory experiments, we developed a method that utilizes comprehension problems during second-language reading and listening as input to generate personalized learning content. We show that through EEG, specifically Event-related Potentials, we can implicitly and reliably detect unknown words during digital reading. In a first user study, we investigate text presented in an Rapid Serial Visualization Presentation (RSVP) format (one word at a time), while a second study presents the text in a full sentence format. Further, we expand this approach to an auditory presentation of text, i.e., narrations. With the detection of word-based vocabulary incomprehension, we will be able to support the user in-situ via the presentation of translations or generate personalized learning content focusing on users' knowledge gaps. The results are a first step toward turning everyday activities such as watching a foreign movie with subtitles into sources for personalized and embedded language learning experiences. We outline the current limitations of the technology and our vision of how EEG could soon be deployed in the wild for implicit comprehension assessment.

Challenge 4: Interruptions and Long Learning Breaks can Negatively Affect Performance

Learning in everyday settings requires the learning application to adapt to the sometimes busy schedule of users and their individual learning routines (cf. [85, 317]). Some might prefer to have ten short learning sessions a day, while others spend one intense focused afternoon every two weeks. The break between two learning sessions can be

fairly short, with only a few minutes between them (what we would call an interruption as it is a temporary shift of attention toward a secondary task), or days up to weeks might elapse (learning break). In our example, Anna spends two hours straight on learning Portuguese, but later only repeats content in shorter sessions. Yet, no matter the time between learning sessions, the memory of priorly acquired content decays eventually if not rehearsed [294]. As current mobile learning applications insufficiently accommodate for breaks and interruptions, the third key challenge is the design of adequate support for irregular schedules and coping with learning in interruption-prone environments. This challenge translates to the following research question:

RQ4: How can we use memory cues to mitigate the negative effects of interruptions in everyday mobile language learning?

CONTRIBUTION 4

4. **Mitigating the Effects of Interruptions and Infrequent Practice** - Connecting individual micro-learning sessions through the implementation of task resumption support can guide users back to the learning task after short interruptions and longer learning breaks.

   A thorough literature analysis shows the broad application spectrum of task resumption support in various domains. By sorting thirty concepts and applications focusing on memory cues into a multi-dimensional design space, we point out well-evaluated designs and uncover research gaps. Drawing on this literature basis and two focus groups (with learners and HCI experts), we examine promising memory cue designs. Moreover, we derive guidelines on integrating cues to facilitate task resumption after interruptions in mobile learning applications. Two subsequent studies (one laboratory experiment and one field study) show that although the effects of the cues on learning performance are negligible, participants find the cues helpful and supportive, especially those that summarize previously learned content or interactively pose questions on the last lesson. A final study provides an outlook on how task resumption cues could be applied beyond language learning. We evaluate users' experience with different cues for various content complexity levels in teaching programming (see Chapter 9). As hypothesized, users prefer a linear complexity mapping - simple content should be supplemented with simple cues: for complex content, complex cue design could be helpful.

This thesis proposes concepts and methods that help embed learning technologies more seamlessly into users' everyday lives. We analyze users' learning behavior, derive requirements, build prototypes, and evaluate applications. The findings are relevant for

several different disciplines, with some excerpts of this thesis being already published in these communities. The disciplines include, but are not limited to:

- **Human-Computer Interaction (HCI)** – as the main focus of this thesis – researches the intersection between the user and technology, particularly their interaction. Chapter 4 (publication [96]) illustrates how interruptions affect users' interaction with mobile learning applications in informal everyday settings. We emphasize the need for implementing mechanisms to mitigate the negative effects of interruptions on learning. Further, Chapter 5 (publication [304]) demonstrates an approach integrating learning tasks into the smartphone authentication process. We argue that the embedding of learning requires finding a balance between frequent interaction opportunities and unobtrusive design.
- **Mobile Computing** – interactions with technology have fundamentally different requirements when they take place in a mobile context. Chapter 3 (publication [305]) analyzes how users interact with mobile learning applications and derives clusters of common usage situations. Specifically, we take an in-depth look at how interruptions affect learning on smartphones (Chapter 4 (publication [94]) and how task resumption support can be tailored to the specific opportunities and limitations of mobile devices (Chapter 7, (publication [300]).
- **Technology-Enhanced Learning (TEL)** – refers to the use of technology as a medium for learning. We evaluate how users learn with mobile devices (cf. Chapter 3, publication [305]) and propose a method to use technology for language proficiency assessment (cf. Chapter 6, publications [302, 303]), and as a medium for the provision of learning content (cf. Chapter 5).
- **Technology-Augmented Cognition** – concerns the goal of using technology to enhance the users' inherent abilities and skills. The field combines insights from Cognitive Psychology on the basic human cognitive processes and investigates how technology can be used to support or enhance them. Chapters 7, 8, and 9 (publication [300]) propose implementing memory cues to augment users' memory. We aim to facilitate the recall of prior lessons to provide the user with context, mitigating the negative effects of interruptions and supporting task resumption after learning breaks.

1.3 Summary of Research Questions

In summary, this thesis targets four core challenges through five research questions (see Table 1.1). **RQ1a** and **RQ1b** inform Part II of this thesis in which we aim to better **Understand User Behavior**. In Part III, we investigate means for **Embedding Mobile Language Learning** with **RQ2** and **RQ3** addressing two possible alternatives.

RQ4 concerns details on how we can realize the **Connection of Micro-Learning Sessions**.

Contributing to solving the first challenge, we first take an in-depth look into how people use mobile learning applications. We identify problems and opportunities in mobile learning in everyday settings and derive patterns of common usage situations (RQ1a). In particular, we investigate the prevalence of interruptions and their frequency in mobile learning, and propose strategies to mitigate their negative effects (RQ1b).

Addressing the second challenge of irregular engagement with learning applications (i.e., learners not adhering to frequent repetition schedules), we explore two techniques of embedding language learning more seamlessly into people's everyday lives. On the one hand, we generate designs to integrate learning tasks into the smartphone authentication process, and perform a comparative evaluation with three prototypes varying in their level of required user initiative (RQ2). Secondly, RQ3 researches the generation of personalized language learning content from incomprehension detected during second-language reading and listening.

In Part IV, we target the third core challenge by examining the use of memory (task resumption) cues to bridge gaps between individual micro-learning sessions caused by interruptions or learning breaks (RQ4). After an extensive literature analysis and the creation of a design space, two user studies (one in the lab and one in the wild) investigate designs for task resumption cues in MLL apps. A follow-up study applying task resumption cues in an application teaching a programming language takes the first step toward generalizing the memory cue approach to more complex learning content.

Table 1.1: Overview of the research questions of this thesis.

RQ	Research Question	Chapter
	Part II: Understanding User Behavior	
RQ1a	How do people use mobile learning applications in everyday settings?	3
RQ1b	How do interruptions affect mobile learning in everyday settings?	4
	Part III: Embedding Mobile Language Learning	
RQ2	How can the integration of learning tasks into the smartphone authentication process foster frequent engagement?	5
RQ3	How can we utilize users' everyday reading or listening activities to generate personalized language learning content?	6
	Part IV: Connecting Micro-Learning Sessions	
RQ4	How can we use memory cues to mitigate the negative effects of interruptions in everyday mobile language learning?	7, 8, 9

1.4 Methodology

In the past thirty years, the prevalence of technology in education has risen significantly. A survey from Cambridge International Global Education Census in 2018[5] based on teachers and students from numerous countries found that 48% of students use desktop computers and 42% use smartphones in the classroom. Similarly to an increasing base of research on technology-enhanced learning, there is also extensive knowledge on the teaching of new languages. However, taking language learning activities outside the traditional classroom and curriculum by using technology is an area of research that changes just as quickly as our technology changes. What does not change is that the learner is the center of the learning activity. Therefore, this thesis follows a user-centered design process in designing and evaluating systems and applications to gain insights into users' experiences and attitudes.

1.4.1 User-Centered Design - Process and Methods

User-Centered Design User-Centered Design (UCD)[6] entails processes and methods of creating designs influenced by end-users [1]. The user can be involved on many levels. We can observe their behavior and interactions in a passive state; or actively, users can state their needs or experiences with a system or report on the usability of such. As the center of the process, the users' behavior and attitudes are the keys to informing the system design. As an iterative process, all phases of the UCD involving creating designs, implementation, and testing, can reveal insights that feed back into the original design idea [1, 226]. For the design and evaluation of the systems and applications described in this thesis, we applied and complemented several quantitative and qualitative research methods [206]:

We performed several **Surveys** as part of an empirical evaluation (Chapters 3, 4, 5, 8, 9). As surveys rely on self-reported data, they provide a subjective perspective on users' experience interacting with a system or application [206]. Therefore, the data gathered unobtrusively through surveys can be complementary to objective data such as log files.

To extract a broad range of perspectives and foster creative thinking through discussion, we conducted **Focus Groups** [206]. This method offers direct conversations with participants and can thus provide data a survey might disregard. In Chapter 7, we applied the Lotus-Flower Method [236], a structured idea-generation technique, combined with open discussions among participants to gain insights into different viewpoints on interruptions during mobile learning and favored task resumption strategies.

[5] Educational Census Survey: https://www.cambridgeinternational.org/Images/514611-global-education-census-survey-report.pdf, last accessed January 3, 2022

[6] ISO-9241-210. ISO 9241-210:2010(en) – Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems, 2010.

The **Experience Sampling Method** (ESM, also called "Diary" study) deploys short in-situ surveys with predefined questions and can have various triggers, e.g., a certain time, event, perception, or emotion [206]. Chapter 4 collects ESM data to gather subjective and timely insights of interruptions during mobile learning in a natural setting. This method prompts users with a short questionnaire after each learning session.

We further applied **Usability Testing** such as high-fidelity prototypes in Chapter 5 to gather insights on alternative user interfaces. This form of testing is often performed in close collaboration with the users to evaluate their interaction with the system or prototype [206].

In the last decade, technology collecting **Physiological Data** (signals and responses from the human body) has become increasingly cheap, robust, and reliable. For example, EEG, as utilized in Chapter 6 for the implicit detection of comprehension problems during reading and listening, has shown promising results when taken outside of controlled laboratory environments and embedded in wearables such as caps [34, 83] or glasses [343].

The majority of the performed evaluations can be categorized as **experimental research**. We postulate one or more hypotheses, define independent and dependent variables and test their relationship by applying significance testing methods [206], and design experimental protocols.

1.4.2 Prototypes and Settings

During the process, we generated prototypes of varying fidelity levels. While some remained conceptual to quickly assess participants' first impression on multiple ideas (i.e., paper sketches), others were developed into high-fidelity interactive prototypes to approximate normal mobile device interaction and gain elaborate user feedback.

We tested some of our prototypes and methods in laboratory conditions to create a controlled setting without external influences and disturbances. Especially for the collection of physiological data, the controlled environment was to ensure that the prototypes would be tested in a comparable setting among multiple participants and thus increase the validity of the collected data. This is particularly relevant for the evaluation of physiological measurements such as EEG, where measurements are easily influenced by different noise and lighting conditions.

Other prototypes were evaluated in field studies in the wild, such as the Android applications we deployed as internal test versions that users could download through the Google Play Store after invitation. These field studies allowed us to collect data from users in their everyday lives. During our evaluations, we applied a wide range of methods collecting subjective feedback (e.g., through user ratings), qualitative data (e.g., through interviews), and quantitative data (e.g., , through data logging). The data

we collected was analyzed using validated methodological and statistical approaches described in literature.

1.5 Ethics

In this thesis, we present research focusing on users' perspectives on technology in their everyday lives. Thus, we employ human-centered evaluation methods, including experimental work together with human subjects. In these studies, we ask for peoples' opinions, collect data on natural user behavior (e.g., their interaction with mobile devices), and quantify cognitive processes such as second-language comprehension.

In all evaluations, we follow the research ethics guidelines stated in the Declaration of Helsinki[7] to physically and mentally protect our participants. We provided transparent explanations on our study goals and procedures and offered extensive information about the study process. Participants provided informed written consent in all cases while having the right to withdraw from the experiment at any point without having to reveal their reason for discontinuation.

We followed the General Data Protection Regulations (GDPR) regulations and did not collect any data before consent was acquired. All data was anonymized and personal information or data identifying the user (such as demographics or contact information) was stored separately from study data on secure university internal servers.

The research described in this thesis was conducted with healthy participants and aimed at assessing their interaction with mobile learning technology and supporting their embedding into everyday settings. We applied non-invasive physiological sensing (e.g., through Electroencephalography (EEG)) to investigate underlying cognitive processes without self-reporting bias.

None of the studies conducted in this work violated any of the eleven criteria of the fast track ethical approval form of the ethics commission of the faculty for Math, Computer Science, and Statistics at LMU Munich[8]. Due to the potentially perceived intimate nature of the collected log data in Chapter 4 (i.e., log of all application names used during the study period without logging any application content itself), we further acquired official ethical approval by our University's ethics commission in this case (see Footnote in Chapter 4 for details).

[7] WMA Declaration of Helsinki - Ethical Principles: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/, last accessed January 3, 2022

[8] The fast track questionnaire can be found here: https://www.um.informatik.uni-muenchen.de/ethikkommission/fast-track-pdf.pdf, last accessed January 3, 2022

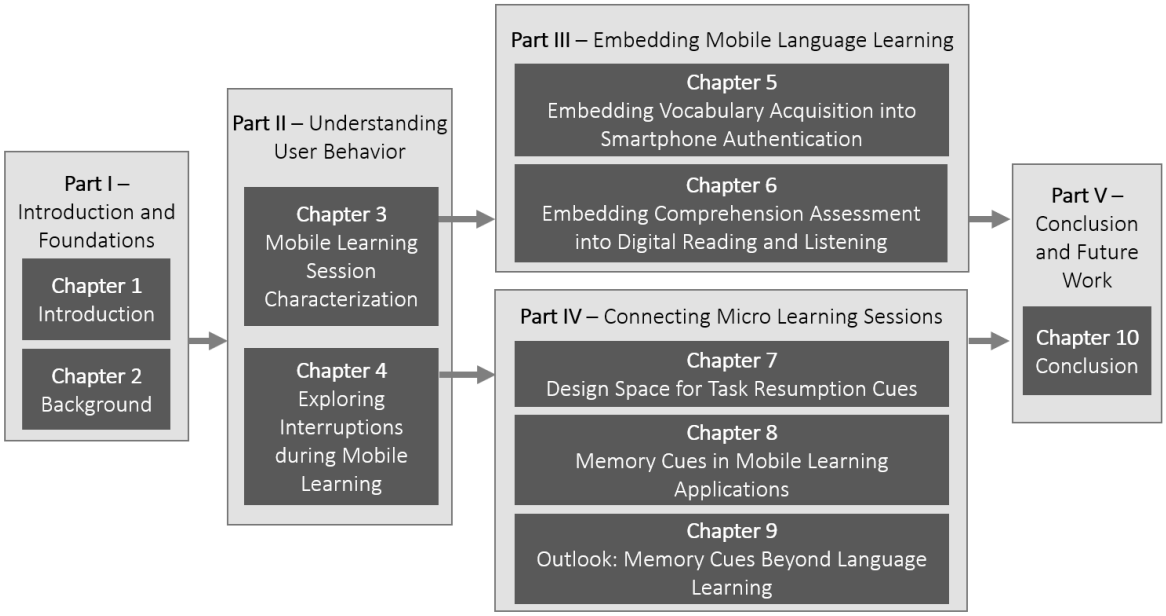


Figure 1.1: Thesis outline

1.6 Summary and Thesis Outline

This thesis comprises ten chapters organized in five parts, as depicted in Figure 1.1.

Part I motivates the work performed in this thesis and lays the theoretical and methodological foundations. In **Part II**, we investigate how people currently use MLL apps in their daily lives and compile challenges that frequently occur in/through the observed user behavior. Parts III and IV cover two of those challenges in detail: irregular use of mobile learning apps and perseverance over time. **Part III** presents a research probe to explore how to increase the frequency of engagement with learning content in everyday settings, and **Part IV** aims to create perseverance over time by bridging interruptions and connecting individual micro-learning sessions using a task resumption approach. In **Part V**, we present a synthesis of the gathered results, a conclusion, and opportunities for future work.

Part I - Introduction and Foundations

Chapter 1 - Introduction: This first chapter presents the motivation, problem description, and vision of how mobile learning can be embedded more seamlessly into everyday settings. It further outlines the research questions that guided this work and summarizes the contributions of this research. Last, this chapter includes a brief outline of this thesis.

Chapter 2 - Background and Foundations: This chapter lays the theoretical foundation for this thesis and introduces related work on everyday smartphone

interaction and (mobile) learning. It presents an overview of research performed in the areas of learning in everyday environments.

Part II - Understanding User Behavior

Chapter 3 - Mobile Learning Session Characterization: Assessing common usage situations for mobile learning application is a first step toward a better understanding of everyday usage of such tools. This chapter illustrates the results of a user study aimed to gather common learning situations, visualizes patterns and opportunities, and outlines potential challenges of learning in everyday contexts.

Chapter 4 - Exploring Interruptions during Mobile Learning: Building on the data collected in Chapter 3, this chapter takes a deeper look into everyday mobile learning app usage with a focus on task switching situations. We describe a user study to detect and classify moments where learning on a mobile device is interrupted and derive four strategies for mitigating negative effects caused by such interruptions – (1) *Ignoring*, (2) *Postponing*, and (3) *Preparing* for interruptions, as well as the provision of (4) *Task Resumption Support*.

Part III - Embedding Mobile Language Learning

Chapter 5 - Embedding Vocabulary Acquisition into Smartphone Authentication: In this chapter, we explore concepts for embedding MLL (vocabulary acquisition) into everyday smartphone interactions. To increase the repetition frequency of the content, we design concepts for integrating short learning tasks into the process of smartphone authentication as one example for a frequent interaction opportunity. A field study compares users' interaction and experience with three concepts representing different levels of embeddedness.

Chapter 6 - Embedding Comprehension Assessment into Digital Reading and Listening: We also explore how we can embed language proficiency assessment into everyday activities such as reading and listening. This chapter presents a method of using physiological sensing (EEG) to implicitly detect users' vocabulary comprehension during digital reading and listening. By extracting unknown words from the users' media content of choice, we have the opportunity to generate learning material tailored to users' interests (learning objectives) and individual proficiency level.

Part IV - Connecting Micro-Learning Sessions

Chapter 7 - Design Space for Task Resumption Cues: This chapter explores a promising interruption mitigation strategy proposed in Chapter 4 in greater detail - task resumption support through memory cues. We present an analysis

of literature from related fields on the application of memory cues to guide users back to their original task after an interruption. We outline a design space and derive six implications for designing task resumption support in mobile learning.

Chapter 8 - Memory Cues in Mobile Learning Applications: Building on the recommendations for task resumption cue designs in Chapter 7, we derive a set of memory cues and embed them into a MLL application. This chapter reports on the implementation and two user studies conducted to evaluate their effect on learning and User Experience (UX) in a controlled laboratory environment and in the wild.

Chapter 9 - Outlook - Memory Cues Beyond Language Learning: Whereas Chapter 8 revealed the need to adapt the complexity of mobile learning task resumption cues to the difficulty level of the content, this chapter represents a first outlook on how task resumption cues could be implemented beyond language learning. We implement a mobile app teaching programming and examine the fit of different cue designs for different task complexities. We evaluate users' experience in a field study.

Part V - Conclusion and Future Work

Chapter 10 - Conclusion and Future Work: In this chapter we synthesize the results of the individual chapters in regards to the research questions posed here. We discuss how this thesis contributes to the vision of embedded MLL and reflect on the limitations that still lie ahead. Lastly, we present opportunities for future work.

2
Background and Foundations

"There is no end to education. It is not that you read a book, pass an examination, and finish with education. The whole of life, from the moment you are born to the moment you die, is a process of learning."

Jiddu Krishnamurti

In line with the quote by Jiddu Krishnamurti, we consider learning a lifelong process, and mobile devices have the potential to support it. In this chapter, we outline theoretical foundations and background on the two main topics of **(language) learning** and **mobile device usage in everyday settings**. While this thesis is mainly rooted in the HCI domain in Computer Science, this section will also extend to the field of Psychology to explain the basic cognitive functions of learning and language acquisition. We will start each section by introducing and defining relevant terms that will be used throughout this thesis. In Section 2.1 we discuss the foundations of learning and explain how the human memory works. In Section 2.2, we describe how ubiquitously mobile devices are used in everyday contexts, independent from the use case of learning, and how frequently we interact with them. Section 2.3 synthesizes both domains to show how learning can work on mobile devices. We explain the paradigm of *Micro-Learning*, which entails breaking down learning content into small units with short interactions [27], and discuss second-language learning as a specific use case. We further outline the central concepts of self-regulation, motivation, and personalization, which greatly affect mobile learning performance (cf. [41, 70, 84]). While this chapter outlines concepts relevant for the larger scope of this thesis, the related work section in each of the following chapters will provide more detailed insights into the topics relevant to the individual chapters.

2.1 Cognitive Processes

The following section outlines the pathway of human information processing. We take an in-depth look into the process of learning and particularly aim to shed light on how memories are acquired, stored, and recalled.

2.1.1 Information Processing and Memory

The terms *memory* and *learning* are closely related. According to the definitions of the American Psychology Association (APA), *learning* can be understood as "[...] *the*

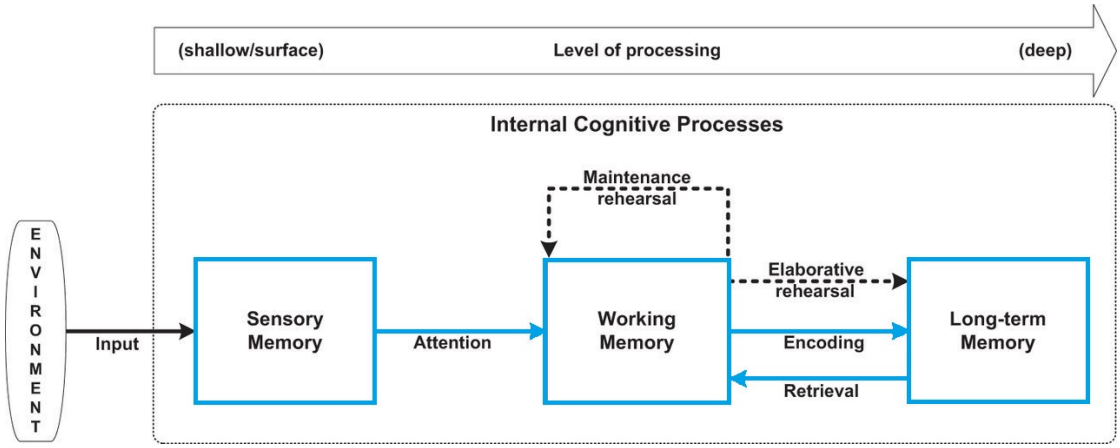


Figure 2.1: The multi-store model by Atkinson and Shiffrin [17] outlines three stages of information processing in the human memory. From left to right: Environmental stimuli, the sensory memory, the short-term or working memory, and long-term memory. Khalil and Elkhider [176] further highlight the different levels of processing and the two types of rehearsal.

acquisition of skill or knowledge". The related concept of memory is defined as "[...] *the expression of what you've acquired*" [11], i.e., the documentation of the learning process. The concepts, however, are not distinct and should be treated as similar in regards to their cognitive processes.

The processing of information and their transition into memories is outlined in the *Multi-Store Model* by Atkinson and Shiffrin [17], extended by Khalil and Elkhider [176] (see Figure 2.1). It shows that environmental stimuli in the form of raw unprocessed information first reach our sensory memory, where they decay in less than one second. Filtered by our attention, a subset of the input reaches the working memory, also called Short-term Memory (STM). Through repetition and rehearsal of the information in the working memory, the information is encoded and thereby transferred to the Long-term Memory (LTM), where it can persist indefinitely. The extension of the Multi-Store Model by Khalil and Elkhider [176] emphasizes the amount of processing involved in the three stages. While the sensory memory deals shallowly with raw data, all knowledge stored in the LTM has been thoroughly processed and embedded into our existing knowledge base.

Contrary to the LTM's unlimited capacities, the STM can store around seven so called *chunks* – bits of information such as numbers, words, sounds, or similar. Early research by Miller [237] showed that the number of items we can keep in the STM is 5 ± 2. This number has been re-evaluated over the last decades, and current research suggests that we can store between three and five meaningful items (for young adults) [73].

In the past, the Multi-Store Model has been criticized for being over-simplistic and for giving the impression that its components are distinct and work in a linear fashion.

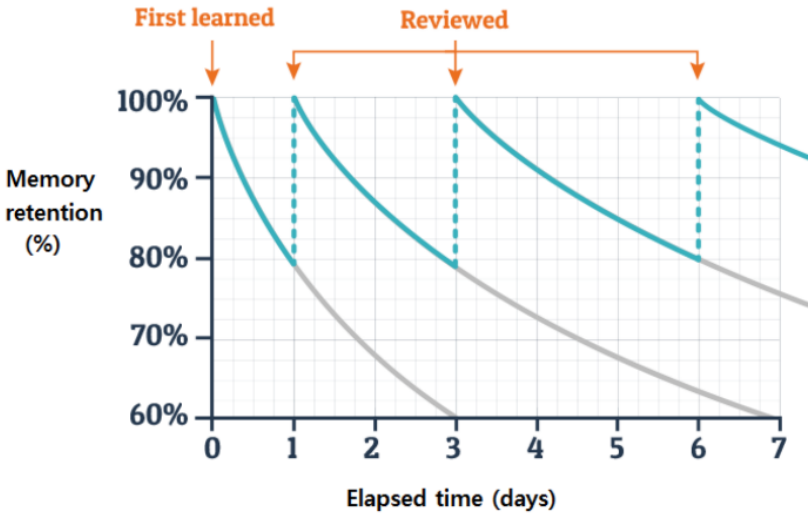


Figure 2.2: Ebbinghaus's forgetting curve and review cycle as depicted in Chun and Heo [64].

We acknowledge the complexity of the processes involved in learning and memory. However, for the scope of this thesis we consider this approach sufficient.

One aspect that is highlighted in this memory model is the necessity for rehearsal of stored knowledge. Memory transience, also known as *forgetting*, describes the decline of memory retention over time. The German philosopher Hermann Ebbinghaus reported on the first evidence of decaying memories in 1885. He observed that, if not reviewed, previously learned content will fade from the memory within days if not hours after acquisition (cf. Figure 2.2). Ebbinghaus's study has been replicated many times inside and outside of laboratories [294]. In a nutshell, information that is not maintained by occasional rehearsal will be forgotten.

Ebbinghaus's studies involved the remembering of nonsense syllables, thus, testing people's *semantic knowledge* (fact-based learning or *factual knowledge*) residing in their LTM. Charles Thompson translated this experiment to *episodic knowledge* (event-based learning). Both knowledge types are subgroups of the *declarative knowledge* (cf. Figure 2.3). In his study, university students were asked to write down one event each day for themselves and one for their room mate, adding its momentary memorability. Thompson evaluated their memories of those events at various time delays. Even though memorable events were encoded better, they were forgotten at the same rate [327]. In contrast to declarative knowledge, implicit or *procedural knowledge*, such as the skill of riding a bike, seems to be the least forgettable and the most time-consuming to acquire. However, if one tries to explain exactly the steps of the procedure instead of just doing it, the information is not as accessible since it is unconscious [294].

In language acquisition, multiple knowledge types play a role. The declarative/procedural (DP) model claims that language is stored as a combination of facts

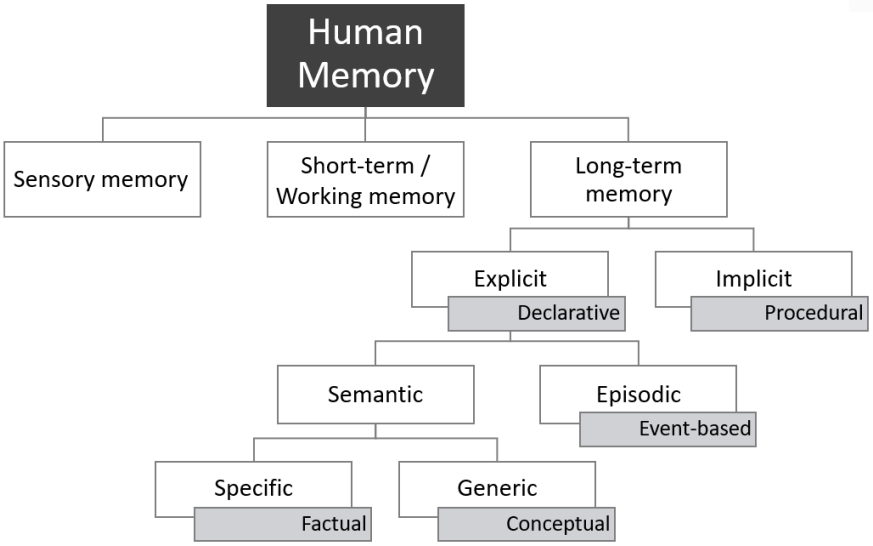


Figure 2.3: The different types of the human memory based on the distinction by Tulving and Schacter [332].

(i.e., declarative knowledge) and sequences (i.e., procedural knowledge) [333, 334, 335]. The differentiation occurs between the memorized "mental lexicon" and a "mental grammar". While the lexicon in its essence stores the vocabulary and phonology, the grammar contains the rules needed to combine words to form complex constructs[1] [334]. Since the vocabulary knowledge is part of the declarative memory, it is prone to fast decay if not rehearsed according to Ebbinghaus's theory. In contrast, the procedural grammar knowledge shows a slower decay but requires more time and effort to acquire in the first place.

2.1.2 Knowledge Repetition and Retention

Over the years, certain techniques have evolved that aim to optimize the repetition and thus, retention of declarative information. An important factor in memory consolidation is the timing of repetition. For example, the "Spaced Repetition" technique, i.e., reviewing an information chunk with increasing time intervals in between, can increase the LTM retention [168]. In contrast, mass repetition practice - "cramming" - is not an effective method for long-term learning. Retrieval of information learned by mass repetition is negatively impacted by the time elapsed until it has to be recalled. The practice has to be spaced appropriately, repeated, and ideally interceded by small tests (so that the retrieval process is practiced as well) [168].

A system that applied spaced repetition is the Leitner index [123]. It is a popular metric for modeling repetition schemes. The system from the 1970s uses flashcards for

[1] Please note that for the scope of this thesis, the model is depicted in a simplified way.

remembering facts and information chunks (i.e., declarative knowledge). The flashcards are sorted into specific boxes representing the levels of how well users remember the content. At the beginning of the learning process, all flashcards are put in box 1. The learner starts with the first card and tries to recall the solution on the back. If it is correct, the flashcard moves to box 2. If it is incorrect, the card remains in box 1. In case a flashcard from box 2 or 3 is answered incorrectly, it returns to box 1. Each box is assigned a certain repetition frequency that corresponds to the learner's proficiency, creating a spaced repetition scheme. A common approach is to double the time between the repetitions with each box. For example, while box 1 is repeated every day, box 2 could be revisited every second day, and box 3 every fourth day. With this method, the learner interacts more frequently with content they are less proficient in.

The spaced repetition practice can be implemented as a physical box for analogue flashcards but can also be translated to a digital equivalent in which learning content is sorted into digital proficiency boxes (e.g., the mobile apps Anki[2], SuperMemo[3] or MicroMandarin [100] and MemReflex [99].

2.1.3 Recognition vs. Recall

The flashcard method (no matter if implemented as analogue or digital cards) differs from other learning tasks such as multiple-choice in the way the users retrieve the memory. In general, we distinguish memory recognition and recall. In the process of recognition, previously learned information is detected among a set of potential options. For example, in a task to translate the second-language (L2) Spanish word 'gata', the learner can be given the option to choose between the native (L1) English words 'dog' and 'cat'. This recognition task is commonly applied in single or multiple-choice test formats.

In case of language teaching, literature further distinguishes between active and passive recognition and recall (see Table 2.1). Active knowledge is described as the generation or recognition of the word form, whereas passive knowledge is concerned with the word meaning. Passive recognition, i.e., the presentation of an L2 asking the learner to select the correct L1 translation out of a set of options, is considered the easiest. It is followed by active recognition, passive recall, and active recall, the last being the most difficult form [205]. It has to be noted that these vocabulary tests assess only a narrow part of users' language proficiency as they do not accommodate for general communicative competencies [205].

[2] Anki App: `https://www.ankiapp.com/`, last accessed January 3, 2022

[3] SuperMemo App: `https://www.supermemo.com/de`, last accessed January 3, 2022

Table 2.1: Distinction between active and passive vocabulary recognition and recall. The colors indicate the respective difficulty for learners [205], from easy (green) to difficult (red).

	Passive	**Active**
Recognition	Presentation of the L2 word, detect the L1 translation from a set of potential distractors	Presentation of the L1 word, detect the L2 translation from a set of potential distractors
Recall	Presentation of the L2 word, generate the L1 translation	Presentation of the L1 word, generate the L2 translation

2.1.4 Memory Augmentation

Various techniques can support the encoding and recall of memories. A common method is the memory *Cue*, which is in general a stimulus that is used to guide behavior [9]. If used for memory recall, it is called *Memory Cue* [140] or *Retrieval Cue* [14]. This technique is applied when learning vocabulary in L1-L2 pairs. In the example of active recall, the presentation of the L1 word acts as memory cues, triggering the recall of the L2 word. Using cues to enhance recall is called *Cued Recall* [49].

Cued Recall is also applied in the common practice of *Mnemonics*. This method creates artificial associations between two pieces of information, one already encoded and one new piece; for example, when a person tries to remember a phone number by breaking it into chunks of numbers that match birthdays, historical events, or other important numbers (cf. [12]).

Memory Augmentation strategies, particularly cued recall, are frequently used in educational settings. In schools, asking the pupils at the beginning of the lesson about the content of the prior lesson is a common practice to reactivate their memory. Research from Psychology highlights the central role of including prior knowledge into the teaching process [185]. Cueing information from the LTM to make it accessible for the STM can occur in an *open* or *specific* format. Open memory re-activation concerns the activation of a broad topic, such as asking the class to brainstorm whatever they can remember from last week's lesson [324]. Specific memory re-activation aims to improve the recall of a specific chunk of information. This is commonly achieved by posing questions [185]. Which strategy to choose depends on the specific situation [324].

2.1.5 The Effect of Technology on Memory

In the last decade, the development of technology towards being a constantly available companion has led to a change in how we deal with information, but having continuous access to information through the Internet has consequences on our memory. Our

organic memory is evolving so that we aim to remember how to locate a chunk of information rather than the actual content. In other words, if users expect to have access to certain information via the internet, their ability to recall the actual information decreases. For example, if we know we can look up the ingredients for a cake recipe on our favorite baking website, we only have to remember how to find the website and not every detail of the recipe. This effect has been coined the "Google Effect" and describes how the internet has become a collectively stored external memory [314].

However, since being able to recall certain information is crucial for creative processes (i.e., brainstorming or idea generation) and also for the successful integration of new information into our LTM, we argue that technology should rather be used to augment the human memory rather than replace it.

Besides the negative effects of technology on memory, several techniques have been investigated that can positively affect the recall abilities. For example, virtual memory palaces [353], using slide decks as a tool to support the recall of previous work meetings Niforatos et al. [248] or multimedia memory cues for life-logging Dingler et al. [87]. We will discuss the design and application of memory cues, particularly for mobile learning scenarios, further and in detail in Chapter 7.2.

2.2 Mobile Interaction

To better understand the challenges and opportunities of learning with mobile devices, this section will further outline the interaction of users with mobile devices in everyday settings independently from the use case of language learning. We will discuss how seamlessly smartphones are nowadays embedded in users' everyday lives and what challenges their ubiquity implies, particularly regarding our limited cognitive capacities.

2.2.1 Interaction Behavior in Everyday Settings

Using a mobile device in an everyday setting means that interactions are diverse. A study by Falaki et al. [103] recorded that users interacted with their mobile device on average between 10 and 200 times per day, whereof each session lasted between 10 and 250 seconds. Interaction further depends on the situation and environment, as users have to continuously divide their cognitive resources between monitoring their surroundings and interacting with the device. Due to our limited cognitive/attention resources, being mobile is costly [256]. For example, when waiting for a bus at the bus stop it is not only about sitting and waiting; we monitor the street to see the bus approaching, we look at the watch to check the time, we observe the environment surrounding us for people approaching. In summary, we continuously interpret all sensory input; visual, auditory, and olfactory, and process all incoming information.

In addition, we might remember that we wanted to make a doctor's appointment or add bananas to our mental grocery shopping list.

Prior work has even shown that half of our interactions with our smartphones take less than 30 seconds, with only one in ten interactions exceeding four minutes [354]. In particular, many interactions do not even reach the threshold of fifteen seconds, what Ferreira et al. [106] defined as "micro-usage" situations. For example, users briefly reply to a message, check their phone for new push notifications, or dismiss an alarm. In some cases, interactions are designed to be short on purpose, such as the *microinteractions* proposed by Ashbrook [16], which do not exceed four seconds. For the purpose of keeping the interaction short and simple, he argues that it can help minimize the interruption caused by a task. Within four seconds, the user can fulfil the task and resume the primary task. While Ashbrook's work is primarly focused on wrist-based interactions, the idea works just as well for smartphone interactions. In the case of learning, microinteractions have been proposed to be embedded into the phone's status bar as a notification [91] or shown on the lockscreen [81].

To make the most out of short interactions, Micro-Learning breaks down the learning material into small content chunks solvable in short interactions. Besides the duration of the interaction, the timing is also crucial. Using attention detection mechanisms, prior work has tried to determine opportune moments to trigger users to interact with their smartphones to not disturb or interrupt them. In moments of boredom [261], task-breaks [250], or activity transitions [145], notifications can be presented, recommended content shown, or learning activities triggered.

2.2.2 Ubiquitous and Embedded Interaction

The vision of the ubiquitous computing age is to design technology that embeds itself invisibly and seamlessly into people's lives. Already thirty years ago, in 1991, Mark Weiser envisioned ubiquitous technologies as follows: "*The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.*" (Weiser [346], p.1). In many aspects this vision has already become reality: a car that knows when the driver is approaching, the smart home system that turns on the lights when it gets dark outside, the smart watch that can differentiate whether you are running, biking, or inline skating.

The basis for this vision is that technology is embedded, either on a technical or on a conceptual level. From a technological perspective, sensors and actuators can be embedded in everyday objects to turn them into smart systems. Conceptually, embedded technology focuses on how we can make the interaction between human and computer more natural and seamless. Examples from research are the design of new input and output techniques, wearable technology, or implicit interaction mechanisms [97, 295].

As this thesis aims to investigate how we can embed mobile learning into everyday settings, we particularly refer to the conceptual embedding of technology. In particular,

the integration of mobile technology in different usage situations so that the interaction becomes implicit or seamless.

2.3 Mobile Learning

This section combines the fundamental psychological concepts described in Section 2.1 and describes research that aims to support learning with mobile technology (cf. Section 2.2). We will define the concepts of mobile and micro-learning and outline the opportunities which the ubiquity of mobile technology offers for learning in our everyday lives.

In this thesis, we consider Mobile Learning as learning with a mobile device. In a very general form, *learning* can be considered as "*the acquisition of novel information, behaviors, or abilities* [...]"(cf. American Psychological Association (APA) dictionary [10]). We emphasize that for the use in this thesis we consider the term *learning* as referring to conscious knowledge of something rather than the subconscious acquisition in line with Krashen [183]. It is possible that information acquisition and learning take place subconsciously. However, attention is the filtering mechanism regulating the incoming data stream. It decides which information chunks will receive deeper processing, making it essential for learning [298]. In general, our attention is limited – we cannot focus on every stimulus we perceive but we need to allocate our attention to what is important.

Mobile Learning is a sub-domain of Technology-enhanced Learning (TEL), and is officially defined by O'Malley et al. [253] as:

> **Mobile Learning** encompasses "any sort of learning that happens when the learner is not at a fixed, predetermined location, learning that happens when the learner takes advantage of the learning opportunities offered by mobile technologies." – *O'Malley et al.* [253] (page 7)

Already back in 2006, Chan et al. [58] anticipated that within a decade, personal and portable learning technologies would become ubiquitous. He was right in his prediction. More precisely, the research and development of mobile technologies for learning has by now far surpassed the goal of ubiquitous presence. Smart tutoring systems can teach sign language [105], augmented reality techniques can teach users motor tasks using elaborate visualizations, and AI chat bots can be teachers and casual discussion partners all in one [260, 340]. These achievements are in large part due to the enormous development of today's technology and its usage, resulting in mobile devices being our primary computer. Small, hand-held computing units with various sensors and capabilities have made the advancements in mobile learning possible in the first place.

When we refer to *Mobile Learning* in this thesis, we consider any situation where learning takes place with the support of a mobile device. The technology acts as a medium to enable a user to perceive and process knowledge independently from time and location. It can include learning in different everyday life settings of a user (e.g., when waiting for the tea water to boil, on the sofa while watching TV). Mobile learning is neither limited to either the common use case of "learning-on-the-go" nor to situations where the mobile device enables the user to learn. The discussion has been ongoing of whether it is necessary to differentiate between mobile and fixed location learning in which users regularly take their device to a certain location to learn (e.g., the university) [80].

In our definition of mobile learning in everyday settings, the essence is that the situation can vary. Sometimes the user learns in the subway, sometimes at home, sometimes on the move. The mobile device itself does not even need to be active in the respective situation but can take a passive role, such as to implicitly collect physiological data about the comprehension or learning process. Therefore, in the scope of this thesis, we consider every waking moment of the user in which the mobile device is present a potential use case for *learning with mobile devices*. This further includes outside activities such as riding a train, having a conversation with a friend, or browsing through the supermarket.

2.3.1 Ubiquitous, Seamless, and Embedded Learning

In contrast to Mobile Learning, the term *Ubiquitous Learning* is concerned with learning using ubiquitous technology. Thus, it is considered a more specific domain of Mobile Learning and can take into account other types of devices than mobiles. The boundaries between these two concepts cannot be drawn clearly, but are extensively discussed in the literature [4]. While this thesis aims at embedding the learning interaction more seamlessly into users' daily lives, it still places mobile devices at the center of the interaction. Therefore, we consider this thesis located in the broader domain of mobile learning. However, the presented concepts and ideas could be generalized and extended to encompass other more ubiquitous technologies and interaction forms, reaching into the domain of Ubiquitous Learning.

Similar to research aligned with the *Seamless Learning* paradigm, we envision building systems and tools that can be used in multiple settings, i.e., in any imaginable situation in users' everyday lives. The concept of seamless learning has been coined as "[...] *utilising affordances of mobile technology, [to] bridge the gap between formal and informal learning, and encourage students to learn in naturalistic settings for developing context-specific competences.*" (Looi et al. [216], page 3). Chan et al. [58] refer to *seamless learning* as students who are learning "[...] *whenever they are curious in a variety of scenarios and [switching] from one scenario to another easily and quickly using the personal device as a mediator.*" In further explanations, Chan et al.

[58] particularly distinguish learning situations according to other people present, such as students, groups, mentors, or professionals. For the current scope of this thesis, we see a similarity in regard to the transition between learning settings, although the distinction between formal and informal settings is as of now negligible.

In the research community, there is not one distinct perception or definition for the term Embedded Learning. While some refer to embedding learning into context (i.e., [215]), others focus on embedding learning into a community (cf. [173]), and yet others consider embedded learning synonymous with learning-while-doing. From the HCI perspective we take in this thesis, we consider embedded learning as learning through embedded technology or embedded interaction. Section 2.2.2 will present further details on the technological perspective of ubiquitous and embedded computing.

2.3.2 Micro (Language) Learning

The Micro-Learning approach is a specialized way of teaching content on mobile devices. Based on psychological research that states the benefits of high repetition counts in contrast to long learning streaks [74], this approach is often used, for example, in language learning. Micro-learning focuses on frequent repetitions while presenting *micro-content* units in *micro-interactions* to help users learn without information overload [44]. In the case of mobile language learning, micro-learning means that lessons are broken down into small information chunks. The user learns a set of consecutive words or phrases ordered by topics. New contents are unlocked gradually. The absence of these explicit contents shows the emphasis on teaching simple and easily processable learning materials in MLLs. In its essence, the two central pillars of micro-learning are (1) *micro-content* and (2) *micro-interactions*, which we will further describe below.

The limitations of the mobile devices create challenges for the design of mobile learning applications. The small screen of mobile devices limits the amount of content that can be presented at once. Thus, in the paradigm of micro-learning, the content is broken down into smaller micro-content units, aiming for a more comfortable interaction on mobile devices [44].

From a memory perspective, these smaller units provide the opportunity for easier processing and frequent repetition. When engaging in learning with mobile devices, higher repetition counts are more favorable than long learning streaks [74]. Especially considering the use of mobiles in uncontrolled environments prone to interruptions and distractions, it is beneficial to keep the amount of information we perceive at once within the boundaries of our short-term memory (i.e., 5 ± 2 [233, 237]). Short learning sessions further cater to users' desire to learn spontaneously whenever and wherever they are, such as on the go or while waiting [53, 85, 91].

Explicit grammar explanations are rare: Heil et al. [135] evaluated 50 mobile learning applications and showed that they were only included in 20% of the applications.

Grammar, which can be classified as procedural knowledge, can be either displayed explicitly or taught implicitly. While implicit grammar knowledge is inferred by the user during language exposure without mentioning terminology, explicit teaching includes either the presentation of rules or corrective feedback with explicit references to grammatical errors [135]. While 19 out of 50 reviewed applications in the presented study contained implicit grammar instructions, only 10 provided explicit grammar presentations. In various theories, languages is not viewed as a sum of knowledge on vocabulary and grammar but rather as a complex cognitive skill. In the monitor model, Krashen outlines five central hypotheses, of which the first one is '*The Acquisition versus Learning Hypothesis*'. It postulates that languages are acquired subconsciously, potentially in informal environments, rather than actively learned [121]. This model somewhat contradicts the DP model, as it states that all parts of languages can be learned.

Besides the amount of content that is presented, it is further important how the content is presented and how users are expected to interact with it, i.e., the User Interface (UI). Mobile devices come with certain requirements such as a small keyboard, which makes writing text more cumbersome on smartphones than on a laptop keyboard [257]. Churchill and Hedberg [65] propose short one-step or micro-interactions with immediate feedback for mobile learning. Further, the design principles for multimedia learning outlined by Mayer [229] apply with some reservations; for example, avoiding redundancies, and structuring applications to be user-paced rather than continuous.

2.3.3 Lifelong Learning

In line with the quote at the beginning of this chapter, we consider learning a lifelong task. However, as the quote already stresses, we need to emphasize that learning does not equal formal education. Especially when we exceed the formal education phase, learning happens either implicitly or on-demand. Today's technology presents people with access to sheer unlimited (new) information and requires them to update their knowledge constantly. According to Fischer [107], learning on-demand is a promising approach to tackle this challenge as, besides other aspects, learners immediately see the usefulness of the knowledge they have acquired. In the example of our scenario in the Introduction, Anna started learning Portuguese after deciding to go on vacation in Portugal. Therefore, her knowledge became relevant to her soon after acquisition.

Already back in the mid-90s, there was a vision of making Lifelong Learning ubiquitous, meaning embedding learning into authentic contexts and activities that learners would be intrinsically motivated to perform [108]. Mobile and ubiquitous technologies today have the opportunity to embed learning seamlessly into our daily lives and support the acquisition of new knowledge over a lifetime. However, the mere availability of information and access to digital learning tools does not necessarily mean that learners will make use of it. Several internal and external factors influence the decision of

whether a learner will engage in a learning activity or not. We outline several of these below, including self-regulation skills, motivation, and personalization:

Self-Regulation in Mobile Learning

Outside of formal classroom environments, being able to independently and actively interact with learning technology, i.e., *self-regulated learning*, is an essential factor for successful online learning [41]. Prior work has shown that students' self-regulation skills vary greatly and therefore some require greater support through technology [201, 202, 341]. For self-directed use of technology in language learning, the perceived usefulness of technology for learning, and perceived compatibility between technology and learning expectations are closely associated with students' technology use [201]. Besides the external support by the technology, self-regulative behavior is closely related to student's inherent skills, such as time-management and the ability to structure learning practice, and their intrinsic motivation [341]. The latter is a type of motivation that comes from the learners themselves and is based on their individual curiosity.

Motivation and Personalization

Demouy et al. [85] found that users often interact with mobile language learning apps out of curiosity for the technology and for entertainment purposes. Having a high level of intrinsic motivation, meaning being motivated to learn about what one finds interesting and not just study for the sake of a test or examination (extrinsic motivataion), is positively associated with learning [84, 161]. Intrinsic motivation, as well as involvement and learning, can be improved by providing students with personalized learning contexts, such as learning with content adapted to their backgrounds and interests [70]. Thus, personalization is important to foster motivation and self-regulatory behavior.

The term *Personalization* is diversely applied in the literature and in regards to educational technology can be understood as:

> "**Personalized learning systems** are learning systems that consider the individual differences of learners and tailor the learning experience of learners to their current situation, characteristics, and needs." – *Graf and Kinshuk* [125] (page 1)

Since learning is a process that is highly dependent on the learner's prior knowledge, the activity of learning in itself is an individual process. Yet, in classical classroom education, all students normally learn in the same setting. In other words, they have the same learning objectives, receive the same learning materials, the same instructions, and have the same time to accomplish a task. The use of novel technologies for learning and the pervasiveness of mobile devices in people's everyday lives now provide the opportunity to support learners on a highly individual level.

In the introduction, we already outlined that personalization in learning has been declared one of the 'Grand Challenges for Engineering' and is part of the 'Seven HCI Grand Challenges' of Stephanidis et al. [319]. In the following, we will outline what facets of learning can be personalized and what the current challenges and limitations are.

The Personalized Learning System of technology-based teaching is often referred to as *adaptation*, in which software "[...] *automatically updates its functionality based on input received or data processed*" (Heil et al. [135], page 41.) In many cases of personalization, digital learning tools provide personalization either through *adaptive assessment*, *adaptive sequence*, *adaptive content*, or a combination of these [101].

When a system adapts its assessment, it means changing the exercises and questions based on students' previous answers (i.e., providing easier or more difficult answers based on previous correct or incorrect answers). Lastly, adaptive sequencing refers to tailoring the order of the learning content presentation to the user [101]. A common example from the domain of language learning is spaced repetition practice according to the Leitner index [123]. As outlined in Section 2.1.2, spaced repetition schedules learning tasks to maximize retention. When a task is incorrectly answered, the repetition frequency is increased and time between repetitions decreased [123].

Adaptive content means that the system is reacting to a student's input and comprehension, for example, by breaking down skills and providing additional explanations and feedback. If the system adapts the content to target personal learning objectives, we move from curriculum-driven learning toward interest-driven self-regulated learning. By setting personal learning objectives and choosing tasks that match individual goals and interests, intrinsic motivation increases. Further, it supports the long-term adoption of a tool or system, making it a central requirement for the adoption of lifelong learning technologies [108].

II

Understanding User Behavior

3

Mobile Learning Session Characterization

In the past thirty years, the development of technology has revolutionized the way we learn. Back in the 1990s, technology had been a helpful add-on to existing practices. In the mid 2000s, with the widespread availability of portable devices, such as mobile phones, personal digital assistants, or laptops, the integration of learning into the everyday lives of users began. Yet, in the context of mobile language learning (MLL), research was still in its infancy and far away from exploiting its full potential [190].

Recent improvements in mobile technology and the trend towards short learning sessions made language learning a ubiquitous activity. Hence, the usage of and research on mobile learning apps steadily increased, fostering learning anytime and anywhere [44]. This increasing autonomy and further advancements in sensor technology and interaction techniques led to more research around the situational learning context. Kukulska-Hulme [192] already emphasized the need to "[...] *review individual learner experiences* [...] *to build up a picture of emergent practices and formulate the implications for the design of language teaching and learning now and in the future.*" (Kukulska-Hulme [192], page 1). This strand of research is concerned with the question of how people experience and make use of the constant availability of learning opportunities that mobile devices offer.

To gain a deeper understanding of learners' experiences, this chapter aim to dive deeper into the situations in which learners use mobile learning apps. We report on the results of an online survey in which users described their subjective perception of common usage situations of mobile learning apps. We report on those situations, describe frequent and less frequent contextual factors, and derive patterns of common usage situations. We discuss the results and outline the potential challenges of mobile learning apps in everyday use.

Thus, we aim to answer the following research question:

> **RQ1a:** How do people use mobile learning applications in everyday settings?

This chapter is based on the following publications:

- Schneegass, C., Terzimehić, N., Nettah, M., and Schneegass, S. (2018). Informing the Design of User-adaptive Mobile Language Learning Applications. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia* (MUM 2018). Association for Computing Machinery, New York, NY, USA, 233–238. DOI: 10.1145/3282894.3282926

Some text passages were taken verbatim from this publication. Further, this section is supported by the Bachelor thesis of Mariam Nettah (see detailed collaboration statement at the beginning of this thesis).

3.1 Evaluations of Mobile Learning Usage

By using mobile learning applications, learning becomes possible anytime and anywhere [44]. The flexible usage presents a great opportunity but also adds to the list of things that compete for our attention. As of now, learning happens mostly on demand, meaning users enjoy the flexibility of learning whenever they have the time and desire to do so [317]. Fischer [110] stated that the true challenge of human-centered computing systems will be to decide which information should be presented when, how, and to whom. By learning more about the users and about their habits when interacting with mobile learning applications, we can design systems that detect and better support learning in opportune moments. First, we have to better understand the necessities and limitations of the users' everyday situations. Which situations do consider an opportune moment for learning?

Since the early days of mobile learning technologies, research has tried to map users' preferences in their interaction with it. *How do learners use mobile applications?* is one of the most frequently asked questions in this domain that many have tried to answer. In their extensive survey and interviews, Demouy et al. [85] show that users aim to fill gaps in the daily schedule with learning sessions and increase the frequency of exposure to the language. The majority of sessions happen informally and rarely planned, thus, adopting their learning habits to their current situation and needs. The convenience of learning spontaneously where ever one goes, hereby personalizing learning in a manner that fits the individual user's daily life, is the central benefit of mobile technology [317]. But what does learning look like when it happens in users' daily lives and what are situations or contexts users' consider a good fit for learning?

Context can include various factors such as location, time, or device, is gathered to better understand how learning takes place in informal settings and create a personalized learning environment adapted to the characteristics of each individual learner and their environment [98].

Looking at a large population of over 4000 users' of the *Busuu* language learning app, Rosell-Aguilar [285] took the first step and evaluated individual characteristics of mobile learning sessions. Through an online questionnaire they surveyed Busuu users about their learning habits, including frequency, session length, and (least) favorite features. The majority of users was novices who used the app several times a week or less (around 75%), whereof only one third of the sessions is planned ahead. The

average duration of each session was in three quarter of the answers reported to be 15 minutes or longer.

However, the individual contextual factors have only limited expressiveness when it comes to describing a learning situation. For the scope of this thesis, we refer to a learning *situation* as a combination of characteristics that describe the particular moment of learning. As described above, the term *context* commonly refers to a variety of situational characteristics such as location, device type, or time of day. It can be subdivided into, among others, users' physical, temporal, task, social, or technical context [166]. In this chapter, we will focus on usage *situations* by viewing contextual factors as related constructs that can be dependent and form patterns. By looking at these patterns, we aim to better understand users' perception of opportune moment for learning and discuss implications for the design of future mobile learning applications.

3.2 Surveying Common Usage Situations

In this section, we report on an exploratory user study aiming to create a deeper understanding of the situations in which learners use mobile learning applications. This evaluation aims to get insights into the subjective users' perspectives on their learning behavior and uncover larger patterns of usage situations that go beyond context descriptions. We chose to conduct an online survey asking users to describe two or more common Mobile Learning Situations. While we offered guiding questions, the survey left room for participants to add further information about situations. We chose the online survey format to attract a diverse set of participants. The anonymity of the survey decreases the social pressure and increases the likelihood of unbiased answers. From the survey answers, we aim to assess individual usage facets (such as location or time of day) and uncover associations between facets and patterns of usage situations.

3.2.1 Sample

We recruited 74 participants via our university mailing list and social media (54 female, 20 male, age range between 17 and 32, $M = 23.30$; $SD = 4.33$) who stated to have used MLL apps at some point in the past. Aside from 3, all had at least a high school degree (28 even a bachelor's degree, 14 a master's degree, and one a doctoral degree). Of these 74 participants, most of them were students (64%) or young professionals. We found these to be a representative group, as students strive to learn new languages for various reasons, as for fun, vacation, or student exchange. Since this study was conducted in Germany (but presented in the English), the sample also contained international students wanting to learn or improve their German skills. The participants stated to speak at least two and max five different languages ($M = 3.95$; $SD = 0.93$) on various proficiency levels (i.e., basic to native). As their first language, 55 people stated

German, and 19 participants stated another native language such as English, Russian, or Turkish. The most common second language was English, with 54 occurrences. In total, participants spoke 28 different languages.

3.2.2 Procedure

In the online survey, users had to describe at least two common situations in which they use a language learning application. The survey offered a table to fill in additional details regarding the learning situations to characterize those situations further. We asked for general information on location, time of the day, device, duration, planning, and frequency as performed in the study of Demouy et al. [85]. Further, the survey inquired about the situation's noise level, users' company during learning, if it was a public or private setting, additional/parallel activities, estimated stress level, and left space to enter additional details the participants considered relevant. Our aim was to include both external and internal factors into the description.

3.2.3 Results

After removing incomplete descriptions, our data set included a total of 131 common learning situations. In the following, we will at first present a summative overview of all dimensions of all facets of the 131 learning situations[1]. Afterward, we will report on our analysis to uncover associations among the different facets. Lastly, we cluster all reported usage situations and outline five commonly stated situations covering 82% of all described situations.

Descriptors The **location** in which most of the learning situations happen is home (74), followed by public transportation (44). Few participants stated public places such as library (4), the university (3) in the waiting room of a doctor (1), in the office (1), or in school (1). The remaining three situations described learning on vacation (1), in the car (1), and "on the way" (1). These locations were further specified as public (43), semi-public (7), or private (79).

Noise levels got characterized most frequently as low (61), followed by medium (38), and high to very high (32).

When looking at the **time of the day**, the majority of the described learning situations occur in the evening (51). Still, many participants characterized scenarios where they learn in the morning (39) or the afternoon. Less often, learning happens at noon (8) or at night (5).

[1] As the survey allowed for the selection of more than one descriptor for a situation, the sum of the individual notions can exceed 131.

The Smartphone is the preferred learning **device** in 98 of the outlined situations, compared to a tablet (10), laptop or desktop computer (24), or books (2) in the category device.

We further looked into the **duration** of learning situations. 93 of the characterized situations last 5 - 20 minutes. The overall range is from 5 to 150 minutes.

In 111 situations, participants have no **company** while learning (or are only around strangers in public environments); some learn while being with their partner (7), family (6), or when friends (5) or colleagues (2) are around.

In most of the described situations, learning is performed while having a low **stress level** (87), followed by a medium (36) or high stress level (8) (none mentioned for two situations).

The participants reported **planning** roughly one-third of all situations in advance (43) and estimated the situations' **occurrence frequency** as multiple times per day (5), daily (39), at least once a week (60), multiple times per month (15), or rarely (10), with two situations being not quantifiable.

This indicates an overall high usage frequency. The participants stated accomplishing various additional activities during the use of MLL apps, as consuming media content (music/TV/videos/Netflix) (24), riding public transport (39), or eating/drinking (5).

As additional information, three participants stated to use headphones during learning and disabling the audio output feature of their learning application. All of them noted this specification for situations when learning on public transportation. One participant further specified that they are standing when learning on public transportation.

Associations among Descriptive Facets While the first part of this chapter reported an exploratory analysis of usage situations through an online survey, the second part of the analysis aims at better understanding the relationship among the reported facets. We hypothesize that the factors defining a usage situation are associated with each other. The location and noise level are likely to be related. For example, learning activities at home are less likely to be accompanied by high noise levels than learning on public transport. Nonetheless, as noted before, most factors can vary according to users' subjective perception of the situation and should be considered observer-relative. Therefore, we will continue to follow the exploratory evaluation approach and look for associations among any of our survey variables. We performed Pearson's Chi-square (χ^2) tests and computed Cramér's V as an indicator for the association's magnitude. We will not report any association for variables where the expected cell count would be below five as it is the requirement for this analysis. For the association between the metric variable *Duration* and the remaining categorical variables, we performed Kruskal-Wallis H tests because the assumption of normality was violated. In the following, we will report on an excerpt of the associations we found in the analysis. For the complete overview, including the associations' strength, please refer to Appendix A, Section A.1.

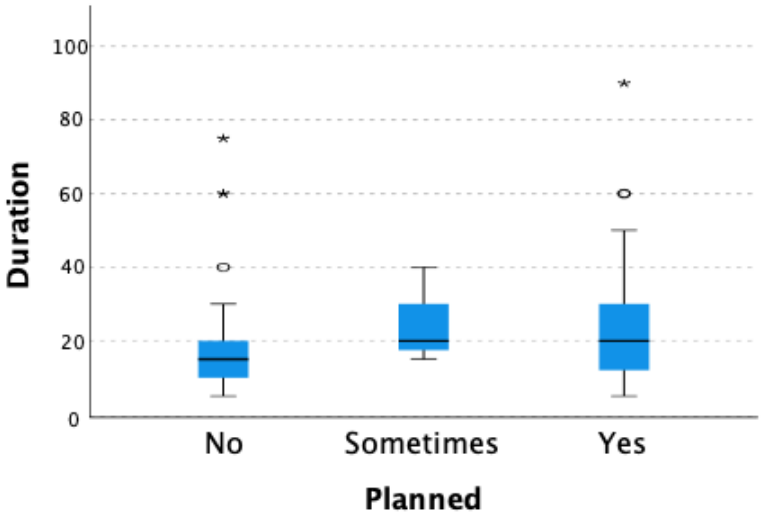


Figure 3.1: Distribution of the reported learning session duration (x-axis) according to the three levels of the variable *Planned*. An independent-sample Kruskal Wallis test revealed a significant difference in duration between planned and not planned sessions.

Our analysis revealed that the variable *Device* is associated with the learning activity *Duration* ($H(3) = 13.66$, $p < .01$), a pair-wise comparison revealed a significant longer learning duration for sessions reported to take place on Laptops/PCs compared to Smartphones (Bonferroni-corrected significance level of $p < .05$). The *Duration* of a session varies in regards to the *Planning* ($H(2) = 10.44$, $p < .01$). Figure 3.1 shows that planned learning sessions show significantly longer duration than unplanned sessions (Bonferroni-corrected significance level of $p < .01$). However, no direct association between Device type and planning can be found.

The variable *Device* is further moderately associated with the variables *Noise* ($\chi^2(21, N = 131) = 84.18$, $p < .001$, $V = .465$). Higher noise levels are mostly reported in learning sessions on smartphones. The *Frequency* with which the learning situation is reported to take place is associated with the factors *Planning* ($\chi^2(12, N = 131) = 23.98$, $p < .05$, $V = .304$) and *Stress* ($\chi^2(18, N = 131) = 39.68$, $p < .01$, $V = .319$). Looking at the cross tables, daily learning is often planned and predominantly reported together with low stress levels.

Furthermore, the *Time* of the reported sessions can be associated with the assigned *Stress* level ($\chi^2(15, N = 131) = 41.06$, $p < .001$, $V = .324$), *Setting* ($\chi^2(10, N = 131) = 34.44$, $p < .001$, $V = .364$), and *Noise* level ($\chi^2(9, N = 131) = 18.41$, $p < .05$, $V = .217$). While the usage situations taking place in the morning are often characterized in combination with a low to medium noise level, the situations in the afternoon and evening predominantly show a very low to low noise level (cf. Table 3.1). The association between *Time* and *Stress* level is similar to the association to the *Noise* level. While the situations in the afternoon are majorly reported to happen with low

Table 3.1: Exemplary cross-table for the Pearson χ^2 analysis between the variables *Time* and *Noise* level.

		Noise				**Total**
		Very Low	Low	Medium	High	
Time	Morning	2	16	17	4	39
	Noon	0	4	3	0	7
	Afternoon	5	13	6	3	27
	Evening	12	25	8	2	47
	Night	0	2	0	0	4
	Any	0	0	4	2	6
Total		21	60	38	11	130

to medium stress levels, learning in the morning is more frequently associated with a medium and sometimes even high stress levels. In the morning, the situation is furthermore often described as semi-public, while afternoon and evening sessions are mostly private.

Clusters of Usage Situations We clustered the situations based on (1) user's company (yes/no, strangers in public transport were not counted as company), (2) whether the learning session was planned ahead or not, (3) users' perceived stress level (low, medium, and high), and (4) the location itself, i.e., home, public transport, or otherwise indoors. Additionally, we differentiate if the location is public or private (whereas e.g., public at home could mean the participant shares home with other people). This clustering resulted in five common learning situation descriptions, marked numerically from 1 to 5 in Figure 3.2, covering 82% percent of all situations described in the survey. These situations can be characterized as follows:

Situation #1 In this cluster of 11 usage situations, the user mainly uses a smartphone as a device to learn with. The session mostly occurs spontaneously in the afternoon, evening, and night at least several times a week, if not daily. A learning session takes on average around 15 minutes and happens in a low to medium stress environment at home. However, learners can be in the company of family or spouses.

Situation #2 In this set of 22 pre-planned learning sessions, users mix between different devices such as smartphones, tablets, or laptops. Learning takes place mostly in the second half of the day and less often in the morning. The situation's frequency ranges from daily to weekly, lasting around 30 minutes on average. The sessions take place in a calm environment, at home, without company, and with low stress level.

			Home		Public Transport		Indoors	
			Public	Private	Public	Private	Public	Private
with company	Planned	Low Stress	0	1	0	0	0	0
		Mid Stress	0	0	1	0	3 (5)	0
		High Stress	0	0	0	0	0	0
	Not Planned	Low Stress	1	9 (1)	0	0	0	0
		Mid Stress	0	2	0	0	0	0
		High Stress	0	0	0	0	0	1
without company	Planned	Low Stress	1	22 (2)	4	2	0	1
		Mid Stress	0	1	8	0	0	0
		High Stress	0	0	0 (4)	0	0	1
	Not Planned	Low Stress	1	31 (3)	9	3	1	0
		Mid Stress	0	5	12	1	1	1
		High Stress	0	0	3	0	0	0

Figure 3.2: The five clusters of situations according to the dimensions location, company, planning, stress level, and setting.

Situation #3 This cluster contains situations in which users learn almost exclusively on their smartphone and is the biggest cluster with 36 reported situations. Learning does not specifically happen at a certain time a day but any time during the day, with a daily to several-times-a-week repetition frequency. The average duration of such a session is around 17 minutes. In contrast to Situation #2, the learners are not planning the sessions to happen, but they take place spontaneously at home with a low or medium stress level.

Situation #4 This broader set of 32 situations contain learning sessions where users exclusively learn with a smartphone in the morning and less frequently in the afternoon. A rather noisy environment characterizes these sessions as learning happens in a public environment, more specifically on transportation vehicles. Users report that during their learning company is present. This can refer to friends or spouses but also to strangers on their bus or train ride. The situations are sometimes planned (e.g., on regular commutes) but can also occur spontaneously, daily to several times a week, for on average around 15 minutes. In this cluster, the stress level can vary from low to high stress.

Situation #5 This final rather small cluster summarizes three learning situations in an indoor yet public setting, such as university, library, or school. Users visit these places between morning and afternoon with the intent to use the learning applications sometimes on their laptops, other times with smartphones. They never have company aside from unknown bystanders, and they estimate these situations to have a medium stress level. Depending on the user, this situation spans over a daily to weekly usage, and takes around 20 minutes.

3.2.4 Discussion

The three different analysis approaches (descriptors, clusters, and facet associations) reported in this section generated interesting insights into users' perspectives on their common usage situations. While the first part outlined the descriptive values of occurrence of the situations' characteristics independently, the second part investigated potential associations between each variables. Lastly, the third part applies a clustering approach to derive patterns of common usage situations. In the following, we aim to synthesize and discuss the insights from the three evaluation parts.

Diversity Can Come From More Than Locations The majority of the reported common and frequent usage situations occur at similar locations, particularly at home and during public transportation. However, the situation descriptions vary greatly even within the same location. Three clusters of usage situations (#1, #2, and #3) contain learning situations that reportedly take place at home. Yet, cluster #2 contains pre-planned learning sessions that last around 30 minutes in a calm and quiet environment. In contrast, cluster #1 includes sessions that last only half as long, can be in the company of family or spouses, and take place spontaneously. Some learners add that for situations taking place at home, they usually learn from their bed, others from their dining table while eating, again others from the sofa while watching TV. While the location is often the predominant descriptor when we think of contextual factors [296], in particular, for learning with mobile devices, this survey reveals confirms that many other factors define a learning situation in an everyday environment. Most importantly, mobile learning should not per se be defined as learning on the go or learning in public settings, as people's home can already create a great diversity in usage situations.

Mobile Learning Does Not Equal Mobile Learning One of the striking differences we observed in the reported usage situations is that there is still a difference in how users learn with smartphones vs. laptops. Both devices are mobile and can be set up nearly anywhere. Yet, the way learners interact with them differs. While the smartphone is still the predominant device in 98 of the 131 collected common mobile learning situations, learning with laptops or PCs is still common (24 reported situations). While a PC can not be considered a "mobile" learning experience, we did not exclude these ten reported situations. Mobile learning applications such as Duolingo allow for usage access from multiple devices and automatically synchronize their progress. We saw that learning on the smartphone is related to increased noise and shorter learning session duration. These findings indicate that there might be a tendency toward learning with smartphones in less controlled environments.

Distractors Are Ubiquitous In the 131 characterized situations, potentially distracting characteristics are not rare. For example, increased medium up to very high noise levels (70 out of 131), increased stress levels (42), uncontrollable situations such as

train rides or waiting rooms (57), or other activities that require multitasking (29) are prevalent in our data set. Especially the indication of learners having the habit of consuming media content such as TV or music during learning automatically diverts some of their attention away from the learning tasks. However, as we established that the perception of these factors is subjective and dependent on the individual learner, future work needs to investigate in detail what can cause interruptions and distractions during mobile learning usage.

Limitations

While this chapter provides first insights into the use of mobile learning applications in everyday settings, the results have to be interpreted with caution. Due to the task's limited scope of asking users to report their common usage situations, the results present a subjective viewpoint on the individual learning behavior. By asking users to report their most frequent usage situations, we explicitly limited the diversity of everyday usage behavior.

As mentioned before, the descriptions of the environment, such as noise, are observer relative and might be viewed differently amongst learners. Similarly, the estimation of learning time and frequency might deviate from the actual learning frequency and be prone to biases such as the recency effect (i.e., the participant attributes greater importance to recent learning situation) or the social desirability bias (i.e., the participant reports situations they think the examiner wants to hear). While we tried to counteract those biases through detailed task descriptions and ensuring anonymity, they can never be fully avoided.

Nonetheless, the subjective view on the frequent usage situations revealed interesting insights. For example, while the actual learning situation might not be noisy (as a sensor would detect it), the user might perceive even a quiet conversation as distracting noise. Accordingly, the subjective view tells us more about the actual situation than an automated detection could. Exemplary, we discovered that users sometimes label learning situations happening during commutes on public transportation as a "private" setting. Similarly, learning while having family, friends, or spouses around does not necessarily disturb learners or make them think of a situation as "public".

Further, the Chi-square analysis provides indications of associations. Yet, the expressiveness of this evaluation has to be seen with limitations. Small sample sizes that can occur in the combination of individual facets might decrease the validity of the conclusions that can be drawn from the analysis. We decided not to cluster individual descriptors too much but rather let the data set reflect the diversity in learning situations, despite some of them having low occurrence frequencies (e.g., learning in a car or learning at night). Therefore, we had to exclude many associations due to violations of the analysis requirements.

3.3 Exploring Design Opportunities for HCI

We conducted a follow-up focus group to transform the knowledge generated from the online survey into HCI insights. This focus group aims to come up with initial design recommendations for mobile learning apps that adapt to the diversity of everyday learning situations as the survey participants have described them.

3.3.1 Sample and Procedure

Four HCI experts, with expertise in, but not limited to, mobile learning, interface design, security and privacy, and decision support, participated. In addition, three of our experts had prior experience in using mobile learning applications in their personal life. After outlining the topic, the discussion was guided by the following four questions/tasks.

1. Which internal and external factors influence the user when learning languages on a mobile phone and why?
2. Categorize these influencing factors according to whether they are external and/or internal.
3. Each pick the one where you expect the highest benefit on learning success and one with the highest benefit for user engagement.
4. Think of ways the (design for the) application could adapt to changes occurring based on the factors derived in step 2 and 3.

3.3.2 Results

The first open question revealed various factors our experts consider potentially influencing learning on a mobile device. The participants of the focus group clustered those in the second step into 13 broad categories they labeled themselves, of which five referred to internal processes (*mood*, *motivation*, *boredom*, *curiosity*, *cognitive load*), and seven target external or situational factors (*weather*, *social*, *interruptions and distractions*, *location*, *hardware*, *comfort*, *privacy*, *necessity*). For example, necessity could mean that a person has to learn to speak English fluently because it is required for their job. In the category *social*, our participants listed reasons such as peers using the same learning app or a "busy work schedule" reducing the frequency of learning activities. *Location* contains examples of learning in public environments such as on public transportation. These situations could demand external attention and holding onto a handle in a moving vehicle allows you only to use one hand. To dive deeper into the actual design of learning applications that target these facets, we asked our participants to mark two categories with a sticker. On the one hand, they should pick

the category where they expect that designing for the optimization of this facet would benefit the learning success. On the other hand, the second category should be picked to target for optimized user engagement.

Our participants agreed on *cognitive load* as the category with the potential to maximize the learning outcome. Cognitive Load is not only influenced by the learning content but also by the learning instruction and external factors such as divided attention due to distractions. In contrast, they expect *motivation* to have the highest benefit on users' engagement. According to our focus group participants, optimizing for the learners' individual motivation would support frequent engagement and perseverance over time.

When discussing how to design for the factors *cognitive load* and *motivation*, the majority of the suggestions revolved around the adaptation of the learning content and UI. For example, to counteract the adverse effects of cognitive overload, the application could present exercise types that require recognition instead of recall (i.e., multiple-choice instead of free text entry). Similarly, it would be desirable for listening tasks if the application could detect if the learner can comprehend the spoken text. In case of difficulties, the speaker's speed or the voice should be adapted automatically to not overwhelm the user. Further, during low attention levels or cognitive overload, the amount of new content introduced should be kept to the absolute minimum. Our participants suggest to instead focus on repetition in these situations. They further recommend de-cluttering the interface and include visual cues as attention-grabbing tools to keep the learners' focus on the task. To keep up the motivation of the learner, the app could track what exercise types the users prefer and include gamification features to make learning more fun. Moreover, ending the learning activity while the learner is in a positive mood was one of the suggestions by our focus group.

3.3.3 Limitations

As our focus group participants were HCI experts, their recommendations are limited to general UCD. While these recommendations provide first ideas on potential support mechanisms for learners in everyday settings, future work needs to evaluate more carefully which suggestions are feasible for implementation. In particular, reviewing related literature from instructional design, cognitive and educational psychology, and

3.4 Implications for Design

Based on the online survey and focus group results, we derived a set of design recommendations. This list is not meant to be exhaustive but is derived to highlight currently underrepresented but promising aspects in frequently occurring usage situations.

Design Recommendation 1: Sustaining Focused Attention by Managing Interruptions

In general, the application should make use of levels of high attention and focus. High concentration increases the chance of information being stored in the LTM [294]. In many environments, such as public transport, split attention is inevitable. Bulling [48] recognized focusing on managing user attention by turning continuous partial attention into sustained attention as an important challenge [48, 284]. By displaying content inferring high user attention (such as interactive tasks [120]) and managing potential interruptions, we could support the user in sustaining high attention levels. As a design recommendation, we propose introducing 'attention grabbers' or visual cues in a public and busy environment if the attention gets drawn away from the learning task. Such cues can redirect the attention back to the screen and have proven to restore the context of the primary task, i.e., learning [159].

Participants suggested individualizing this process during the focus group and investigating changes in users' attention levels regarding their current situation. Exercise types or topics could be perceived differently between people. When looking at our survey's most common usage situations, high attention is often occurring with the absence of distractions and interruptions, which are most likely to take place in private environments. In 24 situations, users stated to be at home with a low stress level. We recommend targeting these situations to present information that needs to be stored in the LTM. Moreover, future work needs to investigate the occurrence and management of interruptions in more detail for the specific use case of mobile learning.

Design Recommendation 2: Turning Distracting Activities into Learning Opportunities The focus group discussion confirmed the importance of user motivation in implementing MLL features, as already pointed out by literature (see related work). It is further essential not to overwhelm the user with tasks that require their active input when their attention declines since a positive mood has a reportedly positive effect on learning [46]. We suggest keeping up the user motivation by piggy-backing learning activities on other activities users perform anyway. This goes in line with the lifelong learning framework by Fischer [108], who emphasizes that "*learning should be embedded in the pursuit of intrinsically rewarding activities*" (Fischer [108], page 9). For example, learners reported to often listen to music or watching TV during learning at home. Here, users choose the content they consume according to their individual interests and preferences. Suggesting to consume content in the foreign language a user wants to learn and providing interactive support has great potential to increase motivation, learning duration, and frequency.

However, especially for spoken text, our focus group participants emphasized the need for monitoring users' comprehension. If the task is too difficult, either because of the choice of words or the voice of the speaker (too fast, difficult to comprehend), the task needs to be adjusted.

Design Recommendations 3: Short and Simple when Necessary, Long and Intense when Possible Controlled experiences have already shown that interruptions have a highly disruptive effect on task performance, error rate, and affective state [22]. When looking at the situations described in the survey, environments that often demand the user's attention can be found during high noise levels or high stress levels. In general, public environments demand more attention than private, as long as users reduce parallel activities to a minimum. When the users are already facing a high cognitive load induced by their environment, the focus group participants suggested reducing the amount of new content. If the users' focus is not solely on the application, the interface should be less cluttered and contain a clear structure. Simple repetition tasks such as multiple-choice vocabulary translations can be easily performed in such environments and still help consolidate vocabulary knowledge. In contrast, when the user decides to invest more time into learning, the application should encourage users' dedication. Planning learning sessions in advance is a sign of self-regulated behaviour, which is essential for successful mobile learning [341]. Therefore, quiet and interruption-free environments can be optimally used to engage in learning more complex content such as grammar knowledge.

3.5 Chapter Summary

This chapter presented an exploratory investigation of common usage situations of mobile learning applications in everyday settings. By performing an online survey, we collected detailed descriptions of users' reported mobile learning usage situations. Regarding our research question **RQ1a**, we contribute an in-depth description of the individual characteristics of the reported situation. We outline frequently used descriptors for facets such as location, company, or parallel activities, analyze the associations among the facets, and derive five clusters of common usage situations. Our results show that learners use mobile learning applications in a diverse manner and that the descriptions of usage situations show a great diversity. With many situations taking place at home, the reported diversity surpasses the narrow-minded idea of location being the determining factor of usage context. In particular, learning that happens at home is not always characterized by quiet and calm environments but can be influenced by family members, stress, or device type.

Combined with the input from a focus group that had the aim to derive design opportunities for HCI, we outline three main design implications: (1) *Sustaining Focused Attention by Managing Interruptions*, (2) *Turning Distracting Activities into Learning Opportunities*, and (3) *Short and Simple when Necessary, Long and Intense when Possible*, that we consider relevant for supporting mobile learning in everyday settings. We consider particularly implication (2) a great opportunity for embedded mobile learning, which simultaneously can improve the personalization of content to users interests. We will further explore this idea in Chapter 6. However, future work is needed to investi-

gate how the situations gather in this chapter can be generalized into user behaviour in the wild. We need to particularly explore, how frequently interruptions occur during everyday learning activities and what their effects on the user are. The following chapter will take a closer look and evaluate interruptions in the wild.

4

Exploring Interruptions during Mobile Learning in Everyday Settings

Chapter 3 reported common mobile learning usage situations and outlined the diversity characterizing learning in everyday settings. Not only do these situations show that people learn wherever they are (i.e., at home, on commutes, or in the library), they also frequently multitask. For example, participants stated watching television or listening to music while learning on their mobile devices. Learning in informal and uncontrolled settings comes with the inherent risk of being distracted or interrupted, inhibiting proper memory encoding and storage in the LTM [294]. Further, interruptions increase task completion time and error rates [22]. To better mitigate these adverse effects of interruptions while learning in everyday environments, we must first examine them in more detail. Particularly, this chapter aims to answer the following research question:

RQ1b: How do interruptions affect mobile learning in everyday settings?

We explore the occurrence of interruptions in everyday mobile learning by applying the experience sampling questionnaires. We implement an Android application that logs all mobile learning sessions and requests the user to enter additional data on the context and interruptions at the end of each session. By inquiring about the users' learning context and their reasons for ending the learning session, we aim to better understand the impact of interruptions on the quality of learning in everyday settings. Further, based on the acquired data, we derive strategies to deal with interruptions in mobile learning and mitigate their adverse effects.

This chapter is based on the following publication:

- Draxler, F.*, Schneegass, C.*, Safranek, J., and Hußmann, H. (2021). Why did you stop?- Investigating Origins and Effects of Interruptions during Mobile Language Learning. Accepted for publication at Mensch & Computer (MuC'21), September 5–8, 2021, Ingolstadt, Germany. ACM, New York, NY, USA, 15 pages.

---

* Both authors contributed equally to this research

Some text passages were taken verbatim from this publication. Further, this section is supported by the Master thesis of Jonas Safranek (see detailed collaboration statement at the beginning of this thesis).

4.1 Related Work

In today's world, multitasking is a common practice. The ubiquitous availability of technology makes it an enormous source of distractions and interruptions. In using a mobile application, an interruption can be defined as an event or action that leads the user to shift the focus away from the application. The term interruption is not specified in regards to duration. For now, we will consider any learning break, no matter if it only lasts a few minutes or several days, as a suspension of the learning task. We will revisit this distinction at a later point in time.

Even though some sources claim the positive effects of multitasking on cost-switching and attentional skills (e.g., [112]), the real-life implications of interruptions on task performance (in this case, the learning task) are predominantly negative [57, 352]. This section will explain the process of interruptions and outline their characteristics.

4.1.1 Phases of Interruptions

The sequence of an interruption is depicted in Figure 4.1. The process starts with the task that is being interrupted – often referred to as the *primary task* –, the interrupting event or task – also called *secondary task* –, and ends with the return to the primary task [329]. The *interruption lag* is the time between the interrupting trigger or stimulus and the shift of attention towards the activity related to the source of the interruption. This phase has the potential for users to prepare for an upcoming interruption. The *resumption lag*, on the other hand, is the time needed to return one's focus on the primary task after moving away from the secondary task [329].

In real-world scenarios, the distinction between primary and secondary tasks is often challenging because series of tasks can be interleaved and convoluted. Nevertheless, we will use this differentiation to explain our use case, as the underlying principles remain the same. Also, secondary tasks can have a different duration and are not all equally demanding. For example, learners can be interrupted because they have to switch trains on a commute. Such an interruption can take several minutes and demand 100% of the users' attention. In a different situation, the interruption occurs because the phone notifies a learner about an upcoming appointment in 30 minutes. If the learner is aware of this appointment, the notification will not attract much attention and the primary task can be resumed after just seconds. We will describe the characteristics of interruptions and their impact on task performance in greater detail below.

The Memory-for-Goals theory [5] views interruptions as a suspension of the primary task's goal. It states that the user can only retrieve a suspended goal with the help of priming through a Memory Cue. This cue acts as a trigger to recall previously stored information from users' LTM [5]. Moreover, Trafton et al. [329] differentiate between two directions to encode a goal, namely: (1) retrospective ("What was I doing before?")

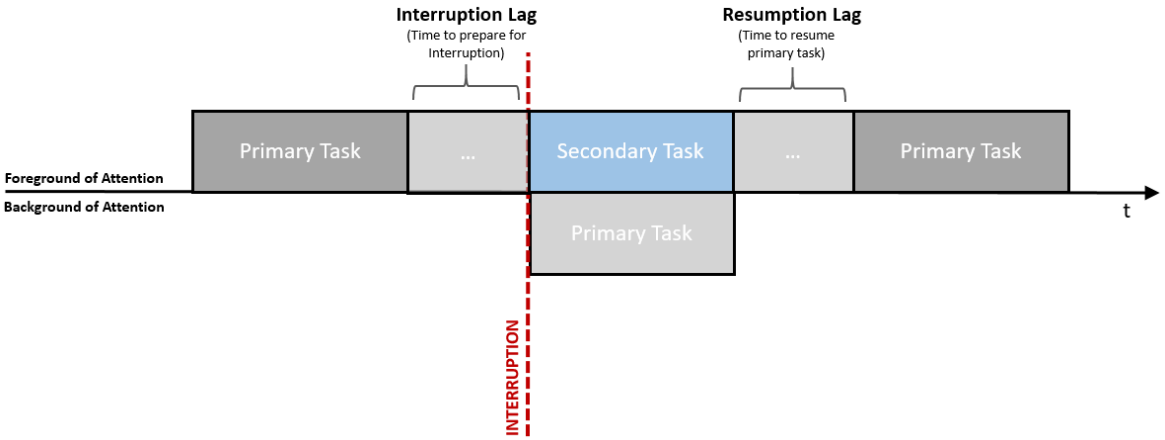


Figure 4.1: A secondary task interrupting a primary task. The time preceding an interruption is called *Interruption Lag*, whereas the time following an interruption is called *Resumption Lag* (terminology cf. [329]); Figure from [300])

and (2) prospective ("What was I about to do?"). Prospective memory cues can be primed in the interruption lag, for example, by taking notes of what you were about to do if the interruption would not have happened and are recalled in the resumption lag. Retrospective cues can still be applied without the need of former priming as any priorly perceived information can function as cue.

4.1.2 Characteristics and Effects of Interruptions

Interruptions can differ among each other in how much they disrupt the primary task. While some go unnoticed, others can severely disrupt task performance and duration [22, 171, 186] and decrease the quality of memory encoding. The disruptiveness further depends on other properties such as source, duration, urgency, anticipation, and how the type of secondary task relates to the primary task in terms of modality and complexity [39, 300]. We will explain these facets in more detail below.

Interruptions can have their **source** in the user (i.e., *self-interruptions*), the device (i.e., *device-internal interruptions*), or the environment (i.e., *external interruptions*) [171, 238]. Self-interruptions are self-initiated [3], for example, caused by boredom, hunger, or mind-wandering. Device-internal interruptions refer to any stimulus originating in the device, for example, calls or app notifications, and external interruptions are caused by any stimulus from the environment (e.g., noise, light, sounds). A study by Katidioti et al. [171] showed that the source of an interruption affects its severity. The task completion time was lower for external than self-interruptions.

In contrast to self-interruptions, external interruptions can rarely be foreseen in advance. **Anticipation** refers to whether an interruption is planned and predictable. If

an interruption is not anticipated, it is more likely to cause stress and affect the performance [69]. For example, while an alarm clock is a planned interruption, a phone call might not.

Similarly, some interruptions can be delayed, while others require immediate attendance. **Urgency** is an important factor when deciding if an interruption can be delayed or even ignored [144].

They further vary in their **duration** - while some are short distinct interruptions, others require the user to take longer breaks. Altmann and Trafton [5]'s memory-for-goals theory states that the presence of a goal fades in memory over time. Immediately after being interrupted, a person can still recall the goal state. The longer the interruption continues, the more investment is needed to recall the goal [147, 239].

Moreover, the severity of an interruption can depend on what task is being interrupted. Their characteristics, particularly in regard to the interruption, can influence how we perceive the interruption. The **modality**, defined as "*a medium of sensation, such as vision or hearing*" [13], refers to senses we use to perceive it. A study by Latorella [204] showed that the adverse effects of auditory interruptions are greater than for visual ones. Particularly, the *similarity* the task and interruptions can influence how easy we can transition between them [7]. For example, two tasks of the same modality can create a conflict of working memory resources. A study by Ledoux and Gordon [207] showed that interrupting a reading task with a visual stimulus decreased comprehension rates. The creation of associative connections during the task can counteract the effects caused by the interruption [102].

Lastly, the **complexity** of the interrupting task can impact its severity. The more demanding the interruption, the longer it takes to resume the primary task [147, 239]. Contrary to those findings, a study by Speier et al. [315] suggests that, in particular, simple interruptions can impact users' arousal and stress level, thereby actually improving the overall performance.

4.2 Implementation for Detecting and Classifying Interruptions

To find out more about interruptions in everyday mobile learning settings, we developed a custom Learning Activity and Interruption Recognition Application (LAIRA). It runs in the background and gathers data on the interruptions that occur during learning. is designed to record all learning sessions of users with their learning app of choice, logging learning sessions, potentially interrupting device events, and issues Experience Sampling Questionnaires (ESQs) after each learning session. LAIRA was developed for Android 7 or higher. Following the recommendations of Ciravegna [66], we implemented a combination of *BroadcastReceivers* and *JobServices* to keep the app

running at all times despite various battery optimisation techniques applied by different device vendors.

After the successful installation of LAIRA, the users manually select the apps they use for learning by selecting them from a drop-down list of all apps on their phone (see Figure 4.2b drop-down menu "Select Learning Application"). By default, the app is in an idle state, in which sensors and receivers are deactivated to reduce battery demand. The only task LAIRA performs is to check app package names to monitor if the user starts one of the predefined learning applications. If such an event is detected, the app switches to its active state and begins logging the activity, interruptions, and context data (for details on the logging see Subsection 4.2.2). To retrieve context data, the open-source AWARE Framework is applied[1]. The data is stored organized as sessions in a Google Firestore Database[2]. The NoSQL-based database is further used to store the ESQ results, mapping them to the user through a unique automatically generated identifier, following LMU's anonymization policies.

4.2.1 App Interface

Interaction with LAIRA is limited to a study dashboard, a timeline view of recorded sessions, and an integrated survey view (see Figure 4.2b). In the study dashboard, users select a learning app from a drop-down list of all apps on their phone. This is necessary for tracking learning sessions (see Section 4.2.2 for more details). In addition, the dashboard gives an overview of permissions granted to LAIRA (coloured green in the figure). The timeline view is intended to increase transparency and trust by showing users what data are recorded, as the required permissions could potentially be used for privacy-invading purposes. Finally, the survey view shows an initial and final survey at the beginning and end of the study period, respectively. Due to this coupling, the survey results and app interaction data are connected without compromising the anonymity of participants.

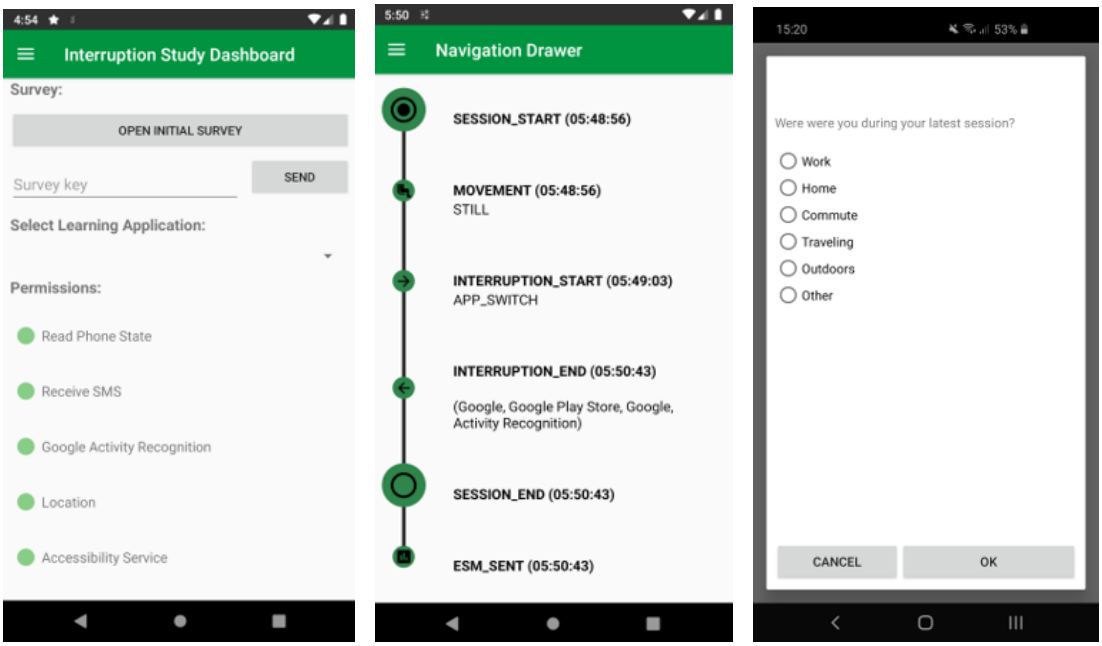
4.2.2 Event Logging

Event logging starts once users have selected a learning app in the study dashboard. Every learning session LAIRA recognizes is saved to the database, including relevant general meta-data such as the current app name, a time stamp, session or event length, session ID, and user ID. The *learning session* events include the following information:

- **Learning App** - Indicating the name of the learning application in use.

[1] AWARE Framework: `https://awareframework.com/`, last accessed January 3, 2022

[2] Google Firestore Database: `https://firebase.google.com/docs/firestore`, last accessed January 3, 2022



(a) (b) (c)

Figure 4.2: **(a):** The main view of LAIRA showing the link to the initial survey, the survey user key input field, the drop-down list to select a learning application to track, and the permissions users have granted to the app. **(b):** The timeline view of all recorded user data. **(c):** An exemplary excerpt of the ESQ prompted after each learning session.

- **Session Duration** - Measuring the time from the session start to the end of the session. We define the end of a learning session either as an active closing action by the user or when the time threshold of inactivity is reached (learning app moved to the background or screen turned off). This time frame in LAIRA is ten minutes and reflects the short learning sessions common in ML. When the user returns to the learning activity after long periods of inactivity, we consider this as the start of a new learning session.

We gather additional information about events that can cause *interruptions*. If multiple interrupting events occur during a single learning session, each interruption is registered as an individual event. Only the last interruption will be inquired upon in the ESQ to keep the effort for the user to a minimum.

- **Notifications & Communication** - This includes push notifications, SMS, and phone calls. We do not process or store any text or voice content but only application package names and metadata. To catch SMS and phone calls, we register a *BroadcastReceiver* and set actions for *android.intent.action.PHONE_STATE* (in particular *RINGING*) and *android.provider.Telephony.SMS_RECEIVED*. To

collect data on incoming push notifications, we implement a *NotificationListenerService*[3] and store the app name, the notification priority, and check if it caused a sound or vibration. The priority (as well as the sound and vibration on Android 7.1 and lower) can influence whether or not the notification is displayed as a heads-up notification[4], which is more likely to distract users than less intrusive notification types.

- **Application Switches** - Switches of applications on the phone can occur for different reasons. If the user switches from the learning app to a different app without prior indication (i.e., a notification), we label the switch as an internal interruption. In this case, we assume that the user decided to start another activity on their own accord. For example, users might want to quickly put an item on a digital shopping list. If the user switches apps due to an SMS, call, or notification, we consider the interruption event triggered by the device.
- **Screen Locks** - Every time the user actively locks their screen or the screen is locked by the phone moving to an idle state, a screen lock event is recorded. The screen lock could indicate an external interruption that cannot be tracked or that the user ended the learning session. We record these events as ambiguous interruptions. Their cause has to be confirmed by the user in the ESQ, otherwise, the label "ambiguous" remains.

All of the events listed above can cause the LAIRA app to register a "Session_End" event. Based on the flow depicted in Figure 4.3, the app categorizes the interruption types and triggers the ESQ. Further, the following *context information* is acquired:

- **Movement type** - We record movement types obtained from the Google Activity Recognition API [5], i.e., *IN_VEHICLE*, ON_BICYCLE, ON_FOOT, RUNNING, *WALKING*, *STILL*, or *UNKNOWN* . To assure sufficient data quality, we only include movements with a confidence level of 90 or greater. A movement type is considered active until a movement type switch occurs.

4.2.3 Experience Sampling

We further collected self-reported data on interruption strength and validity through the Experience sampling Method (ESM) [336]. To create the Experience Sampling

[3] Notification Listener Service: `https://developer.android.com/reference/android/service/notification/NotificationListenerService`, last accessed January 3, 2022

[4] A floating window that is shown at the top of the screen for a short moment when the device is unlocked, `https://developer.android.com/guide/topics/ui/notifiers/notifications`, last accessed January 3, 2022

[5] Google Activity Recognition API: `https://developers.google.com/location-context/activity-recognition`, last accessed April 15th, 2021

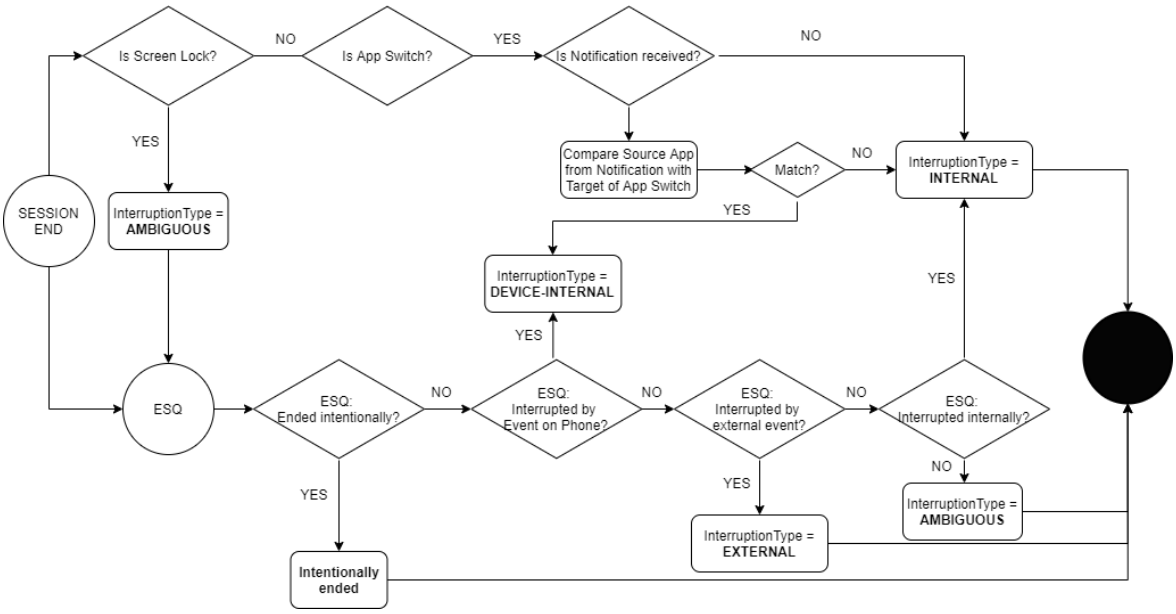


Figure 4.3: This flow chart depicts the LAIRA app's process of categorizing interruptions that terminated the learning session once the Session_End event has been registered and the ESQ answered, dismissed, or expired.

Questionnaire (ESQ) prompts, we used the *ESMFactory* class provided with the AWARE framework. LAIRA triggers new ESQs via a push notification ten minutes after the last interaction with a learning app was recorded. The ESQs only comprise multiple-choice questions that allow for quick completion; some of the questions are dynamically adapted based on recorded events. Clicking on the ESQ notification opens a pop-up dialogue with the following questions on the learning session. The questions used for determining the interruption type are also displayed in Figure 4.3.

- **Where were you during your latest session?** Options: Work | Home | Commute | Travelling | Outdoors | Others
- **Were you alone or in company during your latest session?** Options: Alone | With one other person | With more than one other person
- **Please confirm the movement type we detected.** (see Google Activity Recognition API)
- If the user received at least one notification during the recorded learning session, we further ask for confirmation of distraction **Did you receive any distracting notifications during your latest session?** Options: Yes | No
- **Why did you end your learning session?** Options: *Device Internal* – "I was interrupted by something on my phone (e.g., a notification, call, SMS, email, etc.)" | *External* - "I was distracted by something external to myself or the

phone (e.g., doorbell, other people, having to get off of train, etc.)" | *Internal* - "I was distracted internally (e.g., tiredness, could not concentrate, thinking of something else, mind-wandering, etc.)" | *Intentional* – "I was done using the app."

- If the user did not answer with "intentional", a follow-up question is shown: **How important was it that you follow up upon the interruption?** Options: "Very important- it was urgent / time-critical" | "Moderate - I had to do it eventually in the near future" | "Not important - I could have ignored it and continued learning"

In the AWARE ESMFactory, users can always dismiss ESQ notifications and they do not have to answer right away. If the user does not fill in the ESQ within three hours after a learning session, it is discarded to avoid bias due to fading memory of the learning session (cf. [336]). We chose this time window in line with findings from prior work indicating that mobile learning apps are rarely used multiple times per day but rather daily or less than daily [305]. Therefore, we aim for high ESQ participation and expect participants to remember the context of their learning session within a three hour window. Additionally, as van Berkel et al. [336] recommend, a previously unanswered ESQ is deleted if the user finishes a new learning session before answering the ESQ of the previous session.

4.3 Field User Study

4.3.1 Study Design

We assessed the mobile learning habits of users with a specific focus on exploring interruptions and their effects during learning[6]. We evaluated occurring interruptions caused by the *device* (e.g., notifications, calls), *external* circumstances (e.g., noise, distractions), or *internal* reasons (e.g., getting tired, mind-wandering). The quantitative data was augmented with responses to ESQs. We try to answer two general research questions:

1. How often and in what contexts do interruptions of the three types (internal, external, device) occur during ML?
2. Does the interruption type, length, or count influence the learning session and the risk of termination?

[6] In this thesis, we will not report on the analysis of mobile learning behaviour performed in this study (cf. Contribution Statement). For more details on the analysis see [96].

4.3.2 Procedure

After registering for our study, participants received an email with detailed instructions for installing the LAIRA app and the study procedure. Furthermore, we provided participants with information on LMU's data protection regulation and asked them to read and sign the consent form. To allow for remote execution of our study, users could start the initial questionnaire from within the application upon giving informed consent. Immediately after the successful installation of the app, the user was prompted with the link to the first questionnaire asking for demographic information and previous experience with mobile learning applications. The questionnaire responses and app logging data were linked using an individual, automatically generated user token.

The participants were encouraged to use their learning application of choice and follow their usual learning routine. After each learning session, we asked them to fill in a brief ESQ to gather additional information on the learning session in particularly the reasons for ending the sessions and potential interruptions. At the end of the four-week study, the application informed the participants about reaching the study end and prompted them with the link to the final survey from within the application. LAIRA does not process notification content or caller names; it only stores the respective app package name. Nonetheless, due to the processing of semi-personal data (i.e., movement types, apps used), we acquired approval from LMU's ethics committee[7] to perform this study. All people received a 25 Euro Voucher for an online store or an equivalent amount of study credit points as compensation for successful participation.

4.3.3 Sample

We recruited 12 participants using our university's mailing lists and social media channels. One participants did not hand in the final questionnaire but used the app correctly over the full course of the study. We include this participant's data in the description of the logging but report the questionnaire results only for the subset of eleven.

The participants' age ranged from 19 to 60 ($M = 27.84$, $SD = 10.43$) years and they all identified as female. Four stated that their highest degree was a high school diploma, three had a bachelor's degree, four a master's degree, one person reported a lower-than-high–school degree. Five participants were full-time students, one was working full-time, four studying and working part-time, and two were unemployed at the time of the study[8].

[7] Ethical approval granted: https://www.mathematik-informatik-statistik.uni-muenchen.de/ethikkommission/index.html; case number EK-MIS-2020-019

[8] As this study ran during the COVID-19 pandemic, we further asked the participants to specify their working situation. All participants who were currently working or studying stated to be able to do so from home. We will discuss the implications of the situation on our study results in the Limitation Section.

To assess the participants' smartphone usage behavior, we asked them to specify their average smartphone, social media, and messaging habits (if they were unsure, we advised them to use their phone's digital well-being feature showing an overview of screen time by app category). In particular, they could select time ranges from 0-15, 15-30, 31-60, 61-120, 121-180, and 180+ minutes per day. For general smartphone usage, the majority of participants (11) selected phone usage times between 60 and 181+ minutes per day. When reporting on social media usage in particular, the answers ranged from 0-15 min (2) to 61-120 min (4), the majority (6) being in between. For messaging applications, most of the participants selected the ranges 16-30 min (5) and 31-60 min (4). The remaining three participants reported up to 120 min daily usage. The participants also stated that they received a mean of 133.58 push notifications per day ($SD = 127.50$, estimated or looked up via the digital well-being feature), with a maximum of 400 and a minimum of three.

All our study participants reported to have prior experience with mobile learning applications. Two participants were currently using the apps extensively, five currently but only occasionally. The remaining participants had used mobile learning apps in the past, four extensively and one rather sporadically.

4.3.4 Results

Our evaluation provides complementary insights from qualitative and quantitative data. For our research question 1, we report on descriptive statistics, while for question 2 we apply hypothesis testing.

There were large differences between the learning habits of individual participants and the number or learning sessions per participant. Therefore, for hypothesis testing, we report Bayesian Analysis of Variances (ANOVAs) and post-hoc tests where we control for the individual participants' as random effects[9] The Bayesian tests additionally allow us to draw statistical conclusions even on small sample sizes (cf. [172]). These measures were computed with JASP [344]. We further apply a Generalized Additive Model for Location Scale and Shape (GAMLSS) that predicts our response variables based on context data. This approach is similar to linear mixed models but allows for modeling based on skewed distributions [316]. Again, we integrate the participants as random effect. The models were computed with R [272] and the GAMLSS package [279].

[9] Even for tests with two conditions only, we used Bayesian ANOVA instead of t-tests in order to control for random effects. The reported Bayes factors BF_{10} indicate the likelihood ratio of the alternative hypothesis H_1 (i.e., a difference between groups) and the null hypothesis H_0 (i.e., no difference between groups) [344]. For example, a Bayes factor of 3 would be interpreted as moderate evidence in favor of the alternative hypothesis and 40 would indicate very strong evidence.

Characteristics of Learning Sessions

In total, we recorded 328 learning sessions with LAIRA, the majority using language learning apps such as Duolingo (218 learning sessions), Babbel (39), Drops (39), or Memrise (10). Further, 22 sessions were recorded on the learning app Quizlet, with which the user can design flashcards for any learning topic. All apps were rated as enjoyable ($M = 4.09$, $SD = 0.51$) and participants reported a good User Experience (UX) with their learning apps ($M = 4.09$, $SD = 0.79$) and the ESQ application ($M = 3.82$, $SD = 0.57$). Overall, they were satisfied with their personal learning progress over the course of this study ($M = 3.73$, $SD = 0.75$).

The duration of learning sessions ranged from 27 seconds to 3223 seconds ($M = 671.4\,\text{s}$, $SD = 577.0\,\text{s}$). Participants supplemented 266 of the remaining 327 learning sessions with additional data through the ESQs. In the remaining 61 cases, the ESQs were either dismissed or removed after not being completed in the three-hour time window after the learning session. It has to be noted that not all ESQs were submitted fully answered as no question was mandatory. Thus, we aimed to increase participants' willingness to state at least some extra information. Our report below includes all the available data.

Characteristics of Interruptions

We differentiate two types of interruptions: (1) interruptions that terminated the learning session and (2) interruptions that only led to a temporary suspension of the learning app. In the latter case, participants returned to the learning app within 10 minutes (see Section 4.2.2). We first report the characteristics of the suspending interruptions and then continue with the terminating interruptions, or *termination events*.

Approximately 39% of learning sessions were interrupted and then continued within 10 minutes, the cut-off time after which we classified a learning session as ended (see Table 4.1). During the 327 learning sessions, we recorded a total of 276 interruptions. There were between 0 and 9 interruptions per session ($M = 0.84$, $SD = 1.53$) and an average interruption lasted 28.0 seconds ($SD = 67.4$s). The shortest suspending interruptions were barely a second long, the longest 8:46 minutes. In sum, we registered 276 suspending interruptions during all 327 learning sessions, of which we classified 197 as internal (i.e., app switches without indication or screen locks), 86 as device-internal (i.e., app switched due to calls, SMS, or notifications), and three as ambiguous (i.e., screen lock). The interruptions our algorithm classified as device interruptions mostly followed a notification issued by a messaging app (52 of 87 cases, i.e., 59.8%).

Session termination events are classified as shown in Figure 4.3: we combine the ESQ data with additional checks for device interruptions and internal interruptions. Of the 266 learning sessions supplemented with ESQ data, users ended 168 intentionally, while 99 were ended after an interruption (36.8%). In 19 cases, participants confirmed that it was really necessary to interrupt the learning session (*Very important - it was urgent / time-critical*). For 38 situations, they selected a moderate level of urgency

Table 4.1: Clustered by the interruption type, this table presents an overview of the number of suspending interruptions in total as well as the minimum and maximum of interruptions in one learning session (more than one interruption possible). Further, the table outlines the length of these interruptions in seconds. Note that the type "ambiguous" contains interruptions where the automated classification could not ultimately determine if the source is internal or external.

Interruption Type	$\mathbf{Count_{Total}}$	$\mathbf{Count_{Max}}$	$\mathbf{Count_{M\ (SD)}}$	$\mathbf{Length_{M\ (SD)}}$
Overall	276	9	0.84	27.95 s (67.36 s)
Device-internal	48	6	0.15 (0.61)	51.94 s (98.94 s)
Internal	225	8	0.69 (1.26)	17.24 s (44.00 s)
Ambiguous	3	1	0.01 (0.01)	44.04 s (66.77 s)

(*Moderate - I had to do it eventually in the near future*) and in 41 situations, the interruption was avoidable (*Not important - I could have ignored it and continued learning*). According to the ESQs, 37 learning sessions were terminated because of external, 33 because of internal, and 29 because of device-internal interruptions. Adding automatic classifications, we arrive at a total of 37 external, 97 internal, 112 device interruptions, and 3 ambiguous (i.e., internal or external) interruptions.

In the 62 cases where both ESQ data and an algorithmic classification were present, these coincided in 20 cases, i.e., 32.2% (cf. Table 4.2). However, it has to be noted that it was not possible for us to uniquely identify internal termination events. In case LAIRA could not ultimately determine if the interruption was caused by an internal or external stimulus, the event was labeled "ambiguous" and later confirmed as "internal" or "external" through the ESQ. As Table 4.2 shows, LAIRA classified a total of 32 interruptions as device-internal that were later associated to external or internal stimuli by the participants.

Moreover, the comparison of ESQ data and the classification showed that of the 99 unintentionally ended learning sessions, our algorithm had associated 61 with a session termination event (61.2%). On the other hand, of the intentionally ended sessions, 73 were classified as device-internal (43.5%) and 51 as internal interruptions (30.3%). In 42 cases (25%), we detected no terminating interruption. Adding automatic classifications of termination events as shown in Figure 4.3 to the ESQ data, we arrive at a total of 37 external, 97 internal, 112 device interruptions, and 3 ambiguous (i.e., internal or external) interruptions.

Effect of Interruption Type on Interruptions and Termination Risk

The diverse nature of interrupting events and secondary tasks also manifests in characteristics of interruptions, such as their duration. For example, a Welch test showed that of the unambiguously detected interruption types (i.e., internal and device interruptions), device interruptions were significantly longer ($N = 86$, $M = 519.4$ s,

Table 4.2: Confusion matrix of ESQ classification and the computationally classified types of termination events.

	detected: ambiguous	detected: device	detected: internal
ESQ: device	0	15	6
ESQ: external	0	18	3
ESQ: internal	1	14	5

$SD = 989.4\,\text{s}$) than internal interruptions ($N = 197$, $M = 172.4\,\text{s}$, $SD = 440.0\,\text{s}$; $t(99.98) = 3.12$, $p < .05$, Cohen's $d = 0.45$; Bayes factor $BF_{10} = 327.0$). Moreover, participants were more likely to end their learning session unintentionally (i.e., any response not equal to "I was done using the app" in the ESQ) when there was an interruption before ($\chi^2(1) = 10.913$, $p < .05$).

Effect of Interruptions on Sessions Length

We further found that the occurrence of interruptions (yes|no) influenced the length of learning sessions (in seconds excluding interruption time). The results of a Welch test show a significant difference ($t(185.127) = 4.864$, $p < .05$; $BF_{10} = 89226.0$) between the length of sessions with interruptions ($N = 128$, $M = 877.6\,\text{s}$, $SD = 628.6\,\text{s}$) compared to sessions without interruptions ($N = 199$, $M = 538.8\,\text{s}$, $SD = 422.6\,\text{s}$). An additional Spearman correlation analysis suggests a positive relationship between the total number of interruptions within sessions and the length of the learning session ($rs(327) = .284$, $p < .05$). In particular, we can see an increase in learning time with an increase of interruptions (see Figure 4.4). Similarly, there was a significant positive correlation between the total length of all interruptions in a session and the length of the learning session ($rs(327) = .239$, $p < .05$).

Effect of Context on Interruptions

Similar analyses on interruptions showed no significant effect of time of day on the number of suspending interruptions (Welch's ANOVA: $F(2, 192.403) = 2.635, p > .05$, see Figure 4.5). However, the time of day did influence the number of learning session termination events per category ($\chi^2(4) = 14.583$, $p < .05$). Device interruptions were the most frequent type at all times of the day, but were particularly disrupting in the afternoons (detected as session termination event in 68.4% of cases). The ratio of internal interruptions was highest in the evening (40% of session termination events).

We also fitted GAMLSSs for the number of suspending interruptions and the total duration of all interruptions in one learning session. For interruption counts, we used a Poisson distribution to allow for 0 values.

For the number of suspending interruptions, we set the environment, company, time of day, and triggered push notifications as fixed effects. Receiving notifications increased

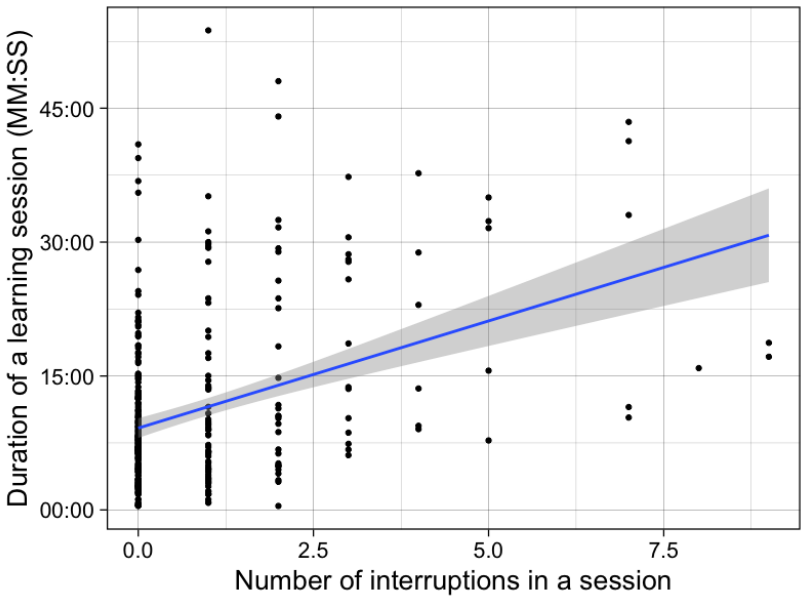


Figure 4.4: Correlation between the number of interruptions that occur in a learning session and the overall task time (excluding interruptions).

Table 4.3: Counts of the interruptions that led to learning session termination according to the ESQs at different times of the day (per interruption type).

Interruption Type	Morning	Afternoon	Evening
Device	13	10	5
Internal	8	12	17
External	6	5	22
Overall (sum)	27	27	45

the predicted number of suspending interruptions ($\beta = 0.53$, $SE = 0.24$, $t = 2.22$, $p < .05$). Besides the intercept ($p < .05$), no other effects were significant. In particular, the occurrence of push notifications led to learners' interrupting their session to open a notifying app with a probability of 31.9% (thus causing a classification as device interruption). Additionally, the number of suspending interruptions at different times of the day is shown in Figure 4.5.

Finally, we fit a model to predict the total duration of interruptions within a learning session from the same factors. The duration was estimated to be highest in the afternoon ($\beta = 0.69$, $SE = 0.29$, $t = 2.35$, $p < .05$ and while traveling or commuting, interruption time was shorter than at home ($\beta = -1.54$, $SE = 0.66$, $t = -2.35$, $p < .05$). The intercept was significant at $p < .001$.

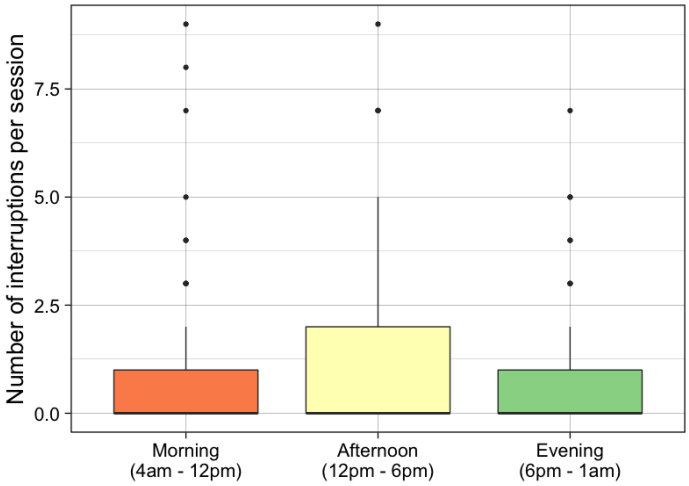


Figure 4.5: Number of interruptions at different times of the day.

Questionnaire: Subjective Assessments of Distractions and Termination

Our user study, in particular the ESQs, helped our participants to self-reflect on their learning behavior. To assess their subjective impression on how they handled interruptions during learning sessions, we presented them with a set of questions at the end of the study. Here, participants stated that they got most easily distracted by external interruptions ($M = 4.27$, $SD = 1.21$, 5-point Likert scale from 1="strongly disagree" to 5="strongly agree") compared to device ($M = 3$, $SD = 1.28$) and internal interruptions ($M = 3$, $SD = 1.48$). Four participants furthermore had the impression that they usually discontinued learning after an interruption occurred, while the majority (7) reported to usually continue learning after a short period of time. When asked how difficult it usually was for them to pick up a learning session after an interruption (5-point Likert scale from 1="very difficult" to 5="very easy"), participants stated medium difficulty, slightly leaning toward easy learning session resumption ($M = 3.45$, $SD = 1.3$). The people who reported problems with resuming the learning task further noted that the main reason was due to loss of focus (P3, P4) and not remembering "[...] *what I was doing before the interruption*" (P2). Those who did not find it difficult to continue a session after an interruption noted that they had established fixed habits (P1) and that content in learning apps is fairly easy (P6, P10).

4.4 Discussion

4.4.1 Limitations

As our user study was conducted during the COVID-19 pandemic, the generalizability of our results is limited in regards to potentially different usage patterns of the application due to the anomalous daily routines of our participants during lockdown and work-from-home phases. Prior work suggests a wide variety of usage contexts for mobile learning applications (e.g., [191, 305]), which we cannot confirm with the data we collected. The majority of the learning sessions we recorded show participants using learning apps at home (311) and only rarely on the go or in a vehicle (11). Still, participants experienced numerous interruptions, with one third of all learning sessions being disrupted. Considering that learning in a home setting is more controlled and quiet than what we can expect from outdoor or public spaces, we estimate that the number of interruptions could be even higher and more diverse for "normal" usage patterns. Similarly, with our sample being gender-biased (all female) and comparably small, a more diverse participant set would be needed to allow the generalization of learning and interruption patterns. Still, this study reveals interesting tendencies and potential patterns in mobile learning and provides implications for the design of mobile learning in light of interruptions.

Furthermore, we annotated the logging data with supplementary information we gathered from the ESQs for the purpose of the study. In particular, we aimed to verify the cause of interruptions and identify the intentional termination of learning sessions. However, we do not know if the participants actually perceived (and remembered) the interruptions detected by LAIRA. Hence, it cannot be guaranteed that an ESQ response always matches the latest interruption. This would also partially explain the low coincidence rate of detected interruption types and the interruption causes selected in the ESQs. Moreover, the use of ESQs is not a feasible approach for everyday-use applications. Based on the data we collected, we can predict device-internal interruptions–caused predominantly by push notifications of messenger apps–but as of yet, fail to successfully distinguish between user-internal interruptions, which are not expressed by actions on the smartphone such as app switches, and external interruptions. The use of additional sensor data (e.g., to detect surrounding noise) or the application of machine learning to train a more sophisticated model of user behavior patterns could facilitate the better distinction of such interruptions.

4.4.2 Observed Interruptions

The data gathered in our study shows that the occurrence of interruptions affects the learning sessions. Even though we excluded the duration of the interruption when calculating the overall learning session length, we found that sessions that have been

interrupted are significantly longer. Furthermore, with an increasing number of interruptions, the overall session time increases as well (again excluding the actual interruption time). This goes in line with prior work suggesting that it takes time to resume a task after an interruption (called the "resumption lag" [329]) and that overall task completion time can increase [22, 171, 186]. These results indicate the need for technology interventions to either reduce the number of interruptions and/or to support the users in dealing with them.

4.4.3 Design Implications: Mitigation Potential

Our study revealed a great variety of interruptions that undoubtedly affect users during their mobile learning sessions. Due to this variety, not all interruptions can be detected automatically and mitigated using the same techniques. Subsequently, we describe four mitigation strategies, focusing on *avoiding* interruptions, *ignoring and postponing* them, preparing for upcoming interruptions, and supporting users in *resuming* the learning task after an interruption.

Avoiding Interruptions The distribution of learning sessions across the day indicates increased usage in the morning and late evening hours. We see two possible reasons for this pattern: (1) common work schedules that entail a dip after 8am and an increase at 5pm and (2) users' circadian rhythm. The circadian rhythm describes fluctuations of alertness and attention over the course of the day, indicating that cognitive and memory performance are highest around 2 hours and 12 hours after waking up, taking a dip in the time in between [75, 297]. Further, variations can occur because of people's individual times of productivity ("night owls" or "morning larks"). In our study, participants learned more often in the evening but session length increased in the afternoon compared to the evening. Furthermore, the likelihood of participants terminating their learning session after being interrupted was highest in the afternoon. Prior work by Schoedel et al. [307] presented the automatic classification of activity behavior based on phone logging data. Similar approaches could be used to recommend individually optimal moments for mobile learning sessions. Hereby, the aim would be to maximize the learners' level of attention and focus by reducing the likelihood of interruptions and the frequency of users succumbing to them.

Ignoring or Postponing Interruptions When LAIRA detected a device interruption, the most frequent source was an incoming push notification. When a notification occurred, participants reacted to this notification in one third of the cases and, thus, interrupted their learning session. In most cases, participants switched to messaging applications. This means that notification management systems (e.g., [111, 157, 251, 264]) that defer notifications until an activity break point is detected would also be a promising approach for mitigating interruptions in mobile learning. This approach is

further supported by the participants' impression that nearly 80% of the interruptions could be either postponed or even ignored.

Preparing for Upcoming Interruptions If an interruption is detectable but not avoidable, at the very least, the learner could be guided to the end of a learning unit and be prepared for an upcoming interruption [45, 329]. Prior work investigated the exploitation of the interruption lag, the time window between noticing and interruption and actually switching to the secondary task. This short time window provides the opportunity to mentally prepare, for example, by consolidating the memory of what a user was currently doing or displaying what they were about to do next [5]. This can be achieved by presenting a summary of what the user just learned or by suggesting them to take written or mental notes to help resume the learning task later on (cf. [124]).

Resuming Learning After Interruptions The data gathered in our study shows that the occurrence of interruptions affects learning sessions. We found that sessions that were suspended by an interruption are significantly longer (even after subtracting the interruption time). Furthermore, the net overall session time increases with the number of interruptions. This goes in line with prior work suggesting that it takes time to resume a task after an interruption (called the "resumption lag" [329]) and that overall task completion time can increase [22, 171, 186]. These results indicate the need for technology interventions to minimize the resumption lag by guiding users back to the original task after the interruption has passed. Task resumption support has shown to positively influence task completion time and error rate after an interruption [293].

4.5 Chapter Summary

This chapter presents an in-depth evaluation of interruptions during mobile learning in everyday settings. With the deployment of LAIRA, we collected valuable information describing learning situations in which learners experienced interruptions. In particular, we were able to gather information on the origin of interruptions, their duration, frequency, and importance. With regards to our research question **RQ1b**, we contribute insights into the actual effects interruptions have on the learning activities. We are able to show that interruptions are a very common and frequent phenomenon in mobile learning and that they are caused by a variety of external and internal stimuli. While almost one third of learning sessions were terminated after an interruption, we notice that participants often report that those interruptions are of lower urgency. Therefore, this chapter discusses opportunities for mitigating the negative effects of interruptions such as postponing interruptions that could be delayed (such as smartphone notifications), preparing for upcoming interruptions if they can be anticipated, or supporting users in resuming the learning task after the interruption. We particularly consider the support of task resumption a promising research approach that will be further investigated in Part IV of this thesis.

III

Embedding Mobile Language Learning into Everyday Actions

5

Embedding Vocabulary Acquisition into Smartphone Authentication

Language learning is a long-term task that requires users to engage with the learning content on a regular basis. Mobile learning applications can enable frequent engagement as they offer the freedom to learn anytime and anywhere [44]. Without a strict curriculum or deadline, however, the decision of when and for how long to learn is often still left to the user. While some strive in the freedom those applications offer, others lack the motivation and perseverance to follow up on their tasks.

In recent years, research explored several approaches of supporting users in their aim to continuously engage with language learning, for example, by making learning content more visible, easy to interact with, or by nudging users in windows of opportunity (e.g., idle moments such as waiting situations). In particular, research investigated the effects of presenting learning tasks pervasively on the lockscreen of the smartphone [81, 113] or using boredom-detection as trigger for the presentation of language learning tasks [91]. Even though those mobile language learning (MLL) apps improved the exposure of learners to the content, many still require the user to actively initiate a learning session and allow for ignoring or dismissing prompts.

In this chapter, we take this approach one step further and explore embedding second-language vocabulary tasks seamlessly into an everyday smartphone interactions. By integrating simple learning tasks into actions we perform anyway multiple times a day on our smartphone, such as unlocking our device, we can create a scenario in which the app initiates learning and a dismissal is as demanding as answering the task. We aim to answer the following research question:

> **RQ2:** How can the integration of learning tasks into the smartphone authentication process foster frequent engagement?

We start by exploring concepts for embedding learning tasks into different means of authentication, such as Personal Identification Number (PIN), pattern, or fingerprint entry. Using high-fidelity prototypes, we gather feedback on the concepts' usability and user experience in a lab-based user study. The results reveal that users appreciate the concept of combining authentication and learning but favor multiple-choice tasks over more complex tasks (sentence-building) or more simple tasks.

Based on the insights from this first user study, we revise our concept and implement an Android application that presents a vocabulary translation task with every unlock event, called *UnlockApp*. We focus on the commonly available fingerprint authentication process and provide the option of using fingerprint sensor gestures as implicit

input method. In an in-the-wild evaluation, we investigate how the different degrees of embedding and self-initiated vs app-initiated exposures impact interactions with the learning content over the day and how users perceive this form of learning in their everyday use. In particular, we compare the users' interactions with the *UnlockApp* to a in-app learning *StandardApp* and an app presenting tasks in continuously visible notifications (*NotificationApp*, similar to the approach of Dingler et al. [91]).

This chapter is based on the following publication:

- Schneegass, C., Sigethy, S., Eiband, M. & Buschek, D. (2021). Comparing Concepts for Embedding Second-Language Vocabulary Acquisition into Everyday Smartphone Interactions. Accepted for Publication in Mensch und Computer 2021 (MuC '21), September 5–8, 2021, Ingolstadt, Germany. ACM, New York, NY, USA, 15 pages. DOI: 10.1145/3473856.3473863

This paper received an *Honourable Mention Award* at the Mensch und Computer conference and was invited to be published as extended version as

- Schneegass, C., Mitrevska, T., Sigethy, S., Eiband, M. and Buschek, D. (2021). UnlockLearning - Investigating the Integration of Vocabulary Learning Tasks into the Smartphone Authentication Process. In submission to the i-com Journal of Interactive Media

Some text passages were taken verbatim from these publications. Further, this section is supported by the Bachelor thesis of Teodora Mitrevska (Section 5.3) and the bachelor thesis and practical work of Sophia Sigethy (Section 5.4.1 and 5.4), see detailed collaboration statement at the beginning of this thesis.

5.1 Related Work

As MLL apps require self-directed learning, it needs to be taken into account that the success of autonomous learning and the adherence to rehearsal schedules depends on the consideration of various dimensions (i.e., context of learners, learner interest and motivation, language proficiency, etc.) [192, 203]. Thus, the overall goal of micro-learning is to maximise the degree of exposure with the learning content due to increased presentation and, thus, frequent repetition of content. In particular for the teaching of languages, the application of the micro-learning approach has proven to improve vocabulary acquisition and recall [53, 100, 331]. Prior work presented different ideas for realizing micro-learning in mobile learning. For example, Dearman and Truong [81] implemented a mobile language learning application that can display learning tasks (i.e., vocabulary translations with three multiple-choice options) on the phone lockscreen, the screen that protects the device from unauthorized usage. The

authors show that users frequently interacted with the application and improved their knowledge of the language over the course of this study. Dingler et al. [91] achieved similar results by implementing and evaluating a vocabulary application that presents tasks in push-notification. Their app "QuickLearn" detects opportune moments for the presentation of the learning content, in particular, when users are bored and use their smartphone without a specific purpose. Dingler et al. [91] concluded that their application increased the number of "quick" learning sessions on the go, which was appreciated by their participants.

5.1.1 Everyday Smartphone and Lockscreen Interaction

For many interactions on smartphones, users do not even unlock their device [142, 143] or only use one application after unlocking [223]. However, unlocking the mobile device is still an action that happens frequently throughout the day. Mahfouz et al. [223] report that users unlock their phone on average 46 times, while other studies report an average of between 25 [142] and 47 unlocks [128] per day per participant. As this interaction has no other purpose than the authentication, prior work suggests combining the authentication event with a microinteraction task. Ashbrook [16] defines microinteractions as interactions that take less than four seconds to complete, making them less interrupting helps users resume their primary task quickly.

An example for such a microinteraction is the integration of data collection tasks into the authentication process, particularly slide-to-unlock [330]. This idea has been proposed as a quick and easy method to enter journaling data [362], perform nutrition tracking [167], or sleep tracking [63]. Slide-to-unlock is, however, only one of multiple common methods of authentication. As our mobile device store an increasing amount of sensitive data, protecting access to them became a common practice.

5.1.2 Authentication Methods and Their Prevalence

In general, smartphone authentication mechanisms can be divided into four categories according to Wang et al. [345]:

1. **Knowledge-based**: Require the user to remember certain information, either text-based or graphical. *Example*: Passwords, PINs, patterns, or graphical passwords
2. **Physiological Biometric-based**: Makes use of unique identifying biological traits of users such as their fingerprint or facial features. This form of authentication can be performed upon explicit prompt or implicit assessment. *Example:* Fingerprint, voice, and face recognition

3. **Behavioral Biometric-based**: Captures users' unique behavioral patters or characteristics. *Example:* Gait recognition, tapping behavior, hand gesture
4. **Mixed Models / Multi-factor**: Combination of two or more approaches to increase security. *Example:* Integrating knowledge factors with biometric authentication

In 2016, Malkin et al. [224] performed an online survey with more than 8000 smartphone users about their unlocking mechanism. While around one third (32.1%) of the participants stated to use slide-to-unlock, 32.5% used pattern authentication, 18.7% PIN, and each around 7% used a password or biometric authentication (e.g., fingerprint). With a decrease in popularity of passwords for mobile authentication due to an increased number of passwords one needs to remember and a comparably high effort and time to input them, visual passwords (i.e., patterns) gained prevalence in the first decade of the 20th century [342]. However, these patterns have a more limited diversity than passwords and are prone to guessing or observer attacks. With the implementation of fingerprint sensors in smartphones, implicit biometric authentication became popular. Although the prevalence is steadily increasing, common techniques (i.e., pattern, PIN, swipe) are still in use for many reasons. For example, users want to protect their privacy by not digitalizing their biometric information such as fingerprints. Further, implicit authentication can be prone to failure [24, 30, 266], with especially the fingerprint entry showing an increased number of errors [269]. Reasons for increased error rate can be reduced sensitivity of a fingerprint sensor due to dirt or water, making knowledge-based methods such as PIN, pattern, or passwords a valid complementary "fallback" method.

For increased security, knowledge-based mechanisms can be included in multi-factor authentication schemes. Thus, even with implicit authentication such as fingerprint input or FaceUnlock becoming more common, PINs, passwords, and swipe-to-unlock do not become obsolete. Findings from studies such as by Qiu et al. [269] show diverse user preferences and authentication behavior, thus, strengthening the argument that offering multiple different unlocking mechanisms for smartphones is necessary.

Therefore, the first part of this chapter will present an exploration and evaluation of embedding different learning tasks into a diverse set of smartphone authentication methods, i.e., PIN, pattern, swipe-to-unlock, or fingerprint input using fingerprint sensor gestures.

5.2 Concept: Integrating Learning Tasks Into Authentication Methods

In the following, we will outline the learning tasks and the authentication methods we explored in more detail. As this works' idea aims for the embedding of learning

tasks into the authentication action – creating one seamless interaction – we will break down both the learning and authentication actions into their individual interactions (i.e., discrete vs. continuous input, single vs. multiple inputs).

5.2.1 Authentication Interaction

We characterized the authentication methods from an interaction perspective, thus, it was less important what the user actually inputs (characters, numbers, etc.) but rather what type of interaction is required. We distinguish the following four types of interactions found in common smartphone authentication processes:

- **Single discrete input**, aka. "tap" – An individual tap that can be used for any button when applied on the touchscreen, to unlock the phone when applied on the unlock hard key or *fingerprint sensor*.
- **Multiple discrete inputs**, aka. "sequence" – A set of consecutive taps that can be used to input a *PIN* or *password* on the touchscreen.
- **Uni-directional continuous input**, aka. "swipe" – A continuous gesture that can be used as *swipe to unlock* on the touchscreen or as gesture input on the *fingerprint sensor* (current sensors recognize four swipe directions)
- **Multi-directional continuous input**, aka. "pattern" - A continuous gesture that connects swipes in multiple directions without lifting the finger from the touchscreen, as it is known from common *pattern* authentication formats.
- **Implicit input** – This format includes all authentication methods that do not require the user to perform an action on the smartphone such as *FaceUnlock* or *voice recognition*.

5.2.2 Learning Task Interaction

We selected three different tasks for the interaction with the language learning content. All tasks are of simple design and focus on vocabulary recognition. They require only micro-interactions with the aim to be feasibly for solving the task on the lockscreen and during the authentication process. We included the commonly used **Multiple-Choice** question format, in which the correct answer is presented accompanied by a number of incorrect answers (also called "lures"). The user has to pick the correct answer from the set. While this format can be used to present multiple correct answers, we decided to implement the simplest version of only one correct answer so only one click (single tap) will be needed to solve the task. With the simple interaction, this task is specifically feasible to be used for single touch authentication methods such as swipe-to-unlock or fingerprint authentication (using fingerprint sensor gestures, see details

in subsection 5.2.1). The number of presented "lure" answers can further be adapted considering the limited space on the lockscreen but needs to be chosen carefully as too many answer options can increase the number of incorrect answers in later testing situations [43, 50] (i.e., "Negative Suggestion Effect" (NSE), for summary see [299]). For new learning content, Roediger III and Marsh [281] recommend to use fewer alternatives in the beginning. Thus, we will present only one to two lure options as addition to the correct answer. Furthermore, at this prototypical stage, it is possible to present either the native word (L1) and ask user for the second-language translation (L2) or present the L2 word and ask for the L1 translation.

As a second format, we implemented a "**Check Word**" task along the vocabulary test yes/no format, for example, applied in the Eurocentres Vocabulary Size Test [234, 235]. This task format aims to present a large number of words while requiring minimal time and effort by the user. In our examples, the L1 words are presented along the L2 translations while the user has to state if he/she knew the translation or not (single tap on a yes or no button). Similar to the the multiple-choice task, the yes/no answer format can be implemented also for single interaction authentications. To control for an overestimation of knowledge caused by users marking unknown words as known, the later application could include phonetically or orthographically similar pseudowords. In contrast to the yes/no test by Meara and Jones [235], this format (with included translation) tests not only the recognition of words but also checks if learners are aware of the word's meaning.

As a third tasks, we include a **Sentence-Building** exercise. Similar to the multiple-choice task design, the user is presented with a set of words from which they are asked to select those that make up a semantically and syntactically correct sentence. Due to the limited space of a smartphone lockscreen that simultaneously displays the authentication process, we chose to present simple sentences with three words that need to be chosen out of a set of six words (three correct words, three lure answers). For example, a task could look as follows:

A	green	**lemon**.
An	**yellow**	tomato.
He	drinking	**water**.
They	**drinks**	wood.

A sentence-building task can thus combine vocabulary recall with applying the correct grammatical construct. In the first example above, picking the correct article (a/an) and remembering the meaning of the words green/yellow and lemon/orange. In the second example, choosing the correct combination of pronoun (he/they) and verb form (drink/drinking) and remembering the meaning of the words water and wood.

(a) **Mockup 1** PIN with Multiple-Choice

(b) **Mockup 2** Fingerprint with Check Word

(c) **Mockup 3** Pattern with Sentence-building

Figure 5.1: Three mockups[a] to visualize our concept for integrating learning into the authentication process. **Mockup 1** outlines a multiple choice task after PIN entry. **Mockup 2** Asks the user to indicate if the word is known (green button with arrow) or unknown (red button with cross). **Mockup 3** shows how the sentence-building can be integrated into pattern authentication.

[a] Mockups created using Figma (`https://www.figma.com/`, last accessed January 3, 2022) and the Noun Project (`https://thenounproject.com/`, last accessed January 3, 2022).

5.3 Prototypes

For the generation of our prototypes we combined all authentication methods described in Section 5.2.1 with all learning tasks presented in Section 5.2.2. This section will give an overview of the prototypes and present an evaluation of users' experience in interacting with them.

5.3.1 Mapping of Learning Tasks and Authentication Methods

In table 5.1 we visualize how our different prototypes address combinations of learning task and authentication method. All interaction concepts were implemented as high-fidelity interactive prototypes (PT1 to PT12) using the Justinmind[2] prototyping tool for mobile applications. The complete set of prototypes (mockups and explanations of functionality) is included in the Appendix A. For example, PT6 is depicted in Mockup 1 (see Figure 5.1a). The users enter their PIN and tap the screen once more to select one out of three translations for the L2 word presented below. Prototype PT3 implements Mockup 2 (see Figure 5.1b) by combining a *check word* task with a *uni-directional continuous* input, i.e., a swipe gesture starting at the on-screen fingerprint sensor to either of two buttons to indicate if the translation is known or unknown. This prototype is also applicable for other uni-directional continuous input types such as slide-to-unlock. Since sentence-building inherently requires connecting multiple words, it is not applicable for single input (discrete or continuous) authentication methods such as swipe-to-unlock. A sentence can be build with a multi-directional continuous input as visualized in Mockup 3 (PT12; see Figure 5.1c). Here, the user enter their predefined authentication pattern and, without lifting the finger, continue the gesture through connecting the words presented below to form a sentence.

Table 5.1: Overview of all prototypes mapped to the three learning tasks (columns) as well as the four interaction types (rows).

	Check Word	**Multiple-Choice**	**Sentence-Building**
Single Discrete	PT1	PT5	(not applicable)
Multi Discrete	PT2	PT6	PT11
Uni-directional Continuous	PT3	PT7, PT8	(not applicable)
Multi-directional Continuous	PT4	PT9	PT10, PT12

5.3.2 Exploratory Evaluation of Prototypes

To investigate user's preferences for our different prototypes, we performed a user experience evaluation with ten in-depth semi-structured interviews. We showed the participants the interactive prototypes of the 12 concepts on a smartphone, asked them to interact with the prototype, and collected their verbal responses. The focus of this user study was to gain insights into the usability and user experience of the concepts.

[2] Justinmind: `https://www.justinmind.com/`, last accessed January 3, 2022

Procedure

At the beginning of each session, we welcomed the participants and informed them about the purpose of the study. After agreeing to the data protection policies along the GDPR regulations, participants signed a consent form. A short questionnaire was presented asking for demographic information such as age, gender, degree of education, current occupation, their experience with mobile learning, their current authentication technique of their smartphone, and its operating system.

Afterward, we introduced the participants to our prototypes and varied the order of presentation between the participants. We viewed the prototypes on a Sony Xperia Z smartphone with a 5-inch screen through the Justinmind app and let them interact independently with one prototype after another. We asked the participants to follow the think-aloud protocol, verbalizing their thoughts about the interface, interaction, and usability as well as problems of understanding. If the user was not able to perform the required interaction, we explained and / or demonstrated it and the user was asked to retry the interaction. We explicitly advised all participants to not input any authentication PIN or Pattern during the study that is similar to their own.

For each prototype, the interview was guided by the following user experience metrics: *Easy of use*, *Enjoyment*, *Interaction Speed*, and *Willingness to use* the prototype as learning method in the future. Those facets were posed as yes/no questions but with the collections of further comments. The participants were furthermore encouraged to state any concerns regarding the prototype as well as potential improvements that come to their mind. Overall, the study required around 60 minutes participation depending on the extend of comments users had about the prototypes. We transcribed the comments of the participants and applied a thematic analysis approach for comments and suggestions. This method is commonly used for identifying, analyzing, and reporting themes found within a qualitative set of data such as interview transcripts [40, 249].

Participants

Ten participants took part in our study (2 female, 8 male) with a mean age of 24.2 ($SD = 1.55$). From those ten participants, nine were currently students and one stated to be a software engineer. In total, six participants reported to have obtained a Bachelor's degree, three a Master's degree, and one a PhD. Six participants owned an android phone, three used iOS, and one user used both operating systems due to owning two devices. The most common used method of authentication was the fingerprint used by eight participants, whereof three stated having implemented a fallback authentication (one stated PIN, one password, and one Pin and FaceUnlock). One participants stated to use a pattern to authenticate and the last one used FaceUnlock. Furthermore, four participants stated to currently learn a language on their mobile device.

Table 5.2: Users' assessment of the prototypes along the dimensions ease of use, enjoyment, interaction speed, and their willingness to use the prototype in their daily lives. Questions were posed in a yes/no format and the answers visualized in the table as "+" when 8 or more users agreed; similarly, a "-" indicates two or less users agreeing.

Task Type	Action	Prototype	Easy to use	Enjoyable	Fast	Would use it
Check Word	SD	PT1	+	+	+	+
	MD	PT2	+	o	+	+
	UdC	PT3	+	+	+	+
	MdC	PT4	+	o	+	o
Multiple-Choice	SD	PT5	+	o	o	+
	MD	PT6	+	+	+	o
	UdC	PT7	+	+	o	o
	MdC	PT8/9	+	+	+	+
Sentence	MD	PT10	o	o	o	o
	UdC	PT11	o	o	-	-
	MdC	PT12	+	o	-	o

5.3.3 Results

We transcribed the responses of the participants and will report on them in the following clustered by the three task types *Check Word*, *Multiple-Choice*, and *Sentence-Building*. The answers of two participants for each one prototype (P6's PT19 and P8's PT10) had to be excluded due to technical problems with the audio recording device. We will report the summative evaluation of the users' answers to the four guiding questions on easy of use, interaction speed, enjoyment, and willingness to use, and report on the open comments in the style of a thematic analysis [40, 249].

Check Word The majority of the participants rated all Check Word prototypes as easy to use in their interaction. For PT2 (multi discrete) and PT4 (multi-directional continuous), each one participants stated that the mechanism lacked intuitiveness. PT1, PT2, and PT3 were unanimously rated as fast, while P7 was the only participant rating PT4 as not fast. In particular, PT4 was not perceived very positively by the participants. Five out of ten indicated that they would not be open to using this prototype in their daily lives, arguing it is "no fun" (P3, repeated by P5). Others would not use it as they state to never use Swipe-to-unlock in general due to low security (P7, P8). In general, participants doubted the efficiency of the check word task and expected a more engaging learning task.

Multiple Choice All participants considered PT5, the simple tap on one out of two answer solutions, very easy to use and very fast. P1 and P9 considered it not very enjoyable, while three participants stated to be not willing to use it due to a lack

of security. Compared to the single tap needed for PT5, participants perceived PT6 with multiple discrete taps needed as not fast (P3, P7, P8, P9). P8 attributes the lack of speed to the PIN authentication method in general. Similarly to PT5, three participants would not use it due to the authentication method and the "*cluttered screen*" (P9). Out of the four multiple-choice prototypes, PT7/8 was best received by our participants. Nine out of ten mentioned they would use it and all considered it fast and enjoyable. Two participants considered it not intuitive enough and expected a tap instead of a swipe gesture. Lastly, for PT9, the multi-directional continuous input, the majority of participants considered it easy to use (9/10) and would use it in their daily lives (8/10). However, P8 noted that it felt slow because one first has to think about the pattern input and then, while not releasing the finger, decide which answer to pick. Similar to the multi-directional continuous input for the other tasks, the screen was perceived as cluttered. As suggestion for improvement participants stated to reduce the number of answer options from three to two.

Sentence-Building Overall, the sentence-building tasks were perceived as least intuitive compared to the check word and multiple-choice tasks. PT11 with its multiple discrete inputs required was considered time consuming by eight out of ten participants and also as not enjoyable (4/10). P4 highlighted that due to the many actions one has to perform, "*it's not fun at all, so many steps just to unlock make this task feel like a chore*". While the continuous multi-directional input (PT12) was rated to be faster and more enjoyable, the majority of the participants still indicated that they are not willing to use it. In particular, P7 mentions that a complicated task like this does not belong on the lockscreen and that the multiple-choice or check word tasks were a better fit.

Limitations

During our study, participants expressed difficulties in rating the interaction with the embedded authentication learning task without being biased by the authentication method used. In general, several participants considered the swipe-to-unlock as insufficiently secure and would not use this mechanism. We stressed that this technique could be integrated into fingerprint authentication and should be rated independently from users' opinion on the authentication itself. However, we can not be ultimately certain that our participants were able to avoid any bias. We further noticed that the alignment of the learning tasks below the authentication interface was not optimal. In particular, for continuous input gestures such as in PT9 and PT12, users had to interrupt the continuous gesture to have a look at the options again. The visibility of the options will be ultimately defined by the end-point of the pattern input – i.e., when the pattern ends in the lower left corner and the user is right-handed, the hand will almost completely cover the answer options.

5.3.4 Summary

In this first part of the chapter, we presented twelve prototypes for combining three learning tasks (check word, multiple-choice, and sentence-building) with four common forms of smartphone authentication mechanisms (represented through their required interaction: single discrete input, multiple discrete, single continuous, multi-directional continuous). A first usability evaluation with ten participants showed that users preferences vary greatly. Most prototypes were reported to be easy to use and fast. While some participants preferred discrete taps, others especially enjoyed the single and continuous swipe input schemes.

The main take-away from this preliminary study is that users desire a quick and easy learning interaction. While they value the idea of combining learning with their authentication, many participants emphasized that the learning should not impede or obstruct the unlock process. In particular, the sentence-building tasks were considered fairly slow and not a good fit for the presentation on the lockscreen. We will not follow-up with this type of task in further evaluations.

Lastly, we noticed that the speed perception varied greatly among participants. While some found the interaction slower than necessary for an authentication method, others perceived the learning task itself as rather quick. Along the glass half-full vs. glass half-empty situation, we argue that the perception strongly depends on the users' perspective on the situation. As the proposed methods combine authentication and learning, it is undeniable that the interaction will take longer than authentication alone. However, when comparing the combined interaction with the action of having to open a stand-alone learning app to learn a vocabulary, it is undeniable that the newly proposed prototypes enable a faster interaction. Thus, for future evaluations we will include a standalone app as control condition for participants to compare the interaction fairly.

5.4 Comparative Evaluation of Embedded Learning Concepts

In this section, we present an in-the-wild evaluation investigating users' experience with different levels of embedding of vocabulary learning tasks into smartphone interaction. We implement and compare three applications:

1. The *UnlockApp* approach that is informed by the results of section 5.3.2 and embeds learning tasks into the smartphone authentication process. We extend the concept of Dearman and Truong's [81] learning wallpaper – by connecting a simple multiple-choice vocabulary task that the user can answer by a single button press with the phone unlock action. Although users can still dismiss the

learning task, this concept is strongly embedded into the unlock action as the app initiates the learning task (see Figure 5.2a. We aim to nudge users toward more frequent learning by lowering the threshold for users to interact with the content.

2. The *NotificationApp* presents a learning task in a continuously shown notification. The notification increases learning task's visibility during interactions with the lockscreen or status bar but still requires users to actively initiate learning (see Figure 5.2c). Our design is informed by the concepts proposed by Dingler et al. [91]).

3. We further implement a *StandardApp*, following a baseline self-contained and self-initiated learning design, which the user has to actively start and quit (see Figure 5.2d).

The key conceptual difference between the three concepts is that the *UnlockApp* nudges the user to interact with the content, while the *NotificationApp* and *StandardApp* require users to actively initiate the learning themselves. However, in contrast to the *StandardApp*, the *NotificationApp*'s constant visibility acts as a continuous reminder for users to engage with the learning content.

All three apps include vocabulary allowing for learning an L2 language with translations into German. In the background, all three applications are connected to a Firebase database, which stores both the learning content (around 450 words per language) and interaction logging data. In the database, each user is assigned a unique id to ensure anonymity.

5.4.1 Implementation

The following sections will outline the implementations of *UnlockApp*, *NotificationApp*, and *StandardApp*.

UnlockApp

In this section, we report on the implementation of an application that embeds learning into the authentication process. As a result of the first study part, we decided to use a multiple-choice learning task with two answer options. The question type was considered easy-to-use and enjoyable (cf. 5.2) and hesitation from participants in regard to potential usage concerned only the prototype's security, particularly when used with a swipe-to-unlock mechanism. However, our participants stressed the importance of a quick microinteraction that does not hinder them in whatever task they were about to perform. Thus, we decide to implement the multiple-choice task combined with the fingerprint authentication, that allows for a combination of fingerprint input and (1) a

single discrete tap on the screen or (2) a single continuous gesture on the fingerprint sensor to select the correct answer.

To integrate the learning task into the authentication action, we display the vocabulary translation interface immediately when the user unlocks the screen (see Figure 5.2a). As the current Android versions do not allow a direct manipulation of the lockscreen for security reasons, we chose to implement a context-registered broadcast receiver. This receiver listens for the unlock event (ACTION_USER_PRESENT indicates that the user is present after the device is woken up[3]) while running continuously in the background and places the app in the foreground when an unlock occurs. Since it is possible that applications are shut down for battery optimisation purposes, we further included a foreground service that allows for automatic restart of the application. This way, the task is presented immediately after the unlock occurs, giving the impression of one seamless interaction.

When an unlock event occurs, the *UnlockApp* asks the user to translate one word from L1 to one of the three languages of choice (L2) – Spanish, French, or Swedish. To keep the learning task as short as possible, the interface only presents the translation task, two potential translation options in the form of buttons. Further, we include a "skip" option to keep learners from guessing the answer, as guessing increases the risk of experiencing the negative testing effect [228] (i.e., remembering the false answer one gave to a multiple-choice question rather than the correct solution).

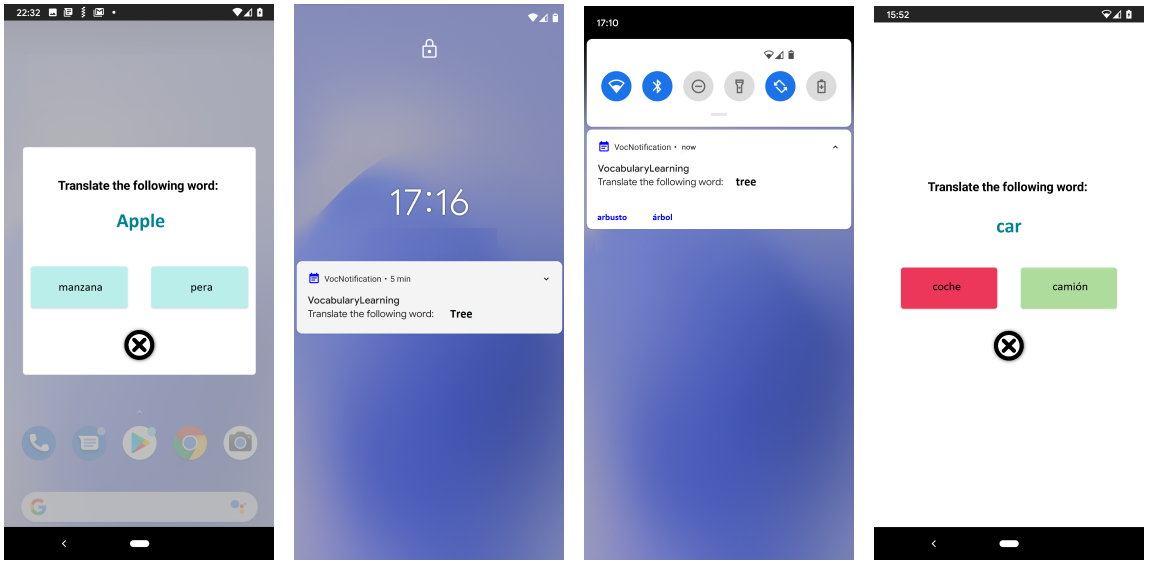
After the user selects one of the two translation options, the *UnlockApp* displays corrective feedback by highlighting the answer option in either red or green. Providing feedback in multiple-choice learning tasks is essential and counteracts the negative testing effect [51]. After a short delay for users to perceive the feedback, the app moves into the background.

We chose a translucent background to not distract the user too much from the primary task they were about to accomplish when unlocking the phone. Prior research by Hodgetts and Jones [146] has shown that when the primary task was still visible during an interruption, participants are able to resume the task faster. To ensure consistency, we used the Android material design and color palettes.

The *UnlockApp* further allows to use a fingerprint sensor gesture (swiping left or right on the fingerprint sensor) as input to answer the tasks. We utilise the Fingerprint Gesture Controller accessibility service[4]. This controller registers gestures performed on the fingerprint sensor, independent from sensor location (depending on the model, the sensor can be located on the back of the phone, on the side, or on the front below or on the screen). As this is still a very rare feature and not supported by many Android devices, we will not evaluate the usage of it.

[3] Android User Intent: `https://developer.android.com/reference/android/content/Intent`, last accessed January 3, 2022

[4] Fingerprint Gesture Controller `https://developer.android.com/reference/android/accessibilityservice/FingerprintGestureController`, last accessed January 3, 2022



(a) (b) (c) (d)

Figure 5.2: (a): The *UnlockApp* displays a vocabulary task immediately after the user performs the authentication. **(b+c):** The *NotificationApp* presents the learning task in a continuously displayed push notification. This notification is visible on the lockscreen (b) and in the (pulled down) status bar (c). Via notification action buttons (also displayed on the lockscreen if interactivity is enabled), the user can select one out of two translations for the respective word. **(d):** The *StandardApp* in the state of showing corrective feedback.

NotificationApp

In this app, learning content is presented in a continuously displayed push notification, both in the (pulled down) status bar of the smartphone (see Figure 5.2c) but also on the lockscreen (see Figure 5.2b), enabling the user to interact with the content at any point in time and answer as many tasks as they like to. The notification displays the task and word to be translated as well as two potential translation options via action buttons[5]. Those buttons are frequently applied by other applications to allow immediate interaction with the app sending the notification. For example, Gmail employs action buttons to enable users to delete or reply to emails directly from the notification. We implemented a broadcast receiver to listen for the input in the notification and check the correctness. In the same style as the *UnlockApp*, green and red coloring of the two potential translations indicates corrective feedback. Afterward, the notification is updated and a new word is shown. Likewise, when the user dismisses a notification or restarts the app, it presents a new word. The notification itself cannot be removed by the user but remains constantly visible in the pulled down status bar.

[5] Android Notification Action Buttons: `https://developer.android.com/training/notify-user/build-notification#Actions`, last accessed January 3, 2022

StandardApp

In contrast to the *UnlockApp* and *NotificationApp*, the *StandardApp* does not show any continuous reminder of the learning task. The users have to actively open the application and engage with the learning content. Once the users open the app, they can complete as many learning tasks (i.e., request as many translation tasks) as they like. The appearance of the application's interface is consistent with the interface of the *UnlockApp*: The app shows the native word, the two translation options in buttons, a dismiss button, and corrective feedback for the task (see Figure 5.2d). In contrast to the *UnlockApp*, the background is fully opaque.

5.4.2 Methodology

Our evaluation follows a within-subject design, where all participants interact with all three different applications to allow them to draw comparisons. Each app is used over the course of one week (seven days) before switching to the next app. The vocabulary progress is preserved so that people can continue in the next app where they left off in the previous one. The order of the apps is counter-balanced across participants to avoid sequence effects.

In our subsequent analysis, we explore differences in users' interaction with the three apps (independent variable), in particular, frequency and duration of interactions, as well as perceived usability and experience (dependent variable). Based on related work, we particularly evaluate differences among the three apps with regard to how much they expose users to the learning content, stating the following main hypothesis:

H_1 Embedding vocabulary tasks into everyday smartphone interactions as in the *UnlockApp* and *NotificationApp* leads to a higher number of learning tasks solved by users over the course of the study compared to the *StandardApp*.

Furthermore, we gather individual feedback on participants' preferences and experiences when learning with the three applications.

Procedure

We provided participants with a detailed installation guide enabling people to download the three applications from the Google Play Store (setup as a non-public test version), after they gave informed consent. The guide also encouraged them to restart the (current) app in case they restart their phones or the app is closed for any other reason and does not restart automatically.

At the beginning of the study, people filled in a questionnaire on demographics and current authentication method, smartphone usage habits, language proficiency, and motivation to learn a new language. In this process, people selected a language to

learn with the apps (Spanish, French, or Swedish). We informed participants that the vocabulary taught in the apps is on a beginner's level and thus recommended to choose a language they are not very proficient in.

During the study, we logged data on the users' interaction with the apps, including the number of solved vocabulary tasks, their correctness, response times and task dismissals (in the *UnlockApp*).

Sample

We recruited 30 participants via University mailing lists, social media channels, and word-of-mouth, nineteen identifying as female and eleven as male. Their average age was 29.8 years ($SD = 15.55$, range 19-78 years) and the majority (18) reported to be studying or working full time in jobs such as Engineer, Physicist, Accountant, Doctor, or Dentist, while two were pensioners. Fifteen stated having a high school degree and twelve a university degree (including bachelor, master, or phd). One third (19 people) reported using fingerprint authentication on their smartphones, and seven reported using a PIN or password, two patterns, and two face recognition. All participants stated to use their device at least multiple times a day (15) if not multiple times per hour (15).

At the time of the study, ten participants confirmed learning or actively improving on a language (four Spanish, three French, two English, and one Japanese). They further stated to be proficient in at least one and up to five foreign languages ($Md = 3$), and reported to have great interest in learning a new language ($M = 6.07$, $SD = 0.85$, Likert-scale from 1="I fully disagree" to 7="I fully agree"). As languages to learn in this study, thirteen people chose Spanish, twelve French, and five Swedish. For their successful participation all people received a 25 Euro Voucher for an online store, or an equivalent amount of study credit points.

5.4.3 Results

We report on the data of 30 people. However, one person did not complete any of the final questionnaires, and two additional people did not submit the questionnaire for the *StandardApp*. Therefore, we report the interaction data of 30 and the questionnaire results of 29 people for the *UnlockApp* and *NotificationApp*, 27 for the *StandardApp*, respectively. For the statistical analysis of the questionnaires, we performed pair-wise exclusion for incomplete data sets.

We used R [272] for significance testing; concretely, (generalised) linear mixed-effects models (LMMs, packages *lme4* [26] and *lmerTest* [200]). The LMMs accounted for individual differences (*participants*) and *app order* via random intercepts. Note that app order was counterbalanced yet we still included it here following best practices. As fixed effects, we included *app*, plus the *number of days* since the start of the study.

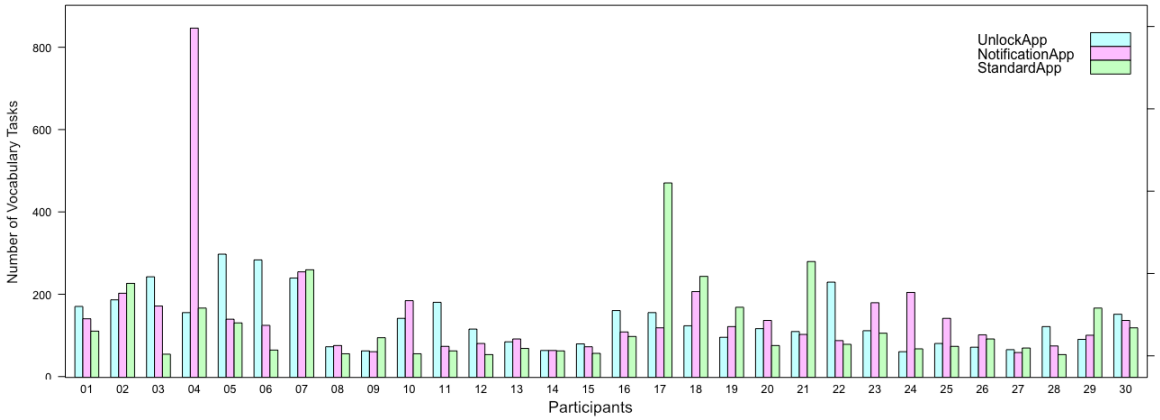


Figure 5.3: Overview of participants' usage of the three apps during the study: (per participant from left to right) the overall number of vocabulary tasks solved in the *UnlockApp*, *NotificationApp*, and *StandardApp*.

Overall Usage

In total, we recorded 7715 answered vocabulary tasks over all participants across the three weeks of our study. The most tasks were answered using the *NotificationApp* (2945), followed by the *UnlockApp* (2604) and the *StandardApp* (2166). On average per day, people used the apps to solve 10-13 vocabulary tasks. Here, the *StandardApp* showed the lowest mean usage ($Md = 2,\ M = 10.28,\ SD = 20.05$), while the *UnlockApp* ($Md = 8,\ M = 12.28,\ SD = 14.73$) and *NotificationApp* ($Md = 5.5,\ M = 13.7,\ SD = 28.43$) were used more frequently. While the *NotificationApp* shows the highest overall usage, the Median is lower than the Median of the *UnlockApp*. This is due to one exceptional case in the *NotificationApp* usage (see P4 in Figure 5.3): P4 reported having used the *NotificationApp* extensively out of enjoyment, with over 800 answered vocabulary tasks (increasing the overall usage count but with little effect on the Median). Since this was intended use, we do not remove this as an outlier from our analysis.

For significance testing, we fitted a generalised LMM (Poisson family) on the answer count data (i.e., number of answered vocabulary questions). The model had *app* as a significant positive predictor (*UnlockApp*: β=.43, SE=.06, $CI_{95\%}$=[.31, .55], p<.0001; *NotificationApp*: β=.36, SE=.06, $CI_{95\%}$=[.25, .48], p<.0001): Therefore, compared to the *StandardApp*, using the *UnlockApp* was estimated by the model to result in $\exp(\beta) = 1.54$, that is, 54% more answered vocabulary questions. Similarly, using the *NotificationApp* was estimated to result in 43% more. Moreover, the model had *day* (since start of the study) as a significant negative predictor (β=-.15, SE=.01, $CI_{95\%}$=[-.17, -.12], p<.0001): The number of answered questions was estimated by the model to decline over the course of the study (estimated as $\exp(\beta) = 0.86$ i.e., -14% per

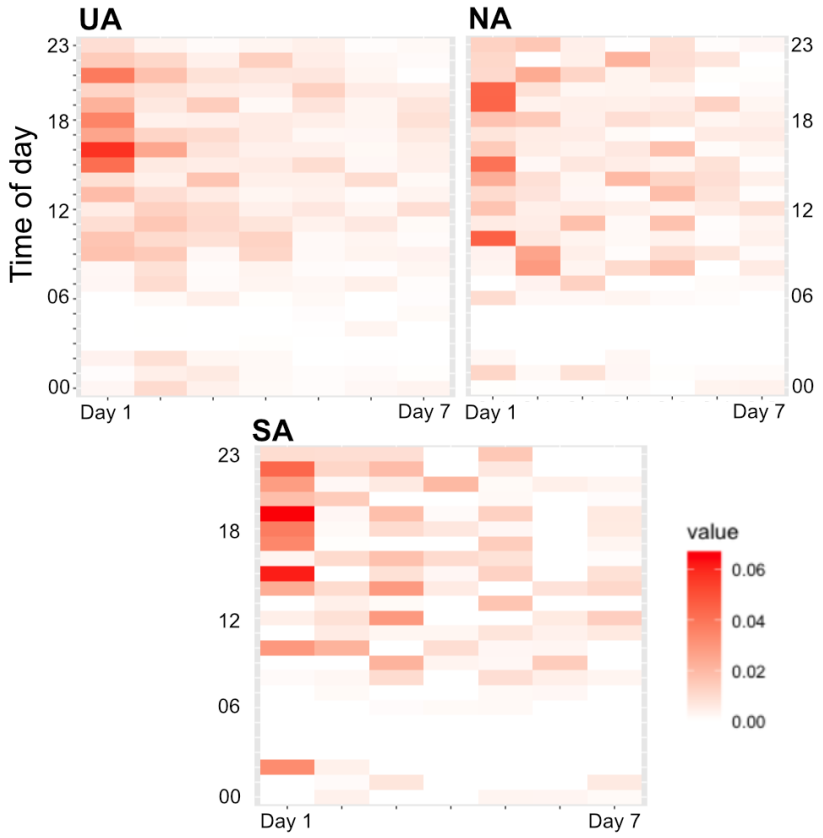


Figure 5.4: The relative distribution of vocabulary tasks solved for the *UnlockApp* (left), *StandardApp* (middle), and *NotificationApp* (right) in percent (0.06 = 6%). The amount of vocabulary tasks is normalised for each user with respect to their overall number of solved tasks and visualised according to the seven usage days (x-axis) and the hours of a day (midnight to midnight, y-axis).

day). The interaction of *day* and *app* was significant for *UnlockApp* (β=-.07, SE=.02, $CI_{95\%}$=[-.10, -.04], p<.0001), but not for *NotificationApp* ($p = .176$).

We further plotted people's interactions with the three applications over the course of the whole day and the week of use: Figure 5.4 visualises the number of tasks solved by the users, indicating that they interacted with each app most frequently on the first day of use. It has to be noted here that participants started the study on different days of the week (Monday: 5, Tuesday: 2, Wednesday: 2, Thursday: 5, Friday: 5, Saturday: 6, Sunday: 5) and the starting weekday then stayed the same for each person across app conditions. Moreover, the plots reveal that interactions with the *UnlockApp* are scattered more across the time of the day compared to the interactions with the *StandardApp* and *NotificationApp*. Furthermore, the plot matches the results of the LMM that the number of answered questions declined from day one to day seven.

Table 5.3: Overview of people's use of the three apps in terms of the number of answered vocabulary questions.

App Type	# Answers Total	# Correct	# Incorrect	% Correct
StandardApp	2166	2026	140	93.54%
NotificationApp	2945	2762	183	93.79%
UnlockApp	2604	2380	224	91.40%

Learning Task Correctness

In general, the users answered more than 90% of the learning tasks correctly over all three applications (see Table 5.3). The correctness rates for the *StandardApp* and *NotificationApp* are slightly higher than the rates for the *UnlockApp*. A Friedman test revealed no significant difference ($p > .05$) for the task correctness of the answers with regard to the three applications.

For the *UnlockApp*, we can further measure the time between when the learning task is presented and the moment the answer is recorded. On average, it took participants 3.4 seconds to answer a task using the *UnlockApp* ($SD = 1.02$). The time to complete the task for incorrect answers is higher ($M = 3.98$, $SD = 2.71$) compared to correctly answered tasks ($M = 3.38$, $SD = 1.02$).

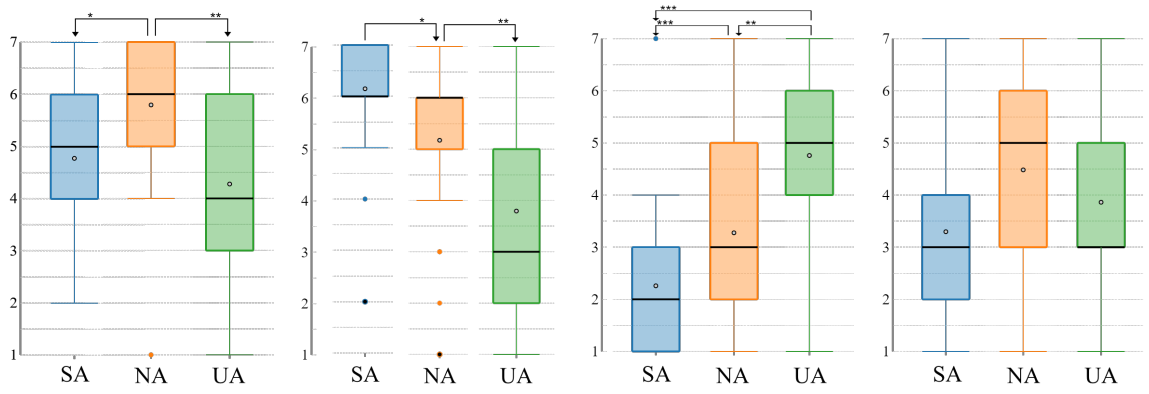
The *UnlockApp* enabled users to dismiss a learning task in case they did not want to answer it. During the seven study days, users chose to skip between zero and 80 tasks, with an average of 9.6 skips ($SD = 14.48$, $Md = 4$).

Favorites, Ratings, and Subjective Results

We asked participants directly which of our three applications they would prefer for everyday use. As their most favorite application, 18 out of 30 people named the *NotificationApp*, followed by seven mentions of the *UnlockApp*. Vice versa, 15 participants stated the *UnlockApp* as their least favorite of the three applications, followed by 13 naming the *StandardApp*.

Moreover, we analysed participants' subjective impressions of the three applications (expressed in the final questionnaire as 7-point Likert scale items from 1= "I totally disagree" to 7= "I totally agree"). For each item we performed a Friedman test with post-hoc Wilcoxon pairwise-comparisons. There was a significant difference in how much participants liked the applications (a: $\chi^2(2) = 6.812$, $p < .05$, see Figure 5.5a). Post-hoc analysis with Wilcoxon signed-rank tests showed a significantly higher score for the *NotificationApp* compared to both the *StandardApp* ($Z = -2.159$, $p < .05$) and the *UnlockApp* ($Z = -3.100$, $p < .01$).

We further asked participants (b) if they felt they used the app only when they had time to learn and (c) if they used the app even when they had no time to learn.



(a) Like (b) When Time (c) Even Without Time (d) Everyday Use

Figure 5.5: Participants' ratings of the *StandardApp* (SA), *NotificationApp* (NA), and *UnlockApp* (UA) for four questionnaire items (from 1= "I totally disagree" to 7= "I totally agree"). **(a):** I liked the [SA/NA/UA]; **(b):** I only answered tasks with the [SA/NA/UA] when I had the time to learn; **(c):** I answered tasks with the [SA/NA/UA] even when I did not have time to learn; and **(d):** I would continue using the [SA/NA/UA] frequently in my everyday life for learning a language. Asterisks indicate all statistically significant differences (* <.05| ** <.01| *** <.001).

Both items revealed significant differences in participants' subjective perception of their usage among all three applications (b: $\chi^2(2) = 16.568$, $p < .001$, see Figure 5.5b; c: $\chi^2(2) = 15.571$, $p < .001$, see Figure 5.5c). For item (b), participants report to be more likely to use the *StandardApp* when they had time to learn compared to the *NotificationApp* ($Z = -2.060$, $p < .05$) and the *UnlockApp* ($Z = -3.960$, $p < .001$), and more likely to use the *NotificationApp* when compared to the *UnlockApp* ($Z = -2.737$, $p < .01$). For item (c), participants report to be more likely to use the *UnlockApp* even when they had no time to learn compared to the *NotificationApp* ($Z = -2.786$, $p < .01$) and the *StandardApp* ($Z = -3.945$, $p < .001$), and more likely to use the *NotificationApp* when compared to the *StandardApp* ($Z = -3.945$, $p < .001$).

We found no significant differences in participants' ratings of the design of the three applications and their intuitiveness of use. Moreover, the Friedman test revealed no significant difference in terms of participants' willingness to continue using the applications in their everyday life. The descriptives, however, show a slight preference in favor of the *NotificationApp* ($M = 4.37$, $SD = 1.85$) compared to the *UnlockApp* ($M = 3.70$, $SD = 2.05$) and the *StandardApp* ($M = 3.30$, $SD = 1.73$), see Figure 5.5d.

Open Comments and Suggestions for Improvement

Regarding the concept of the *NotificationApp*, people reported positive as well as negative impressions: One participant considered the continuous presentation not optimal and rather wants to take the time to actively engage with language learning (P15). P7

perceived the display of the notification on the lockscreen as distracting. Two other participants suggested increasing the delay between the presentation of two vocabulary tasks so that the app is not running all the time but just presents a new word from time to time (P21, P27). On the other hand, some participants positively emphasised the unobtrusiveness of the notification app, saying it is "[...] *not distracting, it provides the opportunity to quickly solve a couple of learning tasks*" (P11) and that it is "*very suitable for daily use*" (P25).To further improve the application, participants wished for the inclusion of grammar knowledge (P2, P5, P14), free text entry of words to foster recall rather than recognition of words (P2), audio output to help with the pronunciation (P5, P8, P14), and gamification features (P6).

Similar suggestions were stated for the *UnlockApp*: Participants wished for free text entry opportunities (P22), pronunciation support (P9, P26), and grammar knowledge (P3, P5, P7, P19). Since vocabulary in itself is not sufficient to learn a new language, P6 considers the app a nice addition for people currently attending language learning courses and P19 recommends using it to freshen up a language. Regarding the overall concept of combining the learning task with authentication, many participants stated positive impressions: While P9 highlighted the simple design, P19 emphasises that "it's good that you always have to solve at least one task and therefore learn continuously". Further, P6 stated to like the app and in particular the idea, similar to P8. However, P8 adds that "[...] *it bothers me that I have to answer the task or dismiss the app when I just quickly want to do something on my phone*". P20 shares the experience and describes that when they just quickly want to use the phone, they are "*not focused enough to answer the questions conscientiously*". To address this issue, other participants suggested including more personalization options. For example, P7 wishes for a feature to adjust the number of vocabulary tasks presented at each authentication (which was fixed to one in the study), describing that if a user currently has more time to learn, then they could increase the number of words presented per unlock event. Further, P15 proposes to define time intervals during the day in which the app is active (P15).

5.4.4 Summary and Limitations

In this evaluation we compared three concepts for embedding vocabulary learning tasks into everyday smartphone use, either integrated into the authentication process (*UnlockApp*), as constant notification (*NotificationApp*), or as standard in-app learning (*StandardApp*). When similar ideas were proposed in the literature, these individual concepts had been evaluated independently of each other and not in a comparative fashion as performed in this work.

However, since our user study was conducted during the COVID-19 pandemic, this evaluation might include anomalies in users' daily routines, mobility, and smartphone usage due to lockdown and work-from-home phases. All these factors potentially in-

fluence the use of mobile language learning applications. Even with the *StandardApp* being a control condition, the generalizability of our results has to be seen with caution.

The focus of our evaluation lies on the users' experiences and users' interaction with the three apps of varying integration into smartphone usage. Although our applications do not constitute an exhaustive representation of all MLL applications, we chose these three as a concise comparison of how different levels of embedding could be implemented. Further, with our focus on user's interaction we can not report on actual vocabulary retention. However, as retention of vocabulary is increased by frequent interaction with the content [74, 86], we expect our applications to positively impact people's vocabulary recall and recognition. Nonetheless, actual vocabulary retention, particularly concerning the consolidation properties of the apps over time, needs to be further evaluated in future work.

By providing people the three applications to use for each seven days during this user study, we gained interesting insights into users' learning behavior. We decided in favor of a shorter within-subject study over a longer between-subject study to enable users to draw comparisons among their interaction with the three apps and to make the best use of the limited sample size. However, we are aware that our study only presents a narrow view onto users' actual behavior. As the users show a great diversity in individual preferences, only a long-term evaluation with a larger sample will be able to show which behaviours will prevail in users' daily lives. Our study data shows a more extensive usage of each application on the first day, declining over the seven days of usage. This pattern could indicate a form of curiosity or novelty effect. Educational technology research has shown that accustomisation with new learning technology can negatively impact learners' preferences regarding technology-based learning [187] and reduce users' motivation [174]. In contrast to the *StandardApp*, the *UnlockApp*'s concept partly counteracts this decline by continuously engaging the learner in solving vocabulary tasks with each authentication event. However, a future long-term evaluation, in particular, of the new *UnlockApp* concept is required to reveal the strength of the novelty effect in everyday usage.

5.5 Discussion

The preliminary evaluation of the user experience of authentication learning concepts as well as the comparative evaluation of embedded learning raised many aspects for discussion. While the ubiquity of smartphones in people's everyday lives offers great potential to increase frequent engagement with learning content, we experienced highly individual preferences around the authentication process and learning interactions. In the following, we will discuss the main issues that arose in our explorations of embedding learning into smartphone interactions.

Increasing Vocabulary Exposure

Our analysis confirms our initial hypothesis, showing that the *NotificationApp* and *UnlockApp* result in a higher number of solved vocabulary tasks per participant. Thus, we conclude that the learning content exposure was higher for these two apps when compared to the baseline, the *StandardApp*. Furthermore, the presentation of the vocabulary tasks after the authentication leads to a more spread out exposure to the tasks across the day. The similar distribution of correctly and incorrectly answered tasks across the three applications gives no indication that users might have been less focused when learning with the *UnlockApp* or *NotificationApp*. Additionally, we observed that users occasionally skip tasks with the *UnlockApp*. This suggests that even though the skipping requires the same amount of effort as selecting an answer (one button press), the users can judge if they have the mental capacity and/or time to engage in the learning or not.

Potential for Adaptation and Personalization

The subjective feedback revealed individual differences regarding people's attitudes toward the three applications: While many state to like the concepts of the *NotificationApp*, opinions on the *UnlockApp* are mixed. Specifically, some people felt distracted by the *UnlockApp*'s vocabulary presentation when they unlocked their phone with a specific task in mind. People do not oppose the concept in general but express their need for further personalization. Suggested adaptation features include the definition of learning time frames for the *UnlockApp* or adjusting the number of tasks presented with each authentication event. Moreover, we see potential for automated mechanisms that learn from people's interactions with the *UnlockApp* (in particular dismissals) and adjust the presentation accordingly. i.e., the app should stop promoting users with learning tasks after unlock events at times during the day when they are frequently dismissed.

Extending the *UnlockApp* Concept to Different Authentication Methods and Smartphone Actions

Based on the results here, we deem it useful to discuss possible extensions of the *UnlockApp* concept to further authentication methods: Concretely, the majority of participants in our study used fingerprint authentication to unlock (yet with phones that do not support fingerprint gestures). Our app and concept already support fingerprint gestures to respond to vocabulary tasks (see implementation section), which can help to embed the learning task even more implicitly into the authentication (i.e., no switch from fingerprint to touch required). Both touch input and fingerprint gestures might also be combined with unlocking methods that do not require further input themselves, such as face unlock, or when Android's smart screen lock feature is enabled (i.e., no unlock at home). Beyond this, we already gathered experiences with knowledge-based authentication methods: One third of our sample stated to use PIN,

password, or pattern authentication. In these cases, the *UnlockApp* also presents the task immediately after the unlock event.

Extending the Learning Features of Embedded Learning Applications

Our application focused on the presentation of vocabulary translations to reduce the complexity of the content. With this, we aim to simultaneously increase the control of potential effects among the three apps. The participants of our study strongly emphasised the demand for additional features such as grammar knowledge or pronunciation exercises in open comments, and compared our application to applications available on the market. To accommodate for people's need for more complex learning tasks, we see two potential solutions: (1) By combining the embedding concepts with a fully-featured learning app, we could address users' preferences for long learning streaks when they have time to spare (i.e., with the *StandardApp*), during which the app could also teach grammar or pronunciation knowledge. Additionally, the *UnlockApp* could refresh users' vocabulary knowledge by continuously presenting translation tasks after each authentication to engage users with a language on a daily basis. Option (2) is to investigate extensions of learning tasks that are short enough to be embedded into the *UnlockApp* or *NotificationApp*. These tasks need to be solvable with simple interactions and with low effort and time. Possible examples include fill-in-the-blank tasks offering a word in two tenses or with two suffix options.

5.6 Chapter Summary

This chapter explored the embedding of vocabulary learning tasks into everyday smartphone interactions. In the first part (Section 5.2), we explored concepts for including learning tasks directly into different authentication methods, as it is a highly frequent yet otherwise useless daily action. The results from our preliminary investigation showed that while participants appreciate the idea of being "nudged" to learn more frequently through the embedded authentication learning approach, they emphasize the need for a quick and simple interaction. If the task is too complex or slow, they can not imagine using it over a longer period. Based on these results, we implemented the *UnlockApp*, which presents a multiple-choice vocabulary task after each unlock event, and performed a follow-up comparative evaluation. The results of this evaluation show that the *UnlockApp* and *NotificationApp* designs, compared to the *StandardApp,* significantly increase the number of learning tasks users answer per day. However, other concepts were favored in situations where participants unlocked their device with a certain goal in mind (*NotificationApp*) or when they aimed for a longer learning streak with more complex learning content such as grammar (*StandardApp*). We end this chapter by discussing implications for embedding language learning tasks into everyday smartphone interactions to increase exposure and prevalence.

6

Embedding Comprehension Assessment into Digital Reading and Listening

The prevalence of the internet gave rise to a steady growth of media content available for everyone in a variety of languages. Especially for studying English, movies, TV series, or audiobooks which are available at streaming services such as Netflix[1] or Amazon[2] are a common tool to improve one's language skills. By changing the audio track of a movie and enabling subtitles, media content can support effective learning [134, 360]. In particular, subtitles have shown to reduce learners' cognitive load during video consumption [189].

Besides being a convenient tool for learning, which is accessible anytime and anywhere, media content also represents the user's interest and hence can increase learning motivation [263]. This is in contrast to the concept of language learning classes, which predetermine learners' schedules and learning content. Media content ensures a high degree of language exposure and can provide interactivity as in pausing and rewinding certain scenes [278]. This interactivity is, in particular, necessary when learners encounter vocabulary they do not understand. However, interacting with translations as in the system proposed by Ma et al. [221]) can interrupt the media experience. When requesting translations on a separate device, it can even lead to "media multitasking". When engaging with more than one medium at once, the effort of multitasking can lead to a decreased recall of the presented content and a worse understanding due to higher cognitive load [337]. Thus, it is likely that *unknown* words are skipped to continue watching the movie or listening to the audiobook, trying to ensure the overall text comprehension.

If we want to support learning of new vocabulary with media content, it is necessary to implicitly assess a person's knowledge gaps [259] without active user intervention. We can assess those knowledge gaps and use them to provide effective learning support by monitoring a user's understanding while engaging with second-language content. Through adaptations in the User Interface (UI) (e.g., lowering the speed of a speaker in an audiobook) we could provide technical support to facilitate comprehension and learning. Moreover, by evaluating a person's comprehension, we can generate personalized learning content for additional post-hoc repetition, targeting exactly those vocabulary the user is struggling with.

In the last two decades, the Implicit Personalization assessment of comprehension by the use of physiological sensing became increasingly researched [28, 29, 117]. For the

[1] Netflix: `www.netflix.com`, last accessed January 3, 2022

[2] Amazon: `www.amazon.com`, last accessed January 3, 2022

estimation of comprehension during reading, eye tracking can give insights on people's understanding. In the context of HCI, eye-gaze analysis has been previously evaluated to assess a learner's language proficiency (cf. [18, 28, 292]). Although eye-gaze analysis already presents a feasible approach for language proficiency assessment [170], this method is limited to visual content presentation. Hence, the evaluation of comprehension during the perception of audio content is not possible. A mechanism that gained popularity in the last decade and has the potential to be applied for implicit assessment of comprehension across multiple modalities is Electroencephalography (EEG). EEG has been used to evaluate language processing (cf. [195, 196]) and proved its potential as implicit input for HCI applications (cf. [133, 310]). EEG has become increasingly robust and easier to handle with the availability of prototypes embedded in caps or glasses to enable evaluation in real-world scenarios [34, 82, 343]. Thus, this chapter aims to answer the following research question:

RQ3: How can we utilize users' everyday reading or listening activities to generate personalized language learning content?

This chapter is based on the following publications:

- Schneegass, C., Kosch, T., Schmidt, A., and Hußmann, H. (2019). Investigating the Potential of EEG for Implicit Detection of Unknown Words for Foreign Language Learning. In *IFIP Conference on Human Computer Interaction* (*INTERACT'19*), pages 293-313, Springer. DOI: 10.1007/978-3-030-29387-1_17 [303]
- Schneegass, C., Kosch, T., Baumann, A., Rusu, M., Hassib, M., and Hußmann, H. (2020). BrainCoDe: Electroencephalography-based Comprehension Detection during Reading and Listening. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. DOI: 10.1145/3313831.3376707 [302]

Some text passages were taken verbatim from these publications. Further, this chapter is supported by the Bachelor theses of Marius Rusu and Andrea Baumann, see detailed collaboration statement at the beginning of this thesis.

6.1 Related Work

6.1.1 Electroencephalography

By measuring electric potentials through electrodes on the scalp, EEG can give insights on a plethora of users' internal processes, such as engagement, workload, attention, fatigue, emotions, flow, or immersion [23, 116]. The evaluation of EEG signals can be performed based on frequency bands or ERPs [219]. The latter refer to changes in signal amplitudes occurring at a precise and consistent time after the presentation of a stimulus [78, 127]. The stimulus triggering an ERP can be motory, visual, auditory, or of any other sense (e.g., hand movements or perceiving audio).

Although EEG has been initially developed for medical applications and required high precision and accuracy, technological advancements within the last decade of both software and hardware have made it attractive for HCI applications [245]. While we still rely on medical-grade hardware and software to explore the feasibility of EEG for specific problems or approaches, researchers have already built a variety of increasingly small, wireless, and low-cost sensing devices for specific applications in everyday scenarios [83]. With research prototypes using printed electrodes connected to portable EEG devices such as the ones used by Debener et al. [82] or Bleichner et al. [34], the integration of EEG in users' everyday context does not seem out of reach anymore. For example, Bleichner et al. [34] showed that they could achieve reliable measurements of specific ERPs. They were able to detect P300s, negative potentials that are reactions often occurring after surprising and unexpected events [267] and related to memory and attention processes [265]. To achieve this, they integrated miniaturized EEG electrodes into a baseball cap and an additional customized earpiece [34]. In a different approach, Vourvopoulos et al. [343] modified a regular pair of glasses to include a low-cost EEG device, the OpenBCI[3], for future use in head-mounted displays. Their work shows promising first results in the investigation of cognitive and sensorimotor tasks by evaluation of frequency bands.

6.1.2 Event-Related Potentials

ERPs can be used to evaluate brain activity during second language processing. These potentials are averaged responses from a group of trials as a reaction to a given experimental stimulus. The assumption is that a certain electrical potential occurs at a consistent time after the presentation of a stimulus to the participant [78, 127] (e.g., a non-word or a word out of context). The application of ERPs to analyze problems during language processing has already been researched extensively in the neuroscience community. Syntactic and semantic problems during reading characteristically elicit N400 ERPs, whereas the N100 ERP is often indicating responses to auditory stimuli.

[3] OpenBCI: www.openbci.com, last accessed January 3, 2022

Changes in amplitudes of these ERPs provide insights on various language processing problems based on individual words or sentence structures. Due to the high temporal resolution, ERPs can reflect responses occurring within a few hundred milliseconds after the stimulus is presented. Thus, ERPs can provide insights about the processing of individual words within sentences. Since the potentials of consecutive words can overlap, serial presentation of words can foster effective detection of ERPs. Either slow rate serial presentation or artificial separation of the words can maintain the correct mapping of stimulus and response [78].

N400 An ERP component that is interesting for the evaluation of language processing and semantic relationships of words is the N400 component. An N400 is a negative deflection of the EEG signal around 250-500 ms (i.e., peaking at about 400 ms) after the presentation of a stimulus [196]. The N400 has shown to reflect on problems during semantically integrating a word into a sentence during reading (cf. [126, 195, 244]) during both visual and auditory word pair and sentence processing [149, 150]. Kutas and Hillyard [196] showed participants reasonable sentences, containing either a word fitting the context or a word that was syntactically correct but semantically incongruous Examples included "*They wanted to make the hotel look more like a tropical resort,* [...] *so they planted [tulips/ palms].*". When reading the word "tulips", the authors report higher N400 in participants' neural responses. Additionally, Holcomb and Neville [150] showed higher N400 amplitudes for the processing of non-words or pseudowords (e.g., "jank", "grusp", "kcsrt") as compared to regular words. For the evaluation of foreign language reading comprehension, Schneegass et al. [303] employ N400 analysis and show significant differences between *known* and *unknown* words during reading. However, this approach is limited to presenting one stimulus at the time and to visual text presentation.

N100 The N100 component is frequently evaluated for the processing of auditory stimuli [277]. It is a typical component responding to the onset of a perceived sound with a negative deflation around 100 ms after the stimulus. It can occur in combination with a P200, an increased amplitude of the signal around 200 ms after a stimulus [277, 351]. The N100 is *known* to be an indicator of the auditory "oddball" effect, which occurs when participants are presented with a set of familiar stimuli, followed by an unexpected stimulus [246]. Zhang et al. [361] investigated the N100-P200 complex for the audio presentation of pseudowords and were able to show significantly stronger negative responses as compared to regular words.

Research Gap: While research in the HCI community has investigated methods to assess general language proficiency (cf. [18, 28, 170, 292**?**]), it has not yet brought forward a method to extract individual unknown words. As this is necessary to create a personalized vocabulary learning set, we draw on extensive research using neurological responses for the detecting of semantic and syntactic incongruities (cf. [150, 195, 244, 277]. We adopt ERPs as a method, priorly used to uncover said incongruities, and

investigate its potential to also detect second-language vocabulary the user can not translate.

6.2 EEG for Word-Based Reading Comprehension Assessment

In this section, we investigate the potential of EEG for the implicit detection of gaps in users' vocabulary knowledge for learning foreign language contents. In particular, we evaluate ERPs [37] to differentiate between known and unknown words in English second-language reading along with a native language baseline. In our experiment, we presented the texts as a Rapid Serial Visualization Presentation (RSVP) approach, which displays text as one word at a time [208].

6.2.1 Methodology

In a first evaluation, we aim to investigate the occurrence of N400s as indicator to uncover gaps in users' second language vocabulary knowledge during digital text reading. We conducted a lab study in which participants were required to read texts on a computer screen while recording their EEG signals. In particular, we investigate the following hypothesis:

H_1 Unknown words will result in higher mean N400 amplitudes compared to known words.

This section will outline the methodology we applied in more detail and present the results of the user study.

Text Difficulty Selection

For this study, we included texts from the corpus of the *Asian and Pacific Speed Readings for English as Second Language (ESL) Learner* [270]. These texts include predefined English language texts on topics related to Asia and the Pacific with a supplementary set of ten single-choice comprehension questions per topic. The texts was specifically chosen because it features frequent words and easy grammar [271] to be easily understandable. We chose to include excerpts from three texts ("Life in the South Pacific Islands", "Buddhism", and "Hong Kong") and translated the first one (further termed *N1*) into the participants' native language to serve as a baseline. N1 included 30 sentences and in total 452 words, which we split into two texts of 15 sentences each. The presentation of either subset was randomized among participants to avoid effects caused by the content.

Table 6.1: Overview of the Lexile measures [210] for E1 and E2 in both their original and revised versions utilized for the study.

	$\mathbf{E1_{orig}}$	$\mathbf{E1_{rev}}$	$\mathbf{E2_{orig}}$	$\mathbf{E2_{rev}}$
Total Sentence Length	29	29	24	24
Lexile Measure	600L - 700L	900L - 1000L	1000L - 1100L	1000L - 1100L
Mean Number of Words per Sentence Length	14.96	15.34	17.36	17.36
Total Word Count	419	445	434	434

The second and third text, named *E1* and *E2*, were in English, the participants' second language. E1 contained 29 and E2 24 sentences (∼450 words per text, for more details see Table 6.1) to generate a sufficient set of trials while not straining the user. The two texts E1 and E2 were randomly assigned to the participants for a within-subject design. We revised each text to contain ten sentences with one uncommon word (e.g., "adscititious"), selected with the help of a thesaurus and a list of unfamiliar words[4]. Including just one difficult word per sentence creates a realistic scenario and prevents the overlapping of ERPs. In regards to the changes performed in the texts, we adapted the comprehension questionnaires for E1 and E2. Each question is meant to check the understanding of one sentences containing a potentially unknown word.

To confirm the difficulty level of the texts, we used the *Lexile Analyzer*[5]. This tool analyzes texts and provides an approximate reading level for it based on the metrics (1) word commonness, which is reported to correlate highly with text difficulty, and (2) complexity of syntax [210]. The Lexile score can range between 200L (L for Lexile) for beginner reading, up to 1700L for advanced texts [210]. Table 6.1 specifies the Lexile Measures for E1 and E2 in the original version and a revised version that includes unknown words. It can be seen that the effect of difficult words on the Lexile score is only noticeable in E1, since E2 already includes many proper names and consists of longer sentences. Since the Lexile Analyzer only supports English texts, we further confirm the understandability of our texts through subjective post-hoc ratings. Participants had to answer comprehension questions as well as specify every word which they could not translate.

[4] Oxford Lexico's Weird and Wonderful Words List: `https://www.lexico.com/explore/weird-and-wonderful-words`, last accessed January 3, 2022

[5] Lexile Framework: `www.lexile.com` - last accessed January 3, 2022

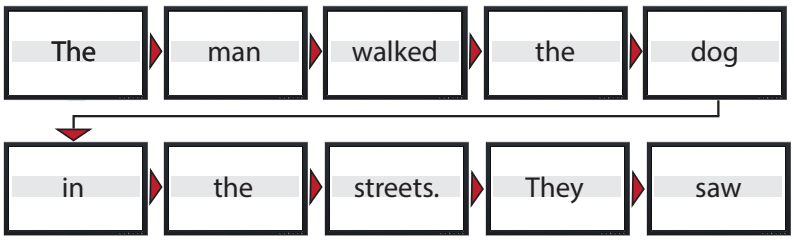


Figure 6.1: In the RSVP approach, text is displayed each one word at a time. The words are centered to minimize eye movements.

Text Presentation

We presented the texts in a RSVP mode, showing each one word at a time on the screen to reduce saccadic eye movements during normal reading behavior [286] (see Figure 6.1). A decrease of eye movements will lead to reduced noise in the data generated by the muscles around the eye. Furthermore, since RSVP only displays one word, the matching of the EEG signal to the dedicated stimulus can be easily performed. The word presentation rate was set to 170 Words per Minute (WPM) based on the findings of [59], who showed that participants' reading comprehension is best at speeds ranging from 171 to 350 WPM. We decided to set the speed to the lower end of this spectrum since our participants are non-native speakers and to minimize overlaps in the signals due to the processing of consecutive words (cf. Section 6.1.2).

Apparatus

Our setup consisted of a display (Dell U2715H; 27 inches; 60 Hz refresh rate) and a Brainvision Liveamp[6] EEG device comprising a sampling rate of 500 Hz. The EEG device provides a bandpass filter ranging from 0.1 to 1000 Hz and does not include a notch filter. Electrodes were placed in accordance to the International 10-20 layout (ground electrode: *Fpz*, reference electrode: *FCz*; see Figure 6.2).

Procedure

At the beginning of the study, we welcomed the participants and handed them a detailed study description. The description contained the study motivation and goal as well as its process. Every participant signed a consent form and randomly picked and crossed out an ID from a prepared sheet of possible IDs to assure adequate anonymity. Then they filled in a short demographic questionnaire asking for age, gender, highest educational degree, vision impairment, and neuronal diseases or disorders. Furthermore, they were asked to set all electronic devices into flight mode so as not to influence the data recording.

[6] Brainvision Liveamp: `www.brainproducts.com` - last accessed January 3, 2022

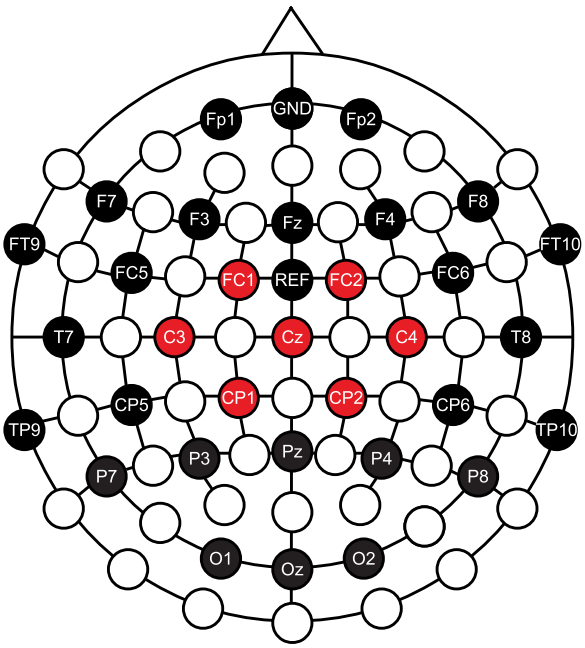


Figure 6.2: EEG electrode layout: the red electrodes located around the parietal lobe were used for analysis.

We carefully explained the EEG system, measured participants' head circumference to select between four different actiCAP sizes (54 cm, 56 cm, 58 cm, and 60 cm), and instructed them to put on the cap. Afterward, we attached 32 electrodes (plus 2 reference electrodes) to the participants' scalp using the actiCAPs' designated 10/20 positioning system [160] (the electrode layout can furthermore be seen in Figure 6.2). We increased the conductivity of all 34 electrodes with high viscosity electrolyte gel and examined their impedance for reliable performance. The experiment started when the impedance of all electrodes reached the threshold of 10 $k\Omega$. The signals were recorded in a quiet, dimly lit experimental room equipped with a desk and a comfortable chair, around 80 cm away from the screen.

At first, participants had to read one subset of the N1 baseline text (15 sentences). Successively, participants had to read one of the two English texts E1 and E2. In total, we recorded participants reading 29 (E1) / 24 (E2) sentences, generating recordings of, depending on the randomization, between 434 and 445 individual words per participants (see Table 6.1).

To evaluate how much additional perceived workload was induced by the second language texts, we presented each participant with two NASA-TLX questionnaires [130, 131], one after the baseline and one after the foreign text. Both texts were followed by a comprehension test consisting of ten questions designed to target the understanding of the sentences that included unknown words. Furthermore, to confirm that participants could not translate the words that were meant to be unknown, we

presented them with a print-out version of the text. We asked them to highlight all the words, which they cannot translate to their native language. In summary, for both texts the following procedure was applied:

1. Read text as RSVP
2. Answer NASA-TLX for this text
3. Fill in ten item comprehension questionnaire
4. Post-hoc rating of unknown words in printed text
5. Short rest phase

Sample

We recruited twelve participants via a university mailing list and an internal social media channel. As a requirement, we asked for our German and English proficiency. A minimum English proficiency of B1 according to the Common European Framework of Reference for Languages (CEFR)[7] was given due to the standards of the german high school diploma. Furthermore, we did not include participants with severe vision problems and neurological disorders. Every study participant was rewarded with a voucher for an online shop. We removed two participants from our evaluation due to technical difficulties.

Within our adjusted sample size of $N = 10$ (5 female, 5 male), the participants' age ranged from 18 - 60 ($M = 31.6$, $SD = 14.41$). They held at least a high school degree (6), some even a master degree (3), or a doctoral degree (1). Due to high school being the lowest minimal educational level of our sample, we can assume a minimum English proficiency level of B2 [72] or more for every participant.

Data Processing

We use Python with the library MNE to process the recorded EEG data[8]. EEG data were bandpass filtered [268] (0.5-40 Hz) to attenuate the influence of artifacts (e.g., blinks, eye, and head movement) as well as the 50 Hz remote power line noise. We consider the electrodes *Cz*, *C3*, *C4*, *CP1*, *CP2*, *FC1*, and *FC2* as the parietal lobe is linked to the processing of spoken and written language [347]. We identified eye blinks using Python MNE and removed them manually. We did not perform an independent component analysis as we employed 32 electrodes and we did not intend to remove the contribution of cortical components that might have been resolved into the summed activity of non-cortical dipoles. To analyze ERPs independently from known and

[7] CEFR: `www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions` - last accessed January 3, 2022

[8] `www.martinos.org/mne/stable/index.html` - last accessed January 3, 2022

unknown words, we slice the data set into triggers for known and unknown words. We look at the first second of neural responses for each word as we are interested in investigating the N400 ERPs that occur between 300 ms and 600 ms after displaying the stimuli.

6.2.2 Results

We statistically analyzed the collected data for differences in ERP magnitudes. We submitted the magnitudes of the averaged N400s for known and unknown words to an ANOVA. Furthermore, we investigated the subjectively perceived workload and reading comprehension. Our results contain the EEG responses to 4390 words, of which 100 are classified as likely to be unknown to the users.

Event-Related Potentials

We divided the measured data into each epoch for known and unknown words. Each epoch has the same duration as a single word is displayed on the screen. We averaged each epoch for every participant and normalized the magnitude of the data to enable person-independent comparisons for native, known, and unknown words. Mauchly's test did not show a violation of sphericity. A repeated measures ANOVA was performed including the three conditions native, known, and unknown words as independent variables and the N400 amplitudes as depending variable. The analysis revealed a significant main effect ($F(2,18) = 13.33$, $p < .001$). A post-hoc test using a Bonferroni correction revealed a significant effect in N400 potentials between known and unknown words ($p < .001$, $d = 2.648$) as well as unknown and native words ($p < .05$, $d = -1.024$). We found no significant effect between the amplitudes of native and known English words. Figure 6.3 shows the averaged N400 across all participants for known and unknown words. The mean amplitude for known words was higher ($M = -1.201$, $SD = 0.662$) compared to the mean amplitude of unknown words ($M = -2.885$, $SD = 0.057$). Figure 6.4a illustrates the difference of the N400 magnitudes for known and unknown words.

Perceived Workload

The workload during the reading of native text was perceived as lower as during English texts. The NASA-TLX is subdivided into six facets of workload: mental, physical, temporal, performance, effort, and frustration [131]. The NASA-TLX was presented as a detailed scale ranging from one (very low demand) to 20 (very high demand). For a general analysis, we added up the individual facets of the NASA-TLX to create one overall score, which therefore had a range from 1 to 120 (6×20, the max value of one facet). This score was lower for native language text ($M = 40.4$, $SD = 19.08$) as opposed to the foreign text ($M = 53$, $SD = 18.66$). We performed a paired samples t-test comparing the overall perceived workload of the two languages (native vs. English).

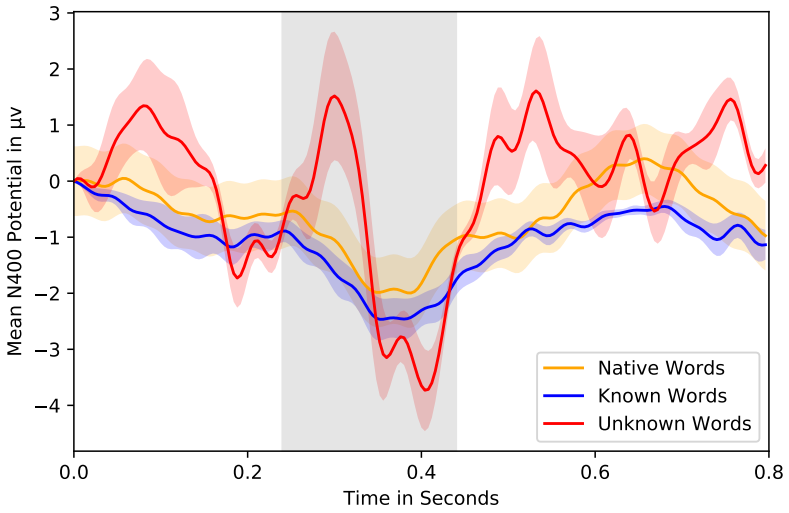


Figure 6.3: N400 measured for known and unknown words. A larger mean amplitude is measured for unknown words compared to known words.

The results revealed a significant difference between the languages ($p < .05$, $t(9) = 2.842$, $d = -0.899$), showing a large effect in the direction of a lower mean workload in the native language texts. A Shapiro-Wilk test showed no indication for a deviation of normality ($p > .05$).

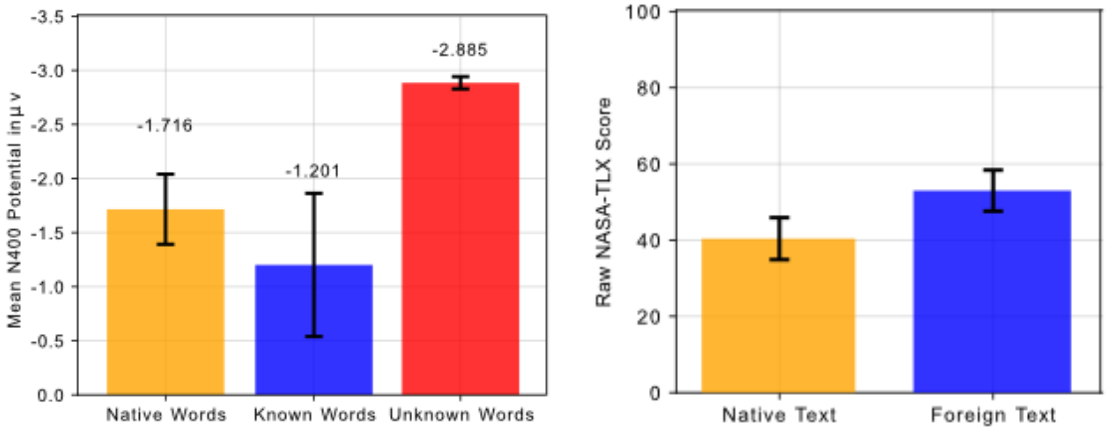
Moreover, paired samples t-tests within the individual facets showed significant differences between the two languages in terms of perceived mental workload ($t(9) = -3.452$, $p < .05$, $d = -0.506$), perceived temporal demand (i.e., feeling rushed; $t(9) = -2.339$, $p < .05$, $d = -0.740$), and perceived performance (i.e., reading and understanding the texts; $t(9) = -2.872$, $p < .05$, $d = -0.899$). All of these results showed negative values for Cohen's d, indicating higher loads for the English texts. Figure 6.4b shows the mean raw NASA-TLX score between both languages.

Text Comprehension

There was each one questionnaire to test the comprehension of the N1, E1, and E2 texts. On average, the participants achieved the best scores in the N1 questionnaire with around 7 correct answers (correct answers out of ten, $M = 6.9$, $SD = 2.13$), followed by E2 ($M = 6.6$, $SD = 1.82$). The text with the least amount of correct answers was E1 ($M = 5.4$, $SD = 2.19$).

Post-hoc Word Review

Participants performed a post-hoc rating of all words, highlighting each word for which they can not come up with a translation. No participant highlighted any word of the N1 text. We can furthermore confirm the low difficulty of the E1 and E2 due to



(a) (b)

Figure 6.4: (a): Mean N400 amplitudes for known and unknown words. Unknown words elicit a statistical significant effect in amplitudes compared to known words. The bars depict the standard error; **(b):** Mean raw NASA-TLX scores for both languages. Reading native languages resulted in less workload compared to foreign languages. The bars depict the standard error.

the fact that not a single word of all three texts was perceived as difficult besides our artificial modifications. When looking at the 10 modified words of E1 in detail, the results show a high consensus among the participants who read the text. Out of the 10 potentially unknown words, 7 are confirmed as unknown by all five participants ($M = 4.6$, $SD = 0.7$). In the text E2, the confirmation of the unknown words turned out to be less distinct. Only three out of the 10 words were highlighted by all participants. On average, the potentially unknown words are highlighted by 4.2 participants ($SD = 0.63$).

6.2.3 Discussion

We conducted a user study to investigate the feasibility of ERPs to detect vocabulary gaps. Our results show a statistically significant main effect in N400 amplitudes between known and unknown words as well as between native and unknown words. In the following, we discuss the implications of our results.

Limitations A major challenge of this approach is the sum of influencing factors, which would reflect in the EEG data when applied in a real-world setting. We minimized these effects by conducting a study in a laboratory setting and were able to control people's attention and task load. Therefore, we do not know to what extent our results are generalizable to other situations. Including additional measurements such as gaze tracking could help to compensate for other influences in the application of EEG data

in everyday settings. Furthermore, we have to examine in small steps the potential of this approach when faced with further stimuli (e.g., by adding video or auditory material). The applicability for other text presentation modes, for example, including a sentence-based text presentation mode as an approximation of subtitles used in videos, needs to be evaluated. Furthermore, we acknowledge that our study employed a low sample size. However, we replicated the methodology from other HCI studies that have successfully employed similar sample sizes [132, 310] and therefore, believe that this study can highlight the potential of EEG for implicit language proficiency detection. We see our work as a first proof of concept and as to be following the path of other EEG research publishing novel ideas for real-world scenarios [133, 290].

Detecting Vocabulary Gaps The results from our study show that we can use EEG data, N400 ERPs in particular, to assess the word-based language proficiency by measuring a significant effect between the amplitudes caused by known and unknown words. This confirms EEG as a valid tool for vocabulary gap detection. Although the descriptive ERP data showed minor differences between the N400 amplitudes of reading native words and reading known second-language words, the statistical analysis did not result in a significant difference. We conclude that N400s, independent from the presented language, have common properties [338]. Going beyond the results of the conducted study, our results encourage further investigations on neural activity of second language processing.

Subjective Workload and Comprehension When comparing the results of the raw NASA-TLX questionnaires we recognize a significant difference in perceived workload when reading native as opposed to English words. This shows that subjectively perceived workload was manipulated in conjunction with our finding in N400 amplitude. The comprehension tests show that there is a difference of complexity when comparing the two English texts. However, when marking the unknown words in the print-outs, the participants reached a 70% consensus. Therefore, we assume that the participants perceived the unknown words as equally difficult to translate in both English texts. We infer that the text comprehension rate is affected by the text difficulty, however, this does not have an impact on the individual N400 measures.

Participants achieved lower text comprehension scores for the text E1 compared to E2 although E2 is more difficult according to the Lexile score. Since the Lexile scores takes several syntactic and semantic factors into account to calculate the text difficulty, a potential threat to validity is posed by participants that were familiar with the text. However, we have observed the effect of N400 for most of the words that were unknown for participants. Thus, we believe that the overall text difficulty does not represent a higher occurrence of N400s.

Differences in other ERP characteristics Besides the characteristic N400 amplitude differences, Figure 6.3 revealed a set of further differences in the EEG signals. Reading

foreign known and unknown words seems to reflect in increased positive amplitudes at around 100ms as well as around 300 ms after stimulus onset. In addition, the signal received for unknown words shows a higher mean potential at around 500-600 ms after the stimulus. The latter could reveal a P600 [194] induced by syntactic continuation problems or checking upon unexpected (linguistic) events [180]. The P600 is related to the P300 [71], which can be a result of the oddball effect discussed in Section 6.1.2. The oddball effect is a phenomenon of inattention blindness and can occur if an unexpected stimulus appears [320]. In our case, the unknown words suddenly interrupted the fluent reading behaviour. It is common, that a P300 occurs simultaneously with an N200 [71]. Further statistical analysis need to evaluate the differences of other ERP components in the recorded signals.

6.2.4 Summary

We conducted a user study to investigate the feasibility of ERPs to detect vocabulary gaps. Our results provide evidence that EEG has the potential to uncover comprehension problems in a controlled reading scenario. We find a statistically significant main effect in N400 amplitudes between known and unknown words as well as between native and unknown words. Although this work supports the assumption that ERPs have a high potential for vocabulary learning support, further evaluation needs to clarify the feasibility during less controlled reading scenarios and the generalizability toward other modalities such as listening. We will discuss the current challenges and limitation of this technology for evaluation in the wild in Section 6.3.3 and portray three use cases of potential application scenarios of our approach in everyday settings (cf. Section 6.4.

6.3 Extending the Approach to Sentence Reading and Listening

In the first part of this chapter, we presented a first evaluation of EEG for the detection of vocabulary-based incomprehension during foreign language RSVP reading. The objective of this follow-up work is to show that we can extend this method to sentence-based text presentation in reading (as compared to RSVP) and to narrated content such as audio books.

We conduct a user study, presenting participants with foreign language content using (1) text on screen (i.e., visual presentation) and (2) verbal narrations (i.e., auditory presentation). Similar to the first study, participants read and listen to English texts which we manipulated to contain several potentially *unknown* words while recording their neural responses. We hypothesize that this manipulation will provoke a measurable neural reaction through greater amplitudes in the N400 reading and N100 while reading or listening, respectively. Furthermore, we show our process of classifying

the neural responses we collected while participants encountered *known* and *unknown* words. By applying this classifier to a subset of our data, we calculate the accuracy and assess the potential of our approach for further use as a real-time comprehension detection tool. We will refer to our approach in the following as *BrainCoDe* (Brain response-based Comprehension Detection) method.

6.3.1 Methodology

Apparatus and Setup

We placed participants in a quiet, dimly lit room to reduce the risk of potential distractions. Participants sat at a fixed distance in front of a 24-inch desktop screen and we recorded their neural activities with a 32-channel EEG. We presented the text content on the screen and used a supplementary eye-tracker to assess the users' focus of attention for the reading trials. Thus, we were able to match the EEG responses to individual words on the screen. We manipulated the texts carefully to contain several potentially *unknown* words. After each condition, we confirmed the manipulation (i.e., the comprehension of the vocabulary) using comprehension and translation questionnaires.

EEG and EOG Recording To record the electric potentials generated by participants' brains, we used a Brain Products Live AMP[9], an EEG device with a 32 channel wireless electrode setup. The sampling rate was set to 500 Hertz (Hz) and the signal was automatically bandpass filtered between 0.1 and 1000 Hz. The electrodes were placed according to the 10-20 layout [160] (ground electrode: Fpz; reference electrode: FCz; see Figure 6.5). Conductive gel was used to reduce the impedance between electrodes and scalp. We ensured that the impedance was set to below 10 $k\Omega$ before starting the experiment.

For the evaluation of the EEG signals, we set four different markers in the software during the EEG recording to map the neural responses to the particular word shown on the screen. Those markers encoded the beginning of a text (marker "1") and a word's estimated difficulty. We differentiated *known* words (marker "2"), potentially *unknown* words (marker "3"), and words that were excluded from the analysis (marker "4"). The latter category contained words to be excluded from our evaluation for two reasons: (A) words that are shorter than three characters as users often skip them during reading [275, 276] and (B) proper names that do not necessarily have a translation since these are difficult to understand in the audio presentation. All markers were encoded into the EEG signal as a simulated keyboard input with a frequency of 8000 Hz.

[9] Brain Products Live AMP: `www.brainproducts.com/productdetails.php?id=63`, last accessed January 3, 2022

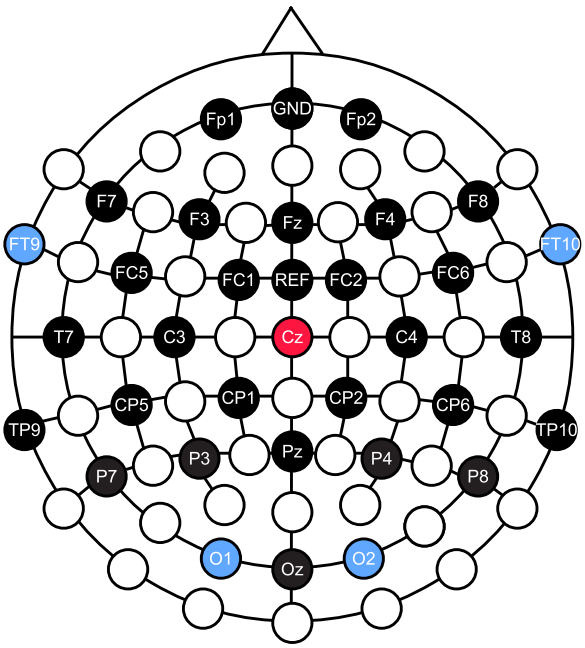


Figure 6.5: For the analysis, we used the Cz electrode (red) as positioned in the 10-20 layout [160]. For the measurement of EOG, we utilized the FT9, FT10, O1, and O2 electrodes (blue) and placed them around the participants' left and right eye.

From our 32-channel EEG setup, four electrodes were used for Electrooculogram (EOG). EOG is used to record electric signals caused by muscles around the eye and works as an indicator of eye movements. Since eye movements are inevitable during reading, EOG enables us to filter the noise generated by muscles to create a clean recording of the actual brain responses. To record EOG, the electrodes were placed on the right and left canthi as well as above and below the left eye as suggested by prior work [122] using adhesive tape for medical use. We chose four electrodes from our setup (right eye: FT10; left eye: above: FT9, side: O2, below: O1), which are least likely to show responses to language processing, i.e., with the great distance to the central parietal area [195]. The remaining 28 electrodes were used for EEG recording.

Gaze Tracking The EOG enables us to filter signals generated by muscles during the reading movements of the eye. However, it does not tell us the user's focus of attention. Knowing which word the user is focusing on is necessary to precisely map the resulting EEG response to the word that caused it. When presenting more than one stimulus at the same time, such as when presenting multiple words on the screen, eye-tracking can be used to map the gaze and, thus, the brain's focus to an individual word. In our setup, we used an EyeLink 1000+[10] which utilizes a video-based recording of eye gaze at 1000 Hz and was calibrated for each participant. A chin-rest is used to avoid

[10]EyeLink 1000+: `www.sr-research.com/products/eyelink-1000-plus`, last accessed January 3, 2022

Table 6.2: We evaluated two native texts, a set of individual words, and four full texts. This table outlines the number of words and sentences per condition as well as the number of difficult words we induced.

	German		English					
Trial Phase	Native Baseline Ge1	Native Baseline Ge2	Indiv. Words IW1	Indiv. Words IW2	Full Text En1	Full Text En2	Full Text En3	Full Text En4
Total Number of Sentences	15	15	-	-	41	46	46	44
Total Number of Words	214	238	30	30	514	517	542	540
Total Number of Difficult Words	0	0	15	15	15	15	15	15
Duration Audio Narration (in min)	1:42	1:28	3:18	3:13	3:12	3:24	3:21	3:10

re-calibration of the eye-tracker and maintain a fluent reading experience. Considering real-life settings, we expect reading to generate few to no head movement and are confident that valid gaze detection can be achieved without a chin-rest (e.g., use of a head-mounted eye-tracking device). Our implementation tracked participants' gaze and annotated the EEG signal accordingly.

Text Presentation In the first part of our study, we assessed the participants' comprehension during *reading* on a computer screen. Texts were presented as a single centered line of black text on a grey background with a font size of 25. We set the maximum number of characters presented in a row to 40 to approximate subtitles in a movie while still being easily readable. If a sentence exceeded the character limit, the sentence was split. If a word would have been split due to the character limit, we pushed the whole word to the next line presentation instead. The system was designed to adapt to participants' reading speed: The next sentence was presented on the screen as soon as the eye-tracker recognized a short fixation of every word with at least three characters. By choosing a fixation time of only 1 ms, we ensured the recognition of each word while maintaining a natural reading flow.

In the second part of our study, we analyzed comprehension during *listening*. For the listening trials, we created narrated speech files using the Google Cloud Text-to-Speech (TTS) engine[11]. In contrast to a human reader, the engine ensured the

[11]Google Cloud TTS Engine: `https://cloud.google.com/text-to-speech`, last accessed January 3, 2022

creation of comparable texts and easy manipulation of individual words. For the speech presentation, we chose the default options suggested by Google's TTS engine, namely a female voice and American pronunciation. We reduced the speaking speed to 90% of the default for easier comprehension and less overlap across the neural reactions in the EEG signal. After downsampling the data to 1000 Hz, we analyzed the amplitudes of the resulting audio file to detect the beginning and end of the spoken words and list their timestamps. The resulting file format contained a list of all words, a timestamp for start and end, and the duration in ms (e.g., "cucumber", 2.258, 2.954, 0.696). In analogy to our process of encoding word markers into the EEG signal for the reading materials, we annotated the narrated words also in the list of spoken words, with the markers 1 (start of text), 2 (*known* word), 3 (*unknown* word), and 4 (excluded from analysis). To validate the accuracy of the timestamps, we visually analyzed the resulting audio file with an open-source audio software. During the listening trials, the screen displayed a red dot for participants to focus their gaze on to reduce unnecessary eye movement [326]. We used consumer earbud headphones to deliver the narrated texts.

Text Materials and Manipulation To ensure the comparability of the two modalities reading and listening, we used standardized textual materials by Quinn and Nation [270], Quinn et al. [271]. Besides a native language baseline, which we used to familiarize participants with our setup and text presentation, we assessed neural responses to individually presented English words and four English full texts, two for each modality (further explained in Section 6.3.1). All texts used in this study were taken from a corpus designed for ESL Learners that includes 15 texts on various topics and complementary multiple-choice comprehension questionnaires [270]. The texts are designed to have easy grammar and feature frequent words [271]. Therefore, they are supposed to be easily understandable by participants with low-level English skills. For the listening trials, the written texts were transformed into narrations with a TTS engine as explained above.

To evaluate vocabulary gaps during reading and listening, we manipulated the texts to include a number of rare and potentially *unknown* words. Hence, we incorporated words from lists containing rare or uncommon words in the English language, which are difficult even for a native speaker (e.g., Oxford Lexico's "Weird and Wonderful Words" List[4] or the "Archaic Words" List[12]). Whether a word was actually *unknown* to the user was later confirmed through translation questionnaires as explained in our procedure. With our setup of easily comprehensible texts we aim to eliminate potential word or grammar difficulties that could interfere with our study's manipulations.

We randomly chose four *Full Texts* from the ESL corpus [270]. Two texts were used for the visual presentation (En 1, En 2) and two for the auditory evaluation (En 3, En 4). The texts had a mean of 44.24 sentences ($SD = 2.05$) and 528.25 words per

[12]Lexico Archaic Word List: `www.lexico.com/en/explore/archaic-words`, last accessed January 3, 2022

text ($SD = 12.81$) (cf. Table 6.2). We manipulated 15 sentences (i.e., one out of three) of each text to contain one potentially *unknown* word. The following sentences are an excerpt of the English text En 1, including two manipulated words:

> "They want to remember their culture and teach their ***progeny*** the old ways."
> "Sometimes the Inuit ***seethed*** their food but often it was not cooked at all."

The percentage of manipulations remained low so that they did not affect overall text comprehension and created a realistic scenario as it could occur during the use of media content. We randomized the presentation of texts within the conditions, to avoid content-related effects.

For the *Native Baselines*, we translated one text carefully into German, the participants' native language. This text was later split into two shorter parts, Ge 1 and Ge 2, with each 15 sentences to serve as baseline for both the reading and listening condition and familiarize the participants with the text and auditory presentation. We decided to split the text to create two short comprehensive texts to not strain the user's attention before starting the presentation of the English content. The native baseline was presented as the first condition in both modalities and did not include any manipulations.

Furthermore, we presented the participants with 60 random *Individual Words* to assess the neural reaction time on a single word basis without influences of overall text comprehension. The individual words can provide insights on potential offsets of response time induced by our setup. The words include verbs, nouns, and adjectives, half of them easily understandable, and half of them supposedly difficult (cf. procedure full-text manipulation).

Comprehension Questionnaires and Translations We assessed the participant's text understanding using 10 pre-validated multiple-choice comprehension questions provided in the corpus by Quinn and Nation [270] (five for the native baselines). For example, for the text "Life in the South Pacific Islands", the following question is given:

> *Thousands of years ago, people came to the Pacific Islands from*
> *a) South America.*
> *b) Asia.*
> *c) Australia.*
> *d) Europe.*

In addition to the comprehension questions, we asked the participants to translate the manipulated and potentially difficult words. After each full text and the individual word presentation, we provided the participants with a translation test, containing

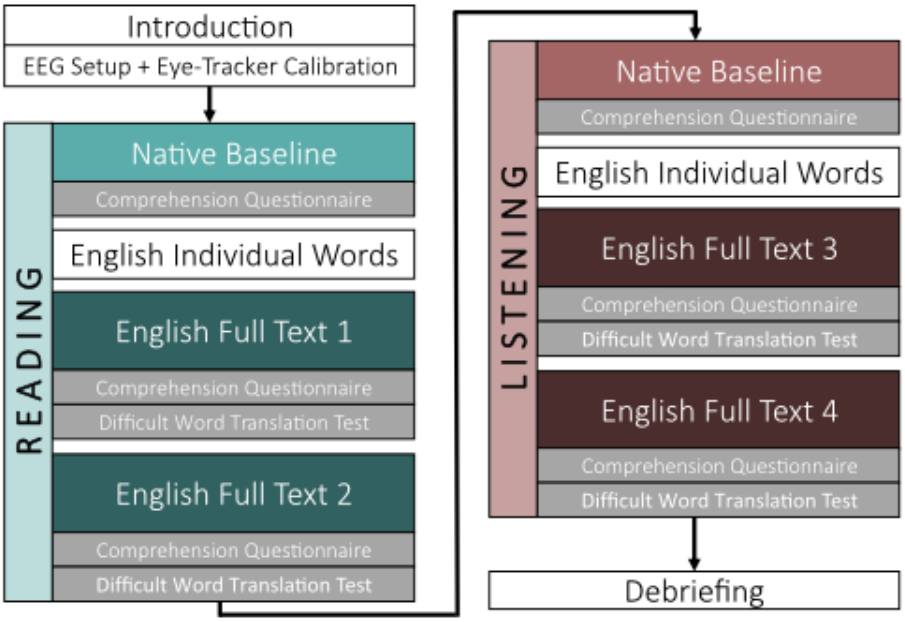


Figure 6.6: After welcoming our participants and preparing our study setup, the participants took part in the evaluation of four reading and four listening trials along the procedure shown in this figure.

a list of the difficult English words with blanks next to them to fill in the correct translation.

Procedure

After welcoming the participants and explaining the process of the user study, they gave informed consent for participation and data handling following the European GDPR. Next, participants chose a random ID from a sheet of prepared user IDs to ensure the anonymization of the data and we introduced them to the EEG and eye-tracking setup. Participants filled in a questionnaire asking about demographic data, including age, highest education level, gender, vision impairment, and history of neurological diseases. We assigned the participants randomly to two conditions, resulting in a changed sequence of text presentation within the two modalities. Afterward, they passed through the reading and listening phase, each including one native baseline text, a condition presenting 60 individual words, and two full texts with additional questionnaires (cf. Figure 6.6). Following the text presentation, the participants were asked to fill in the respective comprehension and vocabulary translation questionnaire. Overall, the participation in our user study took around 110 minutes (including electrode setup, debriefing, and cleaning the electrode caps). As a study compensation, participants could choose between a 20€ voucher for an online shop or study credit points.

EEG Data Processing

To analyze the recorded data, we used the Python MNE library[13] and resampled the raw EEG data to 250 Hz. Afterward, the data was high pass filtered at 1 Hz and low pass filtered at 125 Hz. The data was then re-referenced to the average of all channels which included the original reference electrode FCz. To clean our data, we used a notch filter to remove the 50 Hz powerline noise. We then extracted the epochs and rejected every epoch with an amplitude of higher than 200 μv around the EOG electrodes to remove ocular artifacts from the analysis. Afterward, the data was high pass filtered with 0.2 Hz and low pass filtered with 35 Hz. Finally, we sliced the epochs into blocks of -0.3 ms and 0.7 ms, where 0.0 ms denotes the onset of the stimulus. We automatically extracted the ERP negativity peaks and their latencies. For detection of the N100 during listening, we located the minimum peak in a 50 ms to 150 ms time window after stimulus onset, using a 10 Hz low pass filter. For the N400 ERP detection during reading, we chose a 350 ms to 450 ms time window, respectively.

Sample

We recruited 16 participants (nine identifying as female, seven as male) through our university's internal mailing lists, Facebook page, and Slack channel. The age range of our participants was 20 to 55 ($M = 24.25$, $SD = 8.09$), with 13 participants having a high school degree, two having a master's degree, and one having a secondary school degree. Five participants stated to wear glasses. They were asked to remove the glasses to increase the eye-tracking accuracy for optimal recognition of reading behavior. All participants reported being able to read the text shown on the screen without any problem. Due to an issue in the study setup, the first four participants had to be excluded from the reading trials since the EEG signal was incorrectly mapped to the words the participants were reading. The mapping worked accurately for the subsequent twelve participants after resolving this issue. This error did not have any influence on the listening trials. Thus, the full sample size for the listening trials remained 16.

6.3.2 Results

In the following, we elaborate on the evaluated sample and investigate differences in the ERP amplitudes for *known* and *unknown words*. Furthermore, we assess the feasibility of classifying ERPs to detect vocabulary gaps in real-time.

Text and Vocabulary Comprehension

The evaluation of the *overall comprehension questionnaires* showed a medium to high text comprehension rate across all texts. Figure 6.7 sums up the results for the two conditions. Randomization group A included seven participants from which two had

[13] `www.mne.tools/stable/index.html`, last accessed January 3, 2022

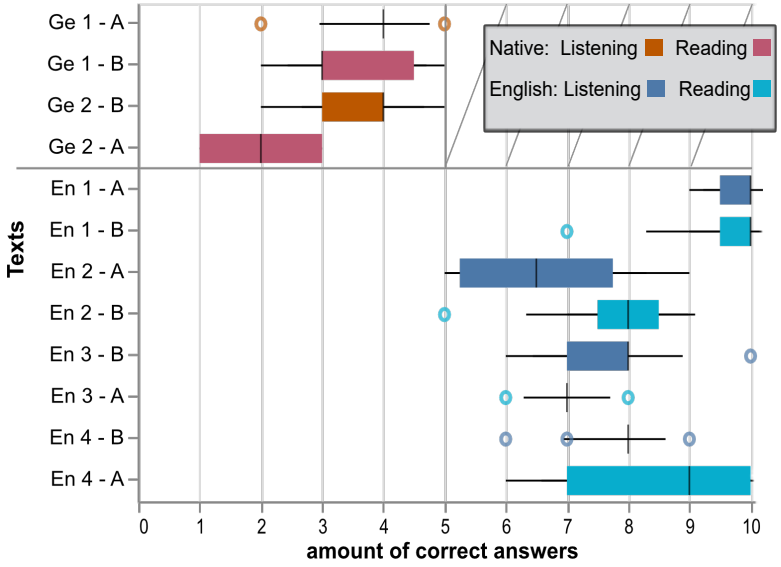


Figure 6.7: Amount of correct answers in the comprehension questionnaires for German (Ge) and English (En) texts across participants of the randomized groups A and B. We compare the results for listening (orange/dark blue) and reading (red/light blue).

to be excluded from the reading trials. Randomization group B contained nine participants, excluding two from reading. For the German baseline, participants achieved a median of four correct answers out of five questions during listening and a slightly lower median of three correct answers during reading for all comprehension scores.

For the English full texts, we notice only minor differences in comprehension scores between the four texts and the modalities. For text En 1, participants achieved a median of ten out of ten correct answers in both modalities; for En 2, they reached eight correct answers for reading and 6.5 for listening. Within texts En 3 and En 4, participants scored a median of eight correct answers for listening, seven and nine for reading respectively. The minimum amount of correct answers across all participants for the English text comprehension questionnaires was five. Thus, we can assume an appropriate level of comprehension for all study participants across the English language texts.

Additional to the overall comprehension, we presented participants with *vocabulary translation tests*, asking to fill in the German word for the potentially *unknown* English words we used as manipulation. The participants were able to translate 7.76% of all difficult words correctly. The majority of questions were left blank, as can be seen in Figure 6.8. Based on the translation tests, we excluded all potentially difficult words, which the participants were able to translate, from our analysis. Thus, we can clearly differentiate between *known* and *unknown* words.

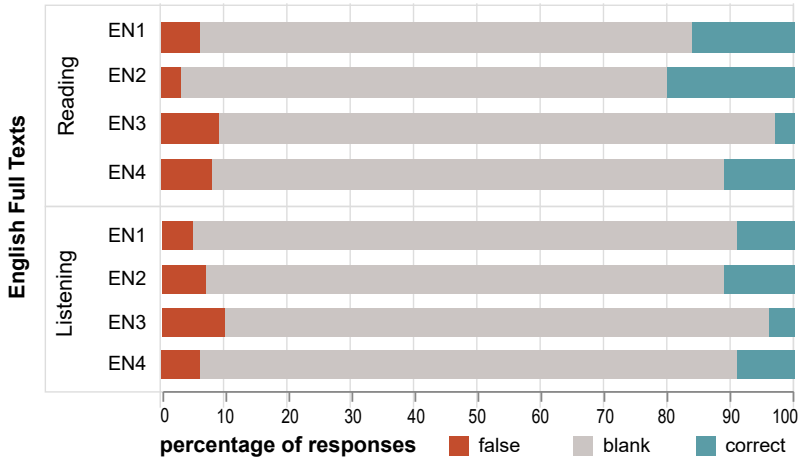


Figure 6.8: Amount of false, blank, and correct answers in the vocabulary translation questionnaires for both modalities.

Evaluating Event-Related Potentials

We statistically analyze the amplitudes of the ERPs generated by *known* and *unknown* English words. Accordingly, we investigate ERPs across both text presentation conditions, individual words, and the full texts for the respective reading and listening trials. We focus our evaluation on the electrode Cz. It is frequently reported in related work that the N400 is larger over the central region of the scalp [196, 197, 198]. For the N100, the Cz electrode also shows higher amplitudes for unexpected stimuli and is also called "vertex potential" [321].

Using the extracted negativity peaks for the N400s and N100s, we average the resulting amplitudes for the full-text conditions of reading and listening. A Shapiro-Wilk test shows a deviation from normality when listening to narrated sentences or individual words ($p < .05$). Thus, we proceed with the non-parametric Wilcoxon-Signed rank test for the analysis of the listening trials. The test reveals a statistically significant difference in the N100 amplitudes between *known* and *unknown* words for individual narrated words ($Z = 78$, $p < .001$; see Figure 6.9a) as well as listening to narrated full text ($Z = 300$, $p < .001$; see Figure 6.10a). The results indicate that there are measurable differences in the auditory processing of *known* and *unknown* words.

A Shapiro-Wilk test does not indicate a deviation from normality when reading sentences ($p > .05$). Therefore, we submit the N400 amplitudes of the reading trials to a t-test for statistical analysis. We find a statistically significant difference in the N400 amplitudes when reading individual words ($t(15) = 14.327$, $p < .001$, $d = 1.531$; see Figure 6.9b) as well as full texts ($t(23) = 23.307$, $p < .001$, $d = 0.758$; see Figure 6.10b) with a large effect size. These results suggest that there are measurable differences in N400 amplitudes, indicating differences when processing *known* and *unknown* words during reading.

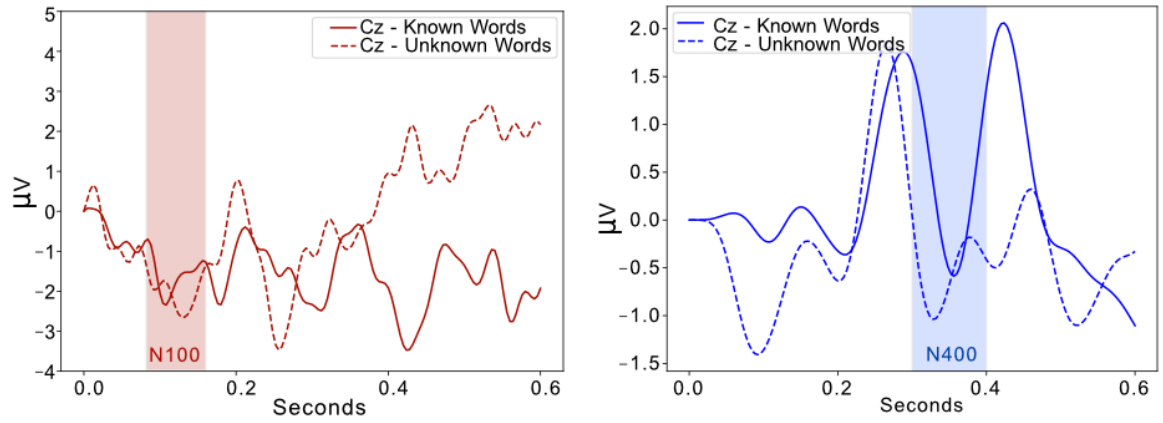


Figure 6.9: ERP responses measured at the Cz electrode for hearing individual narrated words and reading written words. **(a):** *Unknown* words generate greater N100 amplitudes during auditory presentation as compared to *known* words. **(b):** For visual presentation, *unknown* words elicit greater N400 amplitudes as *known* words.

Predicting Vocabulary Gaps

The results from our study show significant differences between the N400 and N100 ERPs when reading and listening to *known* or *unknown* words. We investigate the performance of person-dependent classification based on the extracted ERP amplitudes.

Features, Instances, and Classifier Performance

We apply the same data processing as in the analysis mentioned before to extract the ERP amplitudes. Thereby, we focus on the N400 amplitude to detect word comprehensions during the reading of sentences and on the N100 amplitudes when detecting vocabulary gaps in auditory narrated text. Separately for reading sentences and audio listening, the N400 amplitudes were used for the reading trials and N100 amplitudes for auditory trials. We labeled words afterward as a *known* word or *unknown* word. If *unknown* words were translated correctly after each trial, they were labeled *known* words. The number of epochs was different for each participant since a different number of epochs was rejected. Therefore, per participant between 450 and 510 epochs ($M = 482.75$, $SD = 18.9$, $N = 11586$) with *known* words and between 12 to 15 epochs ($M = 13$, $SD = 1.3$, $N = 312$) with *unknown* words were taken into account during the reading trial. Between 487 and 509 ($M = 498.5$, $SD = 16.3$, $N = 11964$) epochs for *known* words and 13 to 15 ($M = 14.2$, $SD = 1.2$, $N = 341$) epochs for *unknown* words per participant were considered for further analysis for the auditory trials. According to the remaining epochs, the N400 and N100 amplitudes were labeled for classification and were the only features used.

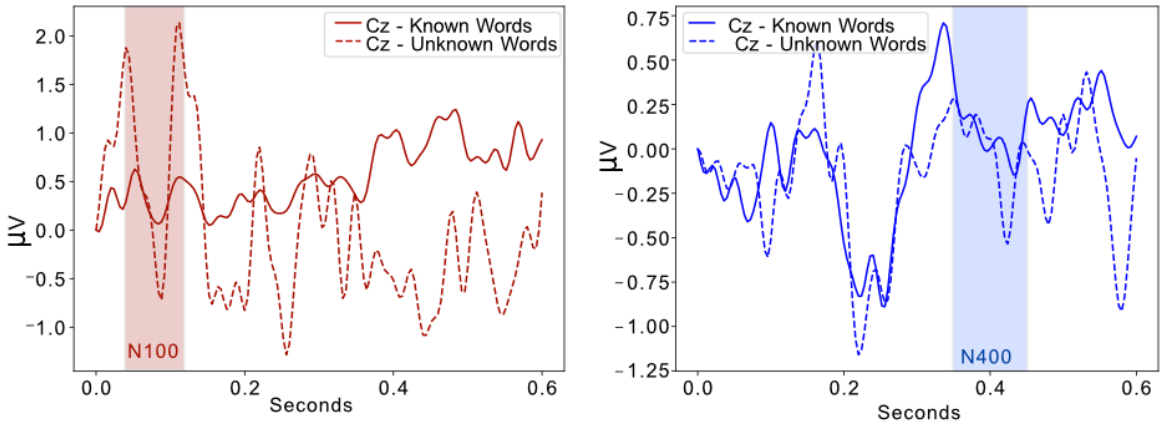


Figure 6.10: ERP responses measured at the Cz electrode for *known* and *unknown* words perceived during narrated full texts and during reading full texts. **(a):** *Unknown* words generate greater N100 amplitudes during auditory presentation as compared to *known* words. **(b):** For visual presentation, *unknown* words elicit greater N400 amplitudes as *known* words.

To sustain a natural scenario, the number of the attributes with the label *known* word is much higher than the number of attributes with the label *unknown* word. Therefore, we slice the number of attributes with the label *known* word to the same number as *unknown* words for each participant. Using scikit-learn[14], we train a Support Vector Machine (SVM) with a radial kernel. We perform a cross-validation on the trained instances with $k = 5$, where $k - 1$ folds were iteratively used for training while the remaining fold was used for evaluation. This process was repeated until *known* words were evaluated with an equal set of *unknown* words. Finally, we calculate the mean from the resulting classification scores to aggregate an overall accuracy. The average overall accuracy in discriminating *known* words and *unknown* words during reading was 87.13% ($SD = 4.5\%$). Auditory trials resulted in an overall accuracy of 82.64% ($SD = 9.6\%$). In both modalities, the majority of wrongly classified words were false positive compared to false negative (listening false positive=10.96%, false negative=6.4%; reading false positive=9.35%, false positive=3.52%; see Figure 6.11). Hence, in case of wrong classification, words are more likely to be classified as *unknown* words, although the word is already *known* during both listening and reading.

6.3.3 Discussion

In this work, we investigated the neural responses during vocabulary comprehension. We found significant differences in the size of amplitudes in N400 ERPs during reading

[14]scikit-learn: `www.scikit-learn.org`, last accessed January 3, 2022

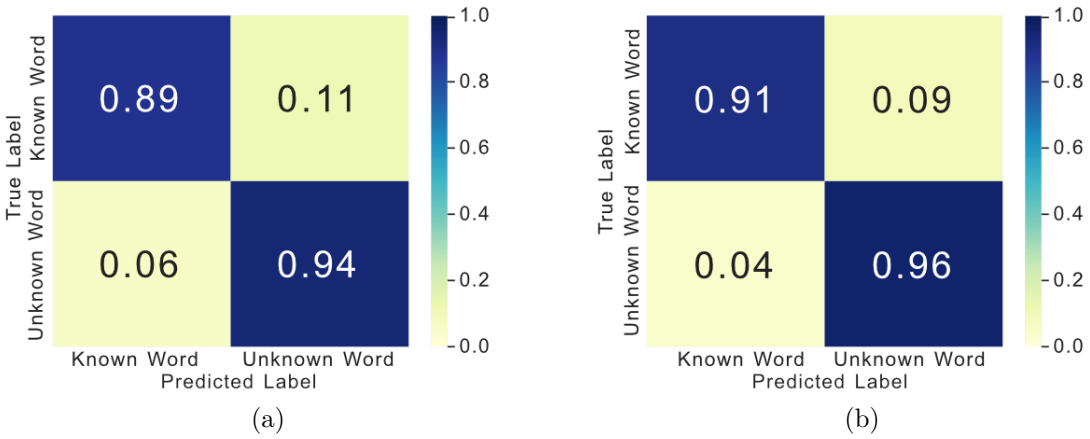


Figure 6.11: Normalized confusion matrix of classifying *known* words and *unknown* words between the **(a):** auditory trials and **(b):** reading trials.

and in N100 ERPs during listening as neural responses to *unknown* words. The classification of the responses performed with an accuracy of above 80%. The *BrainCoDe* concept worked effectively for the detection of vocabulary comprehension problems in second-language reading and listening. We discuss the opportunities and limitations of our approach to be utilized in real-world scenarios.

Limitations Like many EEG-based studies, our evaluation was done in a quiet environment with no distractions. This was suitable to obtain data with limited noise from outside sources to be able to explore the EEG data. However, the discrepancies between the experimental conditions applied in our study and real-life situations can not be neglected. We need further evaluation to explore the feasibility of *BrainCoDe* to detect comprehension problems in everyday scenarios with potentially influencing environmental factors such as noise or complex media such as movies.

When comparing the results of the raw NASA-TLX questionnaires we recognize a significant difference in perceived workload when reading native as opposed to English words. This shows that subjectively perceived workload was manipulated in conjunction with our finding in N400 amplitude. The comprehension tests show that there is a difference of complexity when comparing the two English texts. However, when marking the unknown words in the print-outs, the participants reached a 70% consensus. Therefore, we assume that the participants perceived the unknown words as equally difficult to translate in both English texts. We infer that the text comprehension rate is affected by the text difficulty, however, this does not have an impact on the individual N400 measures.

Participants achieved lower text comprehension scores for the text E1 compared to E2 although E2 is more difficult according to the Lexile score. Since the Lexile scores takes

several syntactic and semantic factors into account to calculate the text difficulty, a potential threat to validity is posed by participants that were familiar with the text. However, we have observed the effect of N400 for most of the words that were unknown for participants. Thus, we believe that the overall text difficulty does not represent a higher occurrence of N400s.

Exploring Language Learning Modalities In our study, we focused on reading and listening as modalities important for language learning. We achieved a classification accuracy of 87.13% when differentiating between reading *known* and *unknown* words on screen. This can be used for situations such as reading an e-book in a new language on a tablet. The analysis of EEG responses in such a situation would still require eye-tracking to be able to pinpoint the exact word the user is looking at at the moment. However, this technology is currently finding its way to commodity tablets, PCs, and smartphones [177, 350].

The *BrainCoDe* approach achieved a classification accuracy of 82.64% for the listening modality. This approach can be used in a real-world learning scenario, such as listening to an audiobook in a second-language with slow narration speed or listening to recorded conversations in an online language class. However, finding comprehension problems using our approach in more complex media such as movies that contain multiple stimuli and modalities would require further investigation. Fortunately for these scenarios, a perfect classification accuracy is not necessary to build successful learning applications, since including false positives, thus, repeating already *known* content, does not hinder learning.

Person-dependent Learning Currently, the evaluation of ERP data only creates expressive results through averaging over many trials and is highly user-specific [218]. In order to use EEG responses as input for applications, a training phase is required [217]. Similar to most state-of-the-art systems utilizing physiological sensors, *BrainCoDe* requires a user-dependent training phase to be able to detect vocabulary incomprehension due to unique manifestations of ERPs for every individual user. Single-trial ERP classification is currently advancing and shows promising results [33] that would make our approach feasible for real-time classification of *unknown* vocabulary in the future.

In addition to inter-person variations of ERPs, the momentary cognitive and physiological state of each user may have an effect on their neural responses per session. The response to repeated *known* and *unknown* words over time may also be different as the user starts to learn the language. Related work in the use of ERPs for language learning also suggests that the neural responses (e.g., N400) of listening stimuli are susceptible to habituation effects, for example, when the user listens to an unexpected stimulus frequently [47]. These factors must be taken into consideration when designing a language learning application utilizing neural responses.

In this work, we focused our analysis on the N400 and N100 ERP components based on prior work. Nonetheless, the closer investigation of other ERP features or potential

feature interactions can yield further insights into comprehension problems during listening and reading and improve the classification accuracy.

BrainCoDe Application Scenarios and Generalization We envision a personal learning application that provides both real-time and post-hoc personalized feedback. The users would start by wearing a Brain-computer Interface (BCI) headset, such as the Emotiv EPOC[15] or the OpenBCI[3], both of which have support for the used electrode setups verified by BrainCoDe. The users would go through a training session by reading and listening to phrases with *known* and *unknown* words to assess their current language level and collect training data, then the media content (e.g., audiobook, movie, e-book, or language learning chat session) would be started. The application can provide real-time or post-hoc feedback. Real-time feedback could recommend the user to pause, rewind a scene of the media content, or repeat a section of the audiobook to increase exposition to the new vocabulary and enhance learning efficiency. As a post-hoc analysis tool, the *BrainCoDe* approach easily facilitates the extraction of *unknown* vocabulary to create a personalized list of contents for the user to learn. Based on a continuous monitoring of the users' language comprehension, *BrainCoDe* could be applied to provide recommendations on media contents adapted to users' proficiency as explored by Yuksel et al. [359] for learning the Piano.

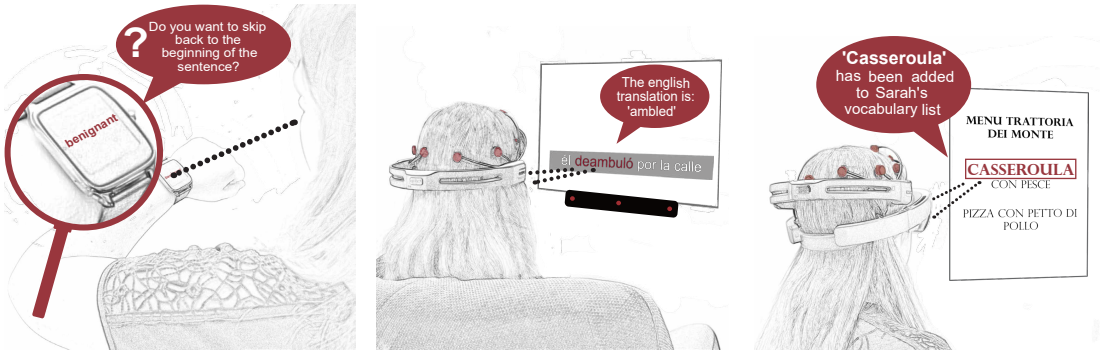
Integrating our approach into applications such as Duolingo[16], we could implement a tool to adapt the content according to the user's knowledge. Thus, the app could present content, which is currently *unknown* to the user, with higher frequency to support learning.

6.3.4 Summary

The detection of gaps in a learner's vocabulary knowledge is a critical step to facilitate effective second-language learning. Since media consumption in different languages is nowadays a common tool to learn and improve on a language, such as in the form of movies, e-books, or audiobooks, the implicit detection of comprehension problems becomes increasingly important. Avoiding interruptions and distractions while consuming media helps to improve the user experience. Understanding what words are unknown to the user allows proactively presenting translations or explanations without interrupting the media consumption. In this section, we showed that by evaluating N100 and N400 ERPs using the single Cz electrode, we can implicitly detect unknown words during foreign language listening just as well as during reading. By building a classifier trained to identify those gaps, we can successfully recognize unknown vocabulary in eight out of ten situations. Thereby, the accuracy of reading (87.13%) exceeds the

[15]Emotiv EPOC: `www.emotiv.com/epoc`, last accessed January 3, 2022

[16]Duolingo: `www.duolingo.com`, last accessed January 3, 2022



(a) Use Case 1 (b) Use Case 2 (c) Use Case 3

Figure 6.12: Depiction of the three use cases for everyday implicit comprehension assessment.

accuracy of detecting known and unknown words during listening to narrated content (82.64%).

6.4 Use Cases for Ubiquitous Language Learning Support

In the following, we present three potential use case perspectives to depict how technology can help embed second-language acquisition into everyday environments. We showcase how our BrainCoDe approach could be embedded into the daily lives of users in an implicit and supportive manner.

Use Case 1 - RSVP Reading on Small Screen Devices Designing efficient reading UIs on devices with limited screen space, such as a smartwatch, is challenging within mobile contexts [90]. The usage of RSVP, where the whole screen presents a single word, is one UI design option that has been evaluated [118]. It has the advantage of presenting text in a reasonably large font and allows reading with little eye movement. However, this UI has one inherent limitation: Being presented with just one word at the time, a user cannot take a step back when encountering problems of understanding due to unknown words. With the insights from analyzing the EEG data, we would be able to detect in real-time when the user encounters unknown words or potentially troublesome concepts during reading as illustrated in Figure 6.12a. By combining a smartwatch and a mobile EEG device we could use an algorithm to dynamically adapt the content and UI. For example, offer real-time support by showing words for a longer time, provide translations, or by show unknown words more frequently in other media contents. The realization of this scenario is, however, depending on the development of portable and pervasive EEG devices, which have the potential to become afford-

able [83], unobtrusive [82], nearly invisible [34], and feasible for applications including natural actions and cognition [79].

Use Case 2 - Media Consumption on Screen Using media content in foreign languages has been often reported to be a useful tool to improve language skills [137]. One area, where our approach can be beneficial, is the presentation of subtitles in videos. Including subtitles in audio-visual content can support the acquisition and the improvement of language skills [129, 247, 339]. There are already tools exploring the potential of subtitle translations (GliFlix [291]), or second screen application to present important concepts of TV shows (*Flickstuff* [178]). By using our EEG-based approach in monitoring users' comprehension during media usage as shown in Figure 6.12b, we can provide an effective tool for real-time and post-hoc vocabulary learning support, such as personalized vocabulary lists. To implement this, one would couple the EEG monitoring and analysis with gaze tracking to detect the current focus of the users and thus, to identify the word in question. The application would then be able to either show instant translation support on the screen or add the unknown word to a list of words to be repeated at a later point in time.

Use Case 3 - Media Content in Ubiquitous Environments with Smart Glasses What we envision for subtitles on a screen could also be transferred into real-world environments. In our everyday life, we encounter signs and texts as they are ubiquitous in our surroundings. With a setup consisting of smart glasses that include a camera, we can link gaze tracking to a mobile EEG. Thus, it will become feasible to also detect signs and texts not understood by the user on many digital and analogue devices such as advertisements or public screens. The general approach to apply comprehension analysis in the physical surrounding is to detect what the user is perceiving and assess their brain's reaction. Figure 6.12c illustrates our vision that this can be realized by smart glasses and a front-facing camera. With the help of optical character recognition [240], it is feasible to monitor text in the users' surroundings [114] and provide individual support.

6.5 Chapter Summary

The findings reported in this chapter provide evidence for the potential of EEG data to support language learning. In regards to our research question **RQ3**, we can conclude that it is indeed possible to utilize everyday actions such as reading and listening for the generation of personalized learning content. In particular, ERPs turn out to be a feasible measurement for the detection of incomprehension on a one-word basis. Having an approach to support the continuous assessment of second-language text comprehension brings us one step closer towards the design of ubiquitous learning support. We tested the approach in the context of text reading and aimed to recognize vocabulary incomprehension to support language learning. Still, this method is not limited

to this application scenario and is of particular interest to many areas of ubiquitous technology. Three possible use case scenarios were outlined in this work, highlighting the additional value an EEG based word comprehension system would offer.

Additionally to the use case of foreign language reading, future work should investigate the transfer of this approach to evaluate spoken language comprehension and clarify the feasibility for real-time support. Hereby, this technique could support real life communication, which is, in particular, important in conversations where the two involved parties show different levels of language proficiency. With the steady improvement of EEG technology, we are confident that physiological signal analysis can contribute significantly to our overall goal of creating a learning experiences that is seamlessly embedded into users' every life.

IV

Connecting Micro-Learning Sessions

7

Design Space for Task Resumption Cues

In Chapters 3 and 4, we investigated everyday usage situations of mobile learning and found a high prevalence of interruptions from the users' environment, mobile device, and the users themselves. Such interruptions direct users' attention away from the learning task, thus, impeding learning.

Researchers have extensively evaluated the possibility of postponing or managing interruptions for mobile devices [60, 156]. One example is to delay notification delivery, such as in the bounded deferral strategy [153]. Incoming notifications are not shown immediately but postponed until a more opportune moment occurs. Using these opportune moments to present potentially interrupting content, for example, delivering notifications in between two tasks, can have a less disruptive effect [22]. Nonetheless, it is not easy to find moments of inattentiveness during a day. A study by Dingler and Pielot [89] evaluated users' attention on mobile devices during the day and concluded that intelligent notification delivery services would have to carefully make use of rare and quickly subsiding moments of inattentiveness.

Since Chapters 3 and 4 emphasized that while many of these interruptions can not be anticipated or avoided, there is the opportunity to support users in resuming the learning task after the interruption. In this chapter, we take an in-depth look into task resumption support and will specifically focus on Task Resumption Cue (TRC) (TRCs) - memory cues that support the recall of priorly acquired knowledge with the goal to support the resumption of the task.

In this chapter, we address the following research question:

- **RQ4:** How can we use memory cues to mitigate the negative effects of interruptions in everyday mobile language learning?

We consider TRCs memory cues with a specific purpose. By reviewing similar concepts from various domains, we show that while prior work has extensively explored memory cues for task resumption in desktop-based and controlled lab scenarios. However, research on the application of such cues in mobile in-the-wild settings is still sparse, particularly when it comes to mobile learning applications. Analyzing the state-of-the-art literature, we derive a set of concepts that are promising for the application in mobile learning scenarios, identify current research gaps, and formulate design requirements for the use of memory cues in mobile learning applications.

This chapter is based on the following publications:

- Schneegass, C. and Draxler, F. (2021). Designing Task Resumption Cues for Interruptions in Mobile Learning Scenarios. In *Technology-Augmented Perception and Cognition*, Dingler, T. & Niforatos, E. (Eds.) (pp. 125-181). Springer, Cham. DOI: 10.1007/978-3-030-30457-7_5 [300]
- Draxler, F., Schneegass, C., and Niforatos, E. (2019). Designing for Task Resumption Support in Mobile Learning. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (MobileHCI'19). Association for Computing Machinery, New York, NY, USA, Article 47, 1–6. DOI: 10.1145/3338286.3344394 [95]

Some text passages were taken verbatim from these publications.

7.1 Related Work

7.1.1 Task Resumption in Existing Applications

Developers and interaction designers started to integrate task resumption strategies into apps long ago. The design may often have been the result of intuition rather than a decision grounded on scientific reasons. Nevertheless, the strategies we identified in common apps do tend to incorporate the principles we mentioned above. They are frequently cue-based, use visual support, and aim to help users get right back to where they left off. Some of them have already been integrated deeply into our daily routines so that by now we are unlikely to even explicitly notice them as resumption strategies.

Reminders

Some learning apps such as Duolingo use a general reminder strategy: they issue notifications to remind users that they should continue with their exercises (cf. Figure 7.1). These reminders do not cue specific memories of the user but function as a mere reminder to engage in learning activities.Similarly, the Facebook messenger Android app[1] uses so-called "chat heads": small bubbles that show conversation partners' profile pictures on top of any screen content and thus remind, if not of conversation content, then at least of the people with whom the users had ongoing conversations (cf. Figure 7.2 (a)).

[1] Facebook Messenger App: https://play.google.com/store/apps/details?id=com.facebook.orca&hl=en, last accessed January 3, 2022

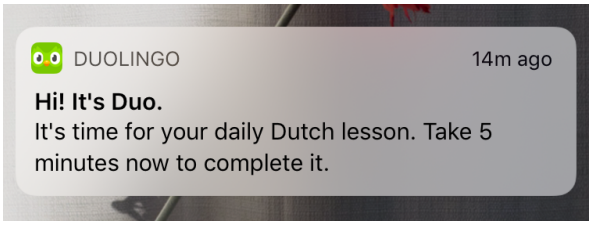


Figure 7.1: Duolingo uses daily notifications to remind users to continue with their exercises.

Regaining Context

Furthermore, we can find methods to regain context. Recent app views on Android and iOS, but also Windows desktop systems show the screen content as recorded just before exiting an app (cf. Figure 7.2 (b)). Depending on the implementation of the app (and also depending on whether system memory or caches have been cleared in the meantime), apps also return to this state when a user resumes it. For instance, browsers typically return to or reload a previously viewed tab. Duolingo returns the position in the skill list where you were before quitting. Netflix includes a "continue watching" option with a progress bar underneath previously viewed programs to indicate where a user stopped watching. Further apps provide a list of recent documents (e.g., Dropbox[2] and Acrobat Reader[3]).

Repetition

In addition to presenting the previous state, some programs also include a recapitulation of past activities. Specifically, audio and video players such as the iOS app "Podcasts" provide the option of rewinding by a couple of seconds so that previous content can be repeated before playback continues.

However, there are also some apps where task resumption, especially after a long interruption, is explicitly not intended. News and map applications, for example, tend to update their interfaces to present new stories or to adjust to the user's current location. Another example is the Facebook app, which currently returns to the news feed upon opening. In these cases, the user experience design was not intended for task continuation, and showing up-to-date information is deemed more important than understanding what was done before.

[2] Dropbox: https://www.dropbox.com/, last access 02/16/2020

[3] Acrobat Reader: https://get.adobe.com/de/reader/, last access 02/16/2020

(a) Recent app views give an overview of previous app states on an iOS device.

(b) Chat Heads in the Facebook Messenger app on an Android device.

Figure 7.2: Examples of existing task resumption cues on mobile devices.

7.1.2 Pedagogical Memory Re-Activation

In school teaching, a common approach to start a lesson is to re-activate the knowledge gathered in the previous learning session. Psychological and pedagogical research emphasizes the central role of including prior knowledge into the teaching process [185]. Re-activating prior knowledge in terms of human cognition and memory can be described as recalling information from the long-term memory (LTM) and holding it in the short-term memory (STM) or working memory for the integration of new information (cf. [21]). In pedagogy, the terms *open* and *specific* re-activation are used to describe the integration of prior knowledge. Open re-activation describes the general activation of a broad topic, e.g., by using visualizations like mind-maps or brainstorming [324]. In contrast, the specific re-activation of knowledge targets a unique information chunk. Therefore, a common technique used for specific re-activation is asking questions [185]. Both strategies can be useful to re-activate a certain memory and can be more appropriate depending on the use case [324].

7.1.3 Interruption Lag

The reaction to the interruption trigger can be immediate or delayed. Immediate interruptions occur, in particular, when a learner is not aware of the interruptions and cannot control them (external and internal, but unplanned). However, engaging with external and internal interruptions can sometimes be scheduled flexibly. For example, users can postpone notifications by setting their phone to flight mode, or by finding a quiet place for studying without distractions.

Being aware of an upcoming interruption furthermore provides the learner with the opportunity to prepare for it. In their study, Trafton et al. [329] left it up to their participants how to make use of the time between the primary and secondary task. When given this opportunity, the majority of study participants used the time to prepare for the upcoming interruption. The effect of the preparation was an overall lower resumption time when proceeding with the primary task [329]. In a study by Brumby et al. [45], a ten-second interruption lag already reduced the number of mistakes made after resuming the primary task. In conclusion, if an interruption can be anticipated, delaying or managing an interruption can be a solution to mitigate the negative effects.

7.1.4 Task Resumption Strategies

When users start to shift their attention back from an interruption task to the primary task, they need time to reorient themselves. Task resumption strategies aim to keep this time as short as possible and to help users regain full awareness of the task context, to keep task completion time and error rate at a minimum. Oulasvirta and Saariluoma [255] stress the importance of interface organization for effortless task resumption. For instance, they suggest to group items based on higher-level concepts that users can interpret. Thus, it is easier for users to encode the workspace in memory, but most importantly also to recover this knowledge when returning to the task: necessary controls are located more easily and faster. A user interface (UI) which the user does not need to remember in detail also decreases the amount of information to be stored in the problem state as defined by Borst et al. [38]. Cades et al. [52] showed that dealing with interruptions can be trained: participants' resumption lag decreased when they were interrupted several times, even when they had had time to get used to the primary task. Woelki et al. [349] looked more closely at the effect of practice time. They found that the disruptive effect of interruptions can be substantially decreased when participants become used to executing a primary task. In Section 4.1.2, it was argued that short interruptions are less disruptive. Therefore, reminding users to switch back to their primary task is a valid strategy for minimizing the effort and cost of task resumption: the decay of goals in memory will not have advanced far, and the problem state will be easier to reconstruct.

Memory for Goals Theory

In the Memory-for-Goals theory [5], interruptions are viewed as a suspension of the primary task's goal. It states that the user can only retrieve a suspended goal with the help of priming through a memory cue. This cue needs to be presented a first time when the current goal is suspended, for example, because of an interruption by a secondary task. The cues are then presented a second time when the goal needs to be resumed to prime the target [5]. Moreover, Trafton et al. [329] differentiate between two ways to encode a goal, namely: (1) retrospective ("What was I doing before?") and (2) prospective ("What was I about to do?"). Prospective memory cues can be primed in the interruption lag, for example, by taking notes of what you were about to do if the interruption would not have happened and are recalled in the resumption lag. Retrospective cues can still be applied without the need of former priming.

Task Resumption Cues

Finally, a major part of resumption strategies relies on cues that aid the user's memory or direct the focus of attention: they range from implicit stimuli such as barely noticeable highlights (e.g., [232]) to complex content cues that restore full task context (e.g., [293]), and can be presented in very different ways and on various channels. For mobile scenarios, where interruptions are frequent, cannot necessarily be predicted, and vary significantly from one occurrence to the next, cues are particularly important. Well-designed cues provide a way of consistently dealing with the challenges that interruptions impose upon users. However, so far, there has been almost no research on mobile TRCs, in particular, for mobile learning, a process that demands consistent attention and where disruptions greatly decrease performance. This lack of available solutions was the motivation for our systematic literature research and exploration of opportunities through a design space which we describe below.

7.2 Literature Review

In this section, we systematically analyze literature from various domains on TRCs. We extract and compare the characteristics of these cues based on five fundamental dimensions. We generate a design space, discussing all possible attributes in each dimension. We will point out well-evaluated cue designs, which have shown their potential for effective task resumption support. Additionally, we will highlight gaps in the design space that leave room for further evaluation and discuss our findings. We will begin this section by presenting the methodology of our literature review approach.

7.2.1 Methodology

We conducted a structured analysis of literature according to the methods used in [325], which we will outline in greater detail below. We started with a great number of publications representing a broad overview of the research field and narrowed down our selection in several steps. Once we had obtained a final set of papers, we evaluated the task settings and interruptions, extracted characteristics of TRCs, and analyzed findings of cue evaluations. This section outlines the procedure we applied for collecting the data points.

Selection of Literature Since HCI is an interdisciplinary domain and includes research from several domains (e.g., computer science, design, or psychology), the publications of HCI are distributed across several online publication libraries. We reviewed full, short, and work-in-progress papers from the ACM Digital Library[4] (ACM DL) as well as IEEE Explore[5], SpringerLink[6], and ScienceDirect[7]. These libraries provide an extensive collection of documents from computer science, but also other disciplines such as medicine, where cues for task resumption are likely to be a relevant research topic. We searched these libraries for groups of terms related to (1) *interruption* and *resumption* and (2) *attention*, *task workflows*, and our focus areas *mobile* and *learning*. From the results of a first round of queries using those keywords, we extracted a third set of terms describing various types of *cues*. Our queries yielded 1207 (query 1) and 1936 (query 2) *unique* results in the ACM DL, 1187 and 548 in IEEE Explore, a total of 9269 in SpringerLink, and 28052 in ScienceDirect across both queries.

The result items were pre-processed by removing duplicates and assigning ratings based on the occurrence of the search terms in title, abstract, or keywords. We weighted the query terms with scores in the range of 0 to 15 according to their significance. The total score of an article was defined as the sum of term scores. We added and reviewed abstracts to all articles with scores above 9 (ACM DL and IEEE Explore) and 30 (SpringerLink and ScienceDirect), respectively[8], and manually classified the best 217/380 results as *not relevant*, *marginally relevant*, and *relevant*. Since two raters performed the rating, we assessed our inter-rater reliability as proposed by Campbell et al. [54] on approximately 10% of those 597 items and achieved a score of $91.\overline{6}\%$. Disagreements on the rating were mostly due to uncertainties regarding the value of (sometimes vague) qualitative results and if to include those for further processing. We discussed these disagreements accordingly to reach a consensus. Thus, three results had to be eliminated upon closer inspection because they either did not include any

[4] ACM DL: https://dl.acm.org, last accessed January 3, 2022

[5] IEEE Explore: https://ieeexplore.ieee.org/Xplore/home.jsp, last accessed January 3, 2022

[6] SpringerLink: https://link.springer.com, last accessed January 3, 2022

[7] ScienceDirect: https://www.sciencedirect.com, last accessed January 3, 2022

[8] Articles obtained from SpringerLink and ScienceDirect generally had higher scores because the result lists contained different types and amounts of meta-information

evaluation of the strategies they proposed, because they only referred to cues defined elsewhere, or because the application area was too far from our context of use.

Finally, we performed additional forward and backward searches using Google Scholar and added a small number of publications referenced in our original selection. Thus, we obtained a final set of thirty articles describing 35 unique cues (two cues each are presented in [154, 212, 258, 287, 312]).

Creating the Design Space Once the selection was complete, we categorized the articles based on the task setting including the device setup, the interruption scenario, and the characteristics of the TRCs they present. For the classification of the cues, we used the dimensions modality, timing, and interactivity, and further attributes derived from the articles themselves. We defined these characteristics in an iterative process to generate a set of descriptive features that enabled us to clearly distinguish the unique cues from one another. Finally, we examined the evaluation state of the TRCs. We specifically noted reported findings regarding the effects of using cues on task performance and error rate. In complement to the categorization, we analysed the articles' metadata, including the primary research area, year of publication, and journal or publication venue.

7.2.2 Results

In this section, we describe the literature sample by summarizing meta information of the 30 publications and 35 cues in our literature set. We outline the setting (such as device and task) in which a TRC is applied and the types of interruption. Moreover, we present our design space derived from the categorization of TRCs as well as details on the included resumption cues. Finally, we complement this with an overview of evaluation states and findings across all cues.

Meta-Analysis of Publications The year of publication ranges from 2002 until 2018 and comprises a set of eighteen full and short conference paper publications, nine journal articles, and four extended abstract submissions (cf. Figure 7.3). The four articles that used mobile devices date from the years 2010, 2015, 2016, and 2018. These recent publications show that the state of technology is comparable to what is available nowadays. The leading publication venue is the *ACM Conference on Human Factors in Computing Systems* (CHI) with nine publications, followed by the *Annual Meeting of the Cognitive Science Society* and the *ACM Conference on Computer-Supported Cooperative Work* (CSCW) with three publications each (see Figure 7.4).

Task Setting and Interruptions Among the thirty different interruption situations described, 21 were using a desktop computer system, two scenarios used mobile devices [227, 358], and two multi-device settings including both mobile and desktop

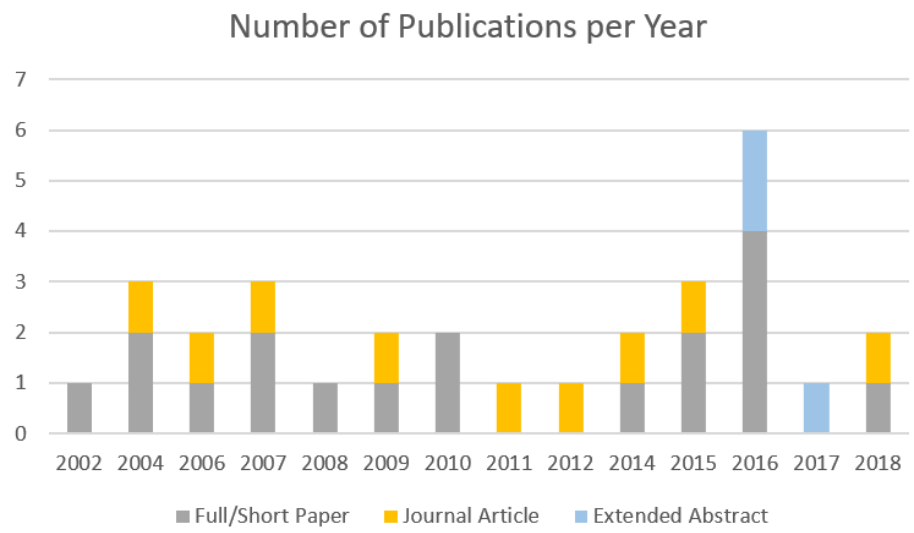


Figure 7.3: The publications used to create the design space sorted by number per year and form of publication (i.e., full and short conference paper, journal article, or extended abstract submission).

devices [62, 175]. Other studies evaluated contexts such as in-car interaction [36] or a smartboard scenario [293].

Furthermore, the primary task setting varied strongly, including programming tasks [258], reading tasks [56, 163], resource allocation and military tasks [67, 148, 308], and various others. Secondary tasks used for the interruptions included phone calls [36, 358], noises [227], video clips [163], search tasks [308], or similar. They varied in their degree of demand and urgency.

In fifteen cases, the interrupting event originated from the system or device itself and was hence classified as device-internal. In contrast, fourteen interruptions were caused by external triggers, and in only one case, the interruption was self-induced: when the participants did not understand a word in a reading task and had to look it up [62]. Across all scenarios, the majority of interruptions were unplanned, meaning that the participants did not expect the interruption to happen at a particular moment. In eight cases, however, the interruption was announced, leaving time for the participants to use the interruption lag for preparation and goal encoding. Table 7.1 gives an overview of task setting and interruptions.

Design Space: Facets & Dimensions

In the following section, we will present the results of our literature review in the form of a design space shown in Table 7.2. This table includes all 30 publications, 35 cues respectively, and classifies them along a set of dimensions as explained below. On the horizontal axis, we aligned the different cue modalities presented in the literature along

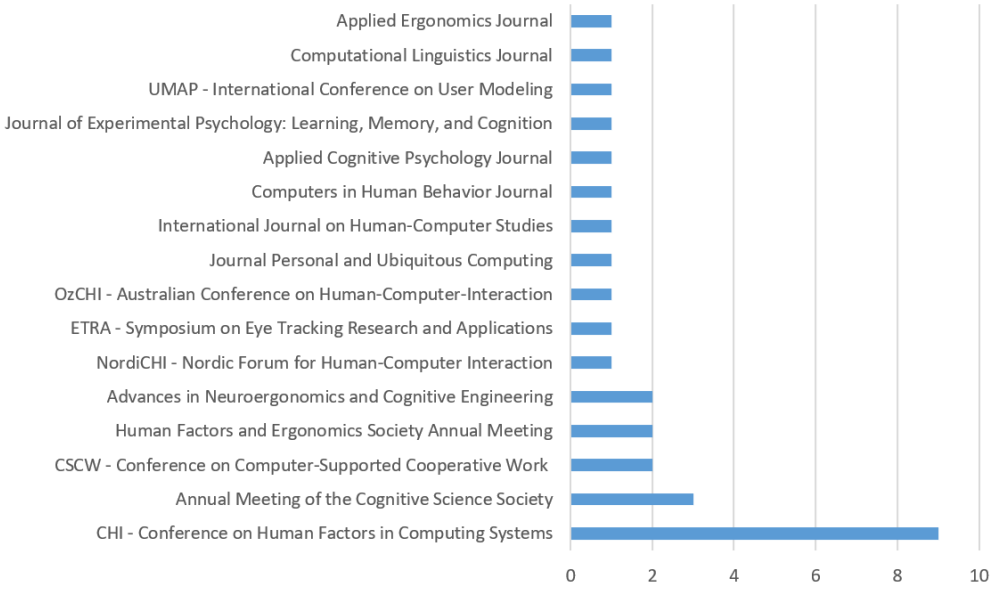


Figure 7.4: Publication venues of all 30 articles presented in the design space and their occurrence frequency in our design space.

with their expressiveness in the case of visual and auditory cues. On the vertical axis, we ordered the publications along four main categories: (1) the purpose of the cue, (2) the level of attention required for perceiving the cue, (3) the timing of cue presentation, and (4) the interactivity of the cue. The differentiation and all descriptions are based on the literature we derived within our analysis and are, therefore, not necessarily exhaustive in terms of what is imaginable with today's technology. In the following, we will describe each of these dimensions in greater detail and state examples for each characteristic included in the design space.

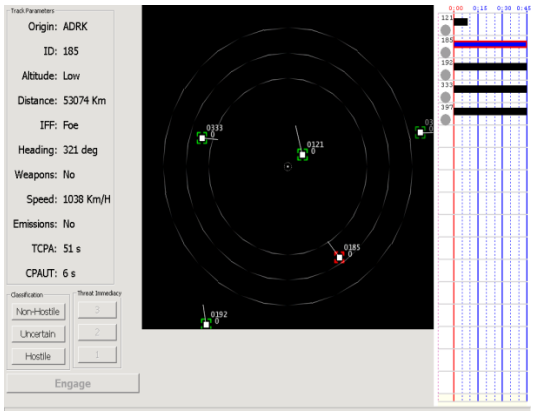
Modality and Expressiveness The modality refers to the signal through which a cue is represented and how it is perceived by the user. Modalities that we encountered in our design space include visual (graphical), auditory, haptic or tactile, or tangible signals. In some scenarios, several modalities are combined as multi-modal cues. It is imaginable to design cues to be perceived by every human sense, including olfactory or thermal cues (i.e., creating a warm or cold sensation and associate a memory with it).

Expressiveness, on the other hand, depends on the amount of information a cue can convey. For example, textual cues can transfer a large amount of information on a small screen, whereas tactile cues (such as a vibration) are limited in their expressiveness. Keeping the expressiveness in mind is particularly important when generating cues that match the capabilities of mobile devices. In this work, we differentiate between *explicit* and *implicit* cue designs. *Explicit* cues present task- or content-related information,

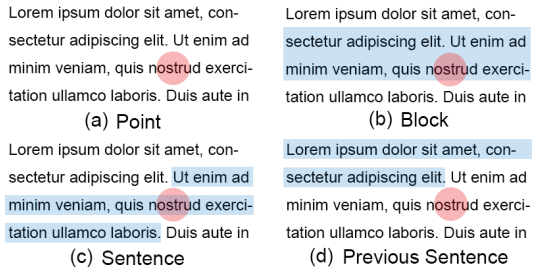
Batch A		Batch B		
Task A1	Task A2	Task B1	Task B2	Task B3
☎: Greeting 1 † ☺:"*Yes*." ☎: "*What is it*?" ☺:"*I am (task label).* ☎: "*Okay, here's your call.*" 🕴:(request) ☺:(response) 🕴:"*Thanks. Bye.*"	☎: Greeting 2 † † ☺: "*I am (task label).*" ☎: "*Okay, here's your call.*" 🕴:(request) ☺:(response) 🕴:"*Thanks. Bye.*"	☎: Greeting 2 † † ☺: "*I am (task label).*" ☎: "*Okay, here's your call.*" 🕴:(request) ☺:(response) 🕴:"*Thanks. Bye.*" ☎: "*Continue your task.*"	☎: Greeting 2' † † ☺:"*I am (task label)*" ☎: "*Okay, you are (task label).* " 🕴:(request) ☺:(response) 🕴:"*Thanks. Bye.*" ☎: "*Continue (task label).*"	☎: Greeting 2 † † ☺:"*I am (task label)*" ☎: "*Okay, you are (task label).* " 🕴:(request) ☺:(response) 🕴:"*Thanks. Bye.*" ☎: "*Continue your task.*"

Legend: ☎:: synthesized voice; ☺: participant's human voice; 🕴: experimenter's human voice
Greeting 1 †"Call from Peter, are you busy with something?"; Greeting 2 † †"Call from Peter, tell me if your busy with something, or say Nothing."

(a) Protocol of explicit auditory cue by Yeung and Li [358].

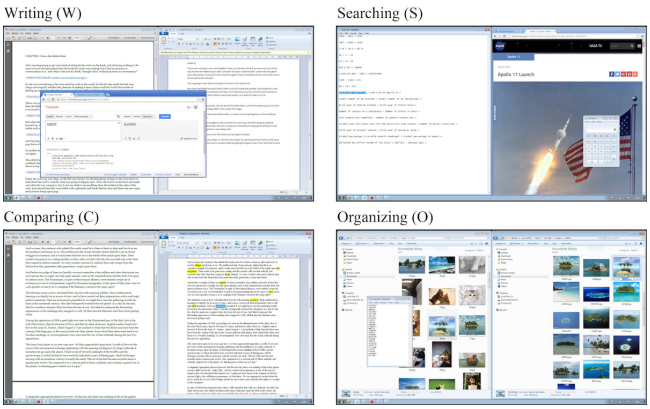


(b) Explicit visual cue: primary task interface remains visible in Hodgetts et al. [148].



(c) Implicit visual cue: gaze markers by Jo et al. [163].

(d) Tangible cue by Okundaye et al. [252].



(e) User-defined workspace cues from Jeuris and Bardram [162].

Figure 7.5: Examples of different resumption cues of different modalities and expressiveness.

Table 7.1: Task setting: device, interruption source and anticipation. There were no instances of unplanned self-interruptions.

	External Interruption		Device-Internal Interruption		Self-Interruption
	Planned	**Unplanned**	**Planned**	**Unplanned**	**Planned**
Desktop PC	Jeuris & Bardramb (2016), Liu et al. (2014), Okundaye et al. (2017)	Franke et al. (2002), González & Mark (2004), Mancero et al. (2009), Rule & Hollan (2016), Scott et al. (2006), Toreini et al. (2018), Yang et al., (2011)	Clifford & Altmann (2004), González & Mark (2004), Liu et al. (2014), Parnin & DeLine (2010)	Altmann & Trafton (2004), Cane et al. (2012), Hodgetts & Jones (2006), Hodgetts et al. (2015), Iqbal & Horvitz (2007a), Iqbal & Horvitz (2007b), Jo et al. (2015), McDaniel et al. (2004), Morris et al. (2008), Ratwani & Trafton (2007), Smith et al. (2009)	Liu et al. (2014)
Mobile		Mariakakis et al. (2015)		Yeung & Li (2016)	
Other		González & Mark (2004), Lindblom & Gündert (2017), Sadeghian Borojeni et al. (2016), Sasangohar et al. (2014)	González & Mark (2004)		
Multi-Device	Kern et al. (2010)			Cheng et al. (2018)	

such as visual or spoken text. *Implicit* cues do not contain content information. They guide the user's attention through the use of highlights such as a sound or the cursor position. They support memory without explicitly presenting task-related information. Examples of cues with different modalities and expressiveness are shown in Figure 7.5.

Visual cues can include any form of graphical display. They can consist of a small object, logo, icon, picture, or text and are intended to remind a user of a certain task or idea. In total, our literature review resulted in 24 visual cues. While text is very explicit and can convey a large amount of information, graphical visualizations such as icons or highlights are very implicit. These can take the form of colored frames around a window [328], or an underlined sentence [227]. Even simpler, gaze points can visualize the last position of gaze fixation before an interruption occurred [62]. *Textual Cues*, or written cues, are a particular form of visual cues, which can include an almost unlimited amount of information. They can be divided into system- and user-generated cues. For example, a system can generate a summary of a previously

Table 7.2: Design Space of task resumption cues.

		Visual		Auditory		Haptic	Tangible	Other
		Implicit	**Explicit**	**Implicit**	**Explicit**			
PURPOSE	**retrospective**	Cane et al. (2012), Clifford & Altmann (2004), Jo et al. (2015), Lindblom & Gündert (2016), Mancero et al. (2009), Mariakakis et al. (2015), Scott et al. (2006), Toreini et al. (2018)	Hodgetts et al. (2015), Lindblom & Gündert (2016), Parnin & DeLine (2010), Sasangohar et al. (2014)		Yang et al., (2011), Yeung & Li (2016)		Okundaye et al. (2017)	
	reminder specific	Iqbal & Horvitz (2007), Lindblom & Gündert (2016), Mancero et al. (2009), Sadeghian Borojeni et al. (2016)	Lindblom & Gündert (2016)					González & Mark (2004), Jeuris & Bardramb (2016), Rule & Hollan (2016)
	reminder unclear	Liu et al. (2014), McDaniel et al. (2004)						
	prospective	Clifford & Altmann (2004), Hodgetts & Jones (2006)	Altmann & Trafton (2004), Lindblom & Gündert (2016)					González & Mark (2004)
	spatial	Cane et al. (2012), Cheng et al. (2018), Jo et al. (2015), Kern et al. (2010), Mariakakis et al. (2015), Ratwani & Trafton (2007)						
	other	Morris et al. (2008)	Parnin & DeLine (2010)	Smith et al. (2009)	Franke et al. (2002)	Smith et al. (2009)		Rule & Hollan (2016)
ATTENTION	**full user attention**	Cane et al. (2012), Clifford & Altmann (2004), Iqbal & Horvitz (2007), Jo et al. (2015), Kern et al. (2010), Mancero et al. (2009), Mariakakis et al. (2015), Morris et al. (2008), Scott et al. (2006), Toreini et al. (2018)	Altmann & Trafton (2004), Parnin & DeLine (2010), Sasangohar et al. (2014)		Franke et al. (2002), Yang et al., (2011), Yeung & Li (2016)		Okundaye et al. (2017)	González & Mark (2004), Jeuris & Bardramb (2016)
	peripheral	Cheng et al. (2018), Hodgetts & Jones (2006), Iqbal & Horvitz (2007), Lindblom & Gündert (2016), Liu et al. (2014), McDaniel et al. (2004), Ratwani & Trafton (2007), Sadeghian Borojeni et al. (2016)	Hodgetts et al. (2015), Lindblom & Gündert (2016)	Smith et al. (2009)		Smith et al. (2009)		Rule & Hollan (2016)
TIMING	**before interruption**	Clifford & Altmann (2004), Hodgetts & Jones (2006), Lindblom & Gündert (2016), Morris et al. (2008), Sadeghian Borojeni et al. (2016)	Altmann & Trafton (2004), Hodgetts et al. (2015), Lindblom & Gündert (2016), Parnin & DeLine (2010)	Smith et al. (2009)	Yeung & Li (2016)	Smith et al. (2009)		González & Mark (2004)
	during interruption	Iqbal & Horvitz (2007), Liu et al. (2014), Ratwani & Trafton (2007)						Jeuris & Bardramb (2016)
	after interruption	Cane et al. (2012), Cheng et al. (2018), Iqbal & Horvitz (2007), Jo et al. (2015), Kern et al. (2010), Lindblom & Gündert (2016), Mancero et al. (2009), Mariakakis et al. (2015), McDaniel et al. (2004), Morris et al. (2008), Sadeghian Borojeni et al. (2016), Scott et al. (2006), Toreini et al. (2018)	Altmann & Trafton (2004), Hodgetts et al. (2015), Lindblom & Gündert (2016), Parnin & DeLine (2010), Sasangohar et al. (2014)	Smith et al. (2009)	Franke et al. (2002), Yang et al. (2011), Yeung & Li (2016)	Smith et al. (2009)	Okundaye et al. (2017)	González & Mark (2004). Jeuris & Bardramb (2016), Rule & Hollan (2016)
INTERACTICITY	**no user interaction**	Cane et al. (2012), Cheng et al. (2018), Hodgetts & Jones (2006), Iqbal & Horvitz (2007), Jo et al. (2015), Kern et al. (2010), Liu et al. (2014), Mariakakis et al. (2015), McDaniel et al. (2004), Ratwani & Trafton (2007), Sadeghian Borojeni et al. (2016), Toreini et al. (2018)	Altmann & Trafton (2004), Parnin & DeLine (2010)					Rule & Hollan (2016)
	simple user interaction	Iqbal & Horvitz (2007), Lindblom & Gündert (2016), Mancero et al. (2009), Scott et al. (2006)	Hodgetts et al. (2015), Lindblom & Gündert (2016)	Smith et al. (2009)	Franke et al. (2002), Yang et al., (2011), Yeung & Li (2016)	Smith et al. (2009)		González & Mark (2004), Jeuris & Bardramb (2016)
	complex user interaction	Clifford & Altmann (2004), Morris et al. (2008)	Sasangohar et al. (2014)				Okundaye et al. (2017)	

read passage of a text before an interruption occurred. In contrast, the users can also generate these cues themselves in the form of electronic or handwritten notes [124]. They write these notes during the interruption lag and thus capture their thoughts about what they were doing before or what they were about to do. Note-taking is highly personal and can vary immensely in the level of detail and explicitness.

The same differentiation of explicitness can be applied to *auditory cues*. On the one hand, implicit auditory cues such as simple sounds can support task resumption [312], whereas, on the other hand, explicit *verbal cues* have a higher expressiveness and can transfer more information. Examples are verbal labels, e.g., words characterizing the current task [358], or complete sentences in discourse systems [115]. Similarly to textual cues, verbal cues can be either self-recorded by the user or system-generated. Within our evaluation, four publications explore the application of auditory cues.

While *haptic* or *tactile cues* may lack the expressiveness of visual cues, they can still function as viable mechanisms to direct the focus of attention [151]. Smith et al. [312] used directional and non-directional vibration cues to guide attention towards tasks. *Tangible cues* also rely on the sense of touch, but use tangible objects. In one example, Okundaye et al. [252] proposed to use tangible Radio-frequency identification (RFID) tag cards to store the current task content to support the switching between tasks. They state that the use of different artifacts is imaginable as well, either physical or digital. Both tangible and tactile cues are mentioned in one publication of our set each, and in both cases, the cues transmit implicit information on the task.

In four cases, we were not able to categorize the cue into one of the modalities, due to the implementation of multiple facets of one TRC. For example, Rule and Hollan [287] describe multiple cues, such as the emptiness of a specific input field, a pink marking, or a cursor location, all used in combination.

Both modality and expressiveness of a TRC heavily depend on the nature of the interruption and the task which is to be resumed. For instance, visual or graphical cues are likely to be a suitable option when the primary task was a reading task [163]. One reason for this is the importance of spatial orientation in reading, i.e., remembering the line and position of the last word read. Not all cue modalities are feasible in every situation. For example, in a noisy and busy environment, it might not be possible for the user to perceive a simple auditory cue. In contrast, a visually highly demanding task could benefit from a non-visual cue.

Purpose We define the purpose of a cue as the type of information that it is intended to convey to the user. Is it supposed to give a retrospective view on what happened before an interruption occurred? Should it bring the steps that are next to be performed back to mind, i.e., give a prospective view? Or should it remind users of the current task and its state, concretely or abstractly?

The TRCs in our literature set were designed for a variety of purposes, which we clustered into six categories. The first two align with the memory-for-goal theory

(see Section 7.1.4), namely (1) *retrospective rehearsal* (indicating the last actions of a user) and (2) *prospective introspection* (showing the user's next steps). In addition, we derived three more categories based on the purposes mentioned in the specific publications: (3) *specific reminder* (a cue reminds the user of resuming a specific task or idea, e.g., [158, 288]), (4) *open reminder* (reminds the user to resume something, e.g., [213, 232]), (5) spatial awareness (guiding the visual focus on a screen, e.g., [56, 62, 163]). We included the category *other* for all publications that specified a purpose of the TRC that did not fit a category for various reasons or for cases where the purpose was left unspecified. Since a cue can have more than one purpose, we took the liberty of assigning more than one category to a single cue, which happened in seven cases. However, this differentiation, again, is based on the authors' descriptions and subjective categorizations. Therefore, it is possible to sort some of the cues into different or additional categories. Which purpose is appropriate in which context largely depends on the nature of a task: for example, in an assembly task, it is less important which steps were already executed, and in what order, but it is more essential to know which step to do next. In a learning setting, cues of many purposes could apply such as summarizing what content was presented before, making a retrospective cue potentially beneficial. Furthermore, reminders could encourage the repetition of content in short time intervals.

Attention Cues also differ in the level of attention they demand from a user. Some are perceived peripherally only, for example, when they are displayed on a secondary screen or are triggered on a device that is independent of the primary task. But the cue can also require full user attention, especially when the primary task and cue share the same presentation device or when user interaction is necessary.

In our literature set, 21 TRCs demand full user attention. Full attention means that a user has to proactively recognize, react to, or interact with the TRC. For example, in the work of Kern et al. [175], the system presents the users with a spotlight indicating their last point of view before switching between tasks. To resume the primary task, the users actively searched for the spotlight. In another example, the users have to actively recognize and trigger a playback of their last actions [159]. Fourteen cues, however, occur in the periphery of the user's attention and might or might not be noticed. For example, in the work of Hodgetts and Jones [147], the interface of the primary task stays visible during the interaction with the secondary task. Furthermore, in the work of McDaniel et al. [232], a small blue dot in the corner of the window reminds the user to resume the primary task. In this particular case, it is up to the user to notice the cue and react to it.

Timing The timing describes the moment of cue presentation during the interruption process (cf. Figure 4.1 for the visualization of an interruption timeline). In the *pre-interruption* phase, the user is in the transition between the primary task and the secondary task, i.e., in the interruption lag. If the interruption is planned or announced, the interruption lag enables the user to prepare for the upcoming secondary task [5].

The *mid-interruption* phase describes the time during the secondary task when the user is not focusing on the primary task anymore, and all cognitive resources are on the secondary task. The *post-interruption* phase is the time when the user returns to the primary task after an interruption. During the resumption lag, the time the user needs for regaining context and continuing with the original task. A common method is to present cohesive cues during both interruption and resumption lag. Moreover, it is possible to combine more than one type of cue and present those at different points of time during the interruption process.

In the literature of this review, there were instances of all three possible presentation times. Hodgetts and Jones [147], for example, examined the use of cues to encode the state of a Tower of London problem before an interruption. In four publications, the cue was presented mid-interruption. In one example, the cue was a progress bar indicating the time spent on the secondary task and thus, motivating the user to resume the primary task [214]. Another cue, which was presented mid-interruption, was to keep the window of the primary task visible during the secondary task [158, 273]. The window was supposed to be a reminder to resume the suspended task as quickly as possible. A major part of the resumption cues occurred after an interruption. Cues presented during the resumption lag are meant to facilitate context recovery [287] or restore visual focus [227, 328]. In other publications, cues are already primed during the interruption lag (pre-interruption) and then presented again in the resumption lag (post-interruption). For example, a content timeline of past actions shown before and after the interruption aims to support both retrospective and prospective goal encoding [258], helping the users to think of their previous and next steps and remember those even after an interruption.

Interactivity Cues require different degrees of user interaction. Peripheral cues typically work without user interaction, as it cannot be ascertained whether they have even been noticed. Attention-demanding cues can just as well work on a presentation-only basis without interaction, but many also require simple or complex interaction. For example, when navigating through an activity log as in [308], a user performs simple actions in the UI. Interactivity is not easy to measure – also because the actual amount of interaction changes from one use case to the next –, below we will therefore only differentiate the levels *no interaction*, *simple interaction*, and *complex interaction* for a typical scenario.

Cues that were only meant to be perceived and require no interaction were gaze markers [175], peripheral light cues [36], or the still visible primary task window during an interruption [6, 146, 158]. Other cues required the user to perform a simple interaction, such as placing a tag [225], or calling out audio labels [358]. In contrast to the number of cues which require simple or no interaction, the resumption cues rarely require more complex interactions. Only in three cases, the user is required to interact with the system in a more complex way: by taking mental or written notes [67], interacting with an event timeline to find out more about past events [293], and by choosing and

	Negative Results	Ambivalent Results	Slightly Positive Results	Positive Results
Quantitative experiments with $N > 15$	Hodgetts et al. (2015)		Jeuris & Bardramb (2016), Kern et al. (2010), Mariakakis et al. (2015)	Altmann & Trafton (2004), Cane et al. (2012), Clifford & Altmann (2004), Hodgetts & Jones (2006), Iqbal & Horvitz (2007a), Iqbal & Horvitz (2007b), Liu et al. (2014), McDaniel et al. (2004), Ratwani & Trafton (2007), Borojeni et al. (2016), Sasangohar et al. (2014), Smith et al. (2009)
Quantitative experiments with $N \leq 15$ or extensive qualitative studies		Parnin & DeLine (2010)	Cheng et al. (2018), Jo et al. (2015), Morris et al. (2008), Okundaye et al. (2017), Parnin & DeLine (2010)	Scott et al. (2006)
Anecdotal evaluation or no evaluation		Franke et al. (2002), González & Mark (2004), Rule & Hollan (2016), Yang et al., (2011), Yeung & Li (2016)	Lindblom & Gündert (2017), Mancero et al. (2009), Toreini et al. (2018)	

Table 7.3: Evaluation states and result tendencies of the publications references in the design space.

exchanging RFID tag cards to regain task context [252]. In general, the complexity of the required cue interaction increases with the complexity of the given tasks. If the user had to deal with a large amount of information at once, i.e., hold many information chunks in the working memory, the cue as well as the possible interactions would get more complex.

However, we subjectively assigned the differentiation between simple and complex user interaction in regards to the complexity of interaction shown in related work. Thus, it remains unclear how effortful users perceive the interaction with a specific cue.

Evaluations and Findings All thirty publications included in this design space were published under peer revision, either in conference proceedings or in a journal. They all presented exciting findings on the design of TRCs but the state of the evaluation and therefore, the generalizability of these findings, varied greatly. Only sixteen of the thirty publications contain experimental results with $N > 15$. Three further papers present preliminary results with small sample sizes ($N \leq 15$). Seven publications report qualitative data based on interviews and questionnaires, while two papers do not evaluate the TRC at all. Nonetheless, we included these publications since they reported interesting ideas based on literature reviews (for an overview of the evaluation states, see Table 7.3).

Among the articles that included an empirical study, the most commonly reported benefit of TRCs was that resumption times (and thus also task completion times) were shorter than without cues [36, 67, 146, 158, 159, 162, 175, 214, 227, 273, 293, 312]. Pilot studies in [163] and [328] also showed that the resumption lag decreased. Similarly, Scott et al. [308] found that in a complex scenario, the time to reach a decision was shorter with an assistive interface. In Parnin and DeLine [258] study, resumption lag was similar in all conditions. However, even in the condition where the interface provided no support, participants were allowed to take notes, and these probably served as an alternative cue. In the experiments described in various studies (cf. [36, 232, 293, 312]), cues were also found to reduce the error rate in the experiment tasks. For instance, Sasangohar et al. [293] reported a "significant increase in the mission commander's decision accuracy for both simple and complex decisions". The positive effect could not be replicated in the evaluation of the cues designed by Hodgetts et al. [148]: the duration of the decision cycles was slower and the "defensive effectiveness" was lower in the cue conditions. However, the cues they used (two types of decision-support systems) were designed in a way that made the interface more complex and even when there was no interruption, performance was worse than in the no-support condition.

The qualitative feedback collected through questionnaires or interviews confirm benefits and reveal some additional aspects. For example, radio dispatchers reported that tagging incidents facilitated their search [225] and echoing task labels after phone calls was considered a helpful reminder [358]. Using the interruption lag for labelling tasks before a phone call made participants feel prepared [288]. Similarly, Morris et al. [241]'s SearchBar eased retrieval of information in a second session, thus reducing the amount of redundant work. The RFID cards that Okundaye et al. [252] used to recover work context were praised for their immediacy when re-accessing information.

The articles also mention a number of issues and challenges that need to be kept in mind when designing TRCs. For explicit cues, the choice of content is crucial. Parnin and DeLine [258], for example, selected method names to visualize programmers' tasks, but the programmers did not consider this an effective means for triggering their memory. Especially when goals are implicit, suitable manifestations are difficult to design [287]. In some cases, participants used their own strategies that possibly interfere with the system design: Mariakakis et al. [227] noted that instead of using their gaze highlight, a participant marked her reading position through scrolling to a fixed positions and others memorized a key phrase as a "mental bookmark". Personal strategies were also observed by Jeuris and Bardram [162]: experts carefully adapted their work environment to their needs and introduced additional cues that supported task switching. [124] mentioned that on the other hand, participants did not always use tools available to them. They hypothesized that in their case, this was due to a lack of visibility of the tool and therefore stressed that artifacts need to be "visible and available". Furthermore, the interface design should not induce stress, so that negative effects of long-term use are avoided [214]. A noteworthy observation regarding the

intensity of cues was that in the work of Yang et al. [355], participants stated that in a more disruptive context, stronger cues were used.

7.2.3 Limitations

Based on our extensive literature survey, we consider our design space a valuable first step towards an inclusive but comprehensible overview of TRCs. However, we are aware that the list of publications and therefore, the design space does not present an exhaustive summary of the field. Thus, our categorization leaves potential for extension. Due to the variety of journals and conferences, we consider our choice of publications a valuable sample across several disciplines such as Human-Computer Interaction, Psychology, and Cognitive Science. Difficulties arose because there was not a fixed set of keywords that the authors used to categorize their work. To still cover a wide range of potentially relevant publications, we extended our list of query terms to include possible variations and followed up on promising references. However, it is possible that further work on TRCs is missing from our analysis due to a different terminology.

7.2.4 Discussion: Research Gaps and Promising Cue Designs

The distribution of publications in our design space shows that the predominant modality for resumption cues so far is visual: 24 out of the 30 articles presented only visual cues (cf. Table 7.2). The almost exclusive application of visual cues also means that there is potential for exploring other modalities. For instance, in mobile settings, tactile cues such as vibration patterns, could be developed further. They are easily noticeable by users themselves but provide privacy in the presence of bystanders and after some training, even patterns that encode an entire alphabet can be learned [220]. So far, little research has been done on the applications of tactile cues, especially in mobile scenarios. In the work of Smith et al. [312], the authors emphasize the tactile cues potential for use in visually busy environments in which visual or auditory cues might be inappropriate, especially for short-term interruptions.

Moreover, both implicit and explicit cues were shown to have positive effects. Ideally, a cue would demand as little attention from a user as necessary to successfully resume a primary task, so the resumption lag is kept short and it is not necessary to further increase the load on the working memory. However, it remains an open question of how explicit a cue needs to be to be effective.

The feasibility of audio cues is strongly depending on the use case. The ability to perceive audio signals in a busy environment can only be ensured with the use of headphones. It is debatable if users consider audio cues an adequate alternative to visual cues. In the design space publications, audio cues are not exhaustively evaluated, e.g., in terms of different purposes or different levels of interactivity.

Overall, clusters in the design space matrix and the evaluation results suggest that applying implicit as well as explicit visual cues can positively influence retrospective rehearsal. This goes in line with the findings of other domains such as supporting life logging by showing contents of a prior meeting [248]. Respectively, presenting a summary of a certain subset of contents learned before an interruption could, therefore, lead to higher recall rates of all contents from this lesson. Additionally, the presentation of cues before and after the interruption has been extensively researched across several modalities with emphasis on visual presentation. If the cue is perceived implicitly, the presentation of it can be either before *or* after the interruption. When designing explicit cues, showing them before *and* after the interruption has shown to be successful [212, 258]. This technique supports the encoding of information during the interruption lag, which increases a memory's chance of being recalled regardless of the interruption's intensity or complexity [254]. However, presenting a stimulus before an interruption requires the recognition of upcoming interruptions and therefore, has its limitations in the use case of mobile learning. Although the quality and robustness of attention sensing have improved over the years, the sensing mechanisms only cover a small niche of potential interruptions during mobile learning. And even if today's technology can sense boredom or disengagement [93], mind-wandering [155], or task transitions [250], their ability to anticipate those interruptions in advance remains limited.

Regarding the task and device setting, we found that although interruptions are particularly frequent in mobile use cases, there has been almost no research in mobile interruption support using TRCs. Our literature research revealed only four publications that considered interruptions on mobile devices [62, 175, 227, 358]. Some ideas that were developed for desktop PCs can probably be translated to mobile scenarios (see Design Guideline 1 in Section 7.4). However, the affordance of devices and applications used on the move differs from static settings and, therefore, additional research is necessary. Besides, in public spaces, different types of interruptions are likely to occur. At the same time, there is potential for exploring completely different ideas.

In general, the empirical studies showed TRCs can decrease task completion time and increase task performance after an interruption. However, in many other cases, the described cues are concepts which are evaluated in preliminary studies with a small number of participants only. The effect of the TRC was also not statistically proven in every case.

There is still room for improvement concerning the comparability of cues. In particular, no study empirically assessed more than two cues in the same task setting; instead, the settings varied significantly between studies. This variance was partly due to the cues serving very different purposes, but even when tasks were similar, the duration and type of interruption tasks differed. For future studies, it would be useful to revisit past publications before deciding on an interruption setting, and – if possible – include one that has previously been used in an evaluation to ensure comparability.

In conclusion, the literature survey performed in this section resulted in a variety of TRC designs. The clustering of publications into a design space along the cues' modality, purpose, attention, timing, and interactivity resulted in an interesting overview of the state of the art in this research domain. The design space highlighted frequently evaluated cue designs and yielded several gaps in research. Those hold potential for new designs of TRCs, but also call for further investigation. As described in the evaluation paragraph of this section, the publications cover a wide array of tasks and use cases, limiting the applicability of these TRC ideas. To design cues that specifically target interruptions in mobile learning scenarios, the second part of this evaluation will present a more user-centered approach.

7.3 Design Idea Generation

In the previous section, we surveyed existing literature and related work regarding the design of TRCs to support recovery from interruptions in various domains and tasks. However, the applicability of those designs for a mobile learning scenario is limited. To supplement the literature-based findings and to explore the characteristics of interruptions in mobile learning scenarios, this section presents a supplementary creative design process of two focus groups of HCI experts and learning app users.

7.3.1 Methodology

We conducted the two focus groups with human-computer-interaction and media informatics experts (having either a masters or doctorate degree in the respective field) and users to come up with novel and previously unexplored ideas for task resumption support. By asking experts' opinion, we aimed to get broad ideas, which would be discussed in the light of what could be realistically implementable. We did not intend to come up with entirely designed products, but with exciting ideas and design artifacts. This open process gives room for the imperfect, yet visionary ideas and creative thought processes. Besides, we asked users without a computer science or design background to come up with ideas which target their everyday problems with interruptions during learning. The following questions guided the focus groups:

1. From your experience - what are the reasons for interruptions in mobile learning scenarios? (User-centered)
2. How can we design task resumption support for common interruptions?
3. What are out-of-the-box / creative ideas for designs of TRCs?

Participants

For our focus group, we recruited four HCI and media informatics experts (3 female) with a mean age of $M = 29.4$ ($SD = 1.0$). All experts have formerly or are currently using learning applications on their smartphones. Moreover, we performed the same focus group procedure with three mobile learning app users with little to no background knowledge in human-computer interaction (2 female, 1 male) with a mean age of $M = 25.7$ ($SD = 2.1$). All users held at least a high school degree and have commonly used learning apps – such as *Duolingo*, *Phase 6*[9], or *Mondly*[10] – on their mobile devices in the past or present.

Procedure

To foster creativity in the generation of new ideas, we aligned our procedure with earlier work of Koelle et al. [179], who applied the *Lotus flower method* [236] (cited in [179]) to derive design ideas for privacy notices for body-worn cameras. The Lotus Flower or Lotus Blossom Method [323] (cited in [179]) is a 3-step design ideation procedure for group brainstorming sessions. With each step, the most interesting ideas are selected and will be the center of the next step to be developed further. This technique is considered structured, easy to use and explain, and useful to promote creative thinking [139, 140, 313] (cited in [179]). The main goal of this technique is to generate a great variety of ideas and take a step back from obvious design solutions.

For the expert group, we formed two subgroups of two people each to facilitate a broader range of ideas in the brainstorming process. We voice-recorded the focus group sessions and archived the sketches from the design phase. We transcribed the interviews and summarized the ratings of the sketches in an evaluation sheet.

The 3-step process (cf. Figure 7.6) was built bottom-up, starting with asking the participants to answer the following question: *What are causes of interruptions during mobile learning in everyday scenarios?*. Each item is written on a sticky note, and similar reasons are grouped. When finished, the participants agreed on three causes of interruptions they felt are most relevant during mobile learning. Those causes became the center of the new brainstorming node in step 2 asking *How could we support the user to resume a learning task when being interrupted?*. Again, the participants collected a set of task resumption strategies and picked the three most interesting and potentially helpful solutions to center step 3, the design phase. We asked the questions *How could the resumption support be designed / implemented in a learning application?* and let all participants draw sketches of as many ideas as they could come up with. Afterward, we evaluated the top 3 ideas of each group by the question *How well do you think does this idea support learners in resuming tasks after being interrupted?* on a 7-point

[9] Phase 6 App: `https://itunes.apple.com/de/app/vokabeltrainer-phase6-classic/id441493173?mt=8`, last accessed January 3, 2022

[10] Mondly App: `https://itunes.apple.com/de/app/mondly-33-sprachen-lernen/id987873536?mt=8`, last accessed January 3, 2022

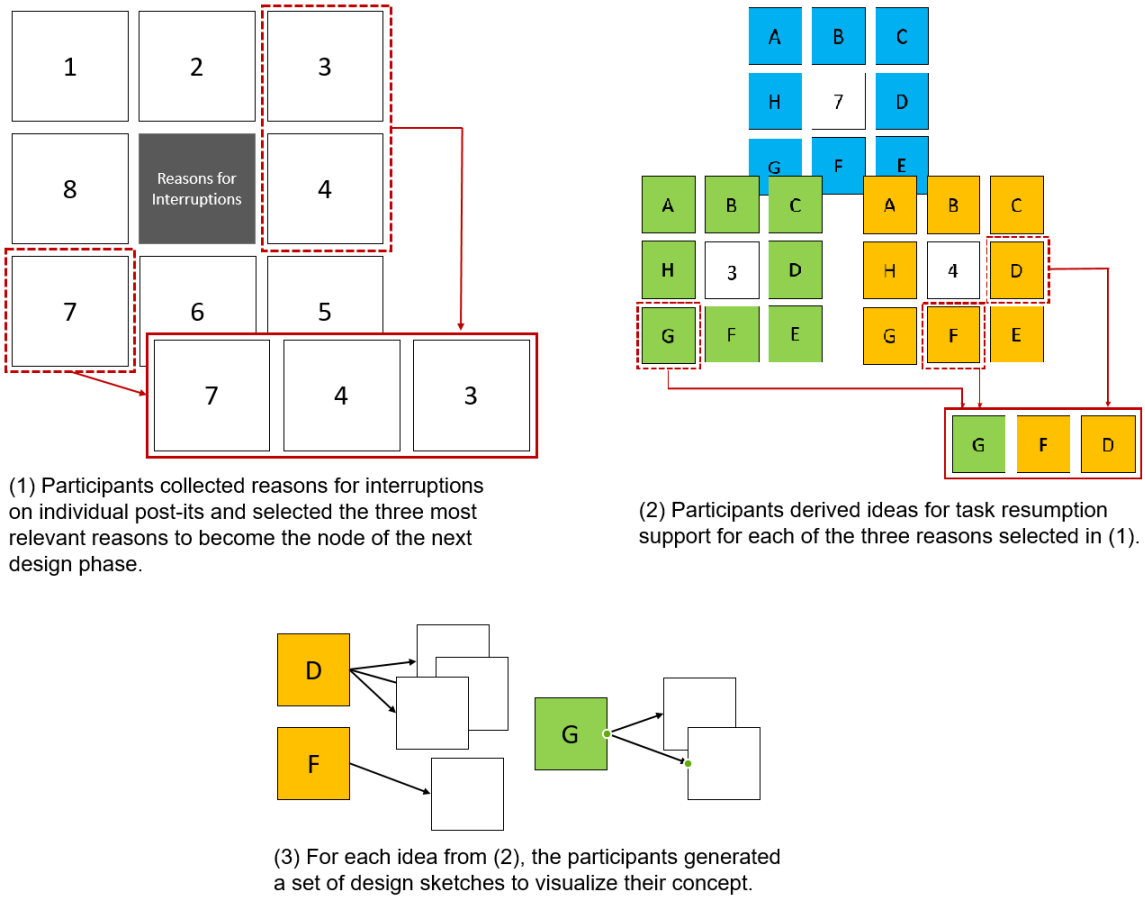


Figure 7.6: The three steps of the design process applied in both focus groups (cf. [179]).

Likert-Scale and an additional in-depth interview asking for details, advantages, and limitations of the chosen approaches. Next, we report our findings. For individual statements, we number our participants consecutively, labeling with an "E" for Expert and a "U" for User.

7.3.2 Results

Causes of Interruptions during Mobile Learning in Everyday Scenarios

This section presents the causes of interruptions the participants named during the focus groups. Since the usage situations of MLAs do not necessarily coincide with those of mobile devices in general, the interruptions potentially deviate, too. So especially for the user focus group, we asked the participants to imagine themselves during a learning activity and describe interruptions that are specific in this usage situation. By doing so, we hope to generate a list of interruptions that are specific to our use case of mobile learning. To characterize the different interruptions, we clustered them according to the

Table 7.4: Clustering of all interruption causes described by the focus group participants, aligned with the categories of interruptions as defined in Section 4.1.2. The column *Device-Internal - Planned* yielded no results and was thus removed.

		Device-Internal	External		Self-Interruption	
		Unplanned	Planned	Unplanned	Planned	Unplanned
Short duration	**Low demand**	sun shines on screen (U2)		loud neighbor (U2)/ uneasy metro ride (U2, U3)		sudden thoughts (E4)
	Medium/high demand	instant message notifications (U3, E3, E4); missing network coverage (E3)	listening to TV in the background (U1)	approached by other person in room (U2); social interactions (E3, E4)	hungry (U2, U3)	hungry (U2, U3); mind wandering (E4), cravings and needs (E4)
Long duration	**Low demand**	advertisement video (U1)				cold (U2)
	Medium/high demand	incoming phone calls (U3, E3, E4); updates and device failures (E1); low battery level (U1)	switching trains (U1); being called into doctors office (U1); daily chores (E3, E4); getting off train on a commute (E1, E2)	doorbell rings (U2, U3); walking the dog (U1, U3); social interactions (E3, E4)	end of learning time slot (E2, E4)	tired (U2); headache (U2)

schema presented in Chapter 4, Section 4.1.2 (aligned with work of [171, 238]) into self-interruptions, device-internal interruptions, and external interruptions. Furthermore, Table 7.4 presents an overview of all interruptions mentioned during both focus groups, which we aligned with our characterization of interruptions (cf. Section 4.1.2) based on detailed descriptions of the focus group participants.

Self-Interruptions The users of learning applications collected a set of 17 reasons for interruptions based on experience. Four of them concerned internal states, such as feeling tired (U2), hungry (U2+U3), cold (U2), or getting a headache (U2). Most of these interruptions are taking place when the participants are learning with their mobile device while sitting or lying on the sofa and disturb the concentration needed for the learning task. These interruption types can be either short or of longer duration and cannot easily be eliminated. The participants noted that these situations occur often and they would first target these internal problems (for example, by getting a blanket to get warm) and then resume the learning task. The experts listed mind-wandering (E4), cravings and needs (E4), sudden thoughts (E1) – such as the idea to look up something –, and the end of a self-assigned time slot reserved for learning (E2, E4) as causes of self-interruptions.

Device-Internal Interruptions and Hardware Moreover, both experts and users noted down a set of interruptions caused by the mobile device and/or the learning application itself. This list included incoming instant messaging notifications (U3, E3+E4), incoming phone calls (U3, E3+E4), and distracting advertisements (U1) within the

learning apps. Two participants furthermore mentioned hardware-related problems, such as updates and device failures (E1), a low battery level (U1), missing network coverage (E3), or the angle of the sunlight making it difficult to read on the smartphone screen (U2). The latter is described as a short-term interruption, which is likely to occur for example, when learning outside or in a train or bus.

External (Environmental) Interruptions Especially when learning at home, the participants noted distractions caused by the environment. Exemplarily, they named the the mailman ringing the doorbell (U2+U3), the neighbor being loud (U2), having to walk the dog (U1+U3), other people in the room, e.g., the partner having a request (U2), or U1 stating that "the TV running in the background causes a distraction". One participant also mentioned that they learn outside the home and named distractions such as in general people approaching them (U2). More specifically, U1 mentioned she is often learning while being in the waiting room of a doctor. Although she knows about the limited learning time in this situation, being called into the doctor's office usually interrupts her in the middle of a task. Additionally, an uneasy metro ride (U2+U3) can already distract users from the learning task, as well as having to switch trains (U1). The experts listed more general external causes such as social interruptions (E3+E4) and daily chores (E1+E2), but also getting off the train on a commute (E1+E2).

Design Ideas for Task Resumption Support

The variety of different causes for interruptions in mobile learning settings noted above shows the need for more fine grained task resumption support as well as their evaluation in mobile scenarios. The ideas for task resumption support generated by participants of both focus groups are summarized in the following:

Design Idea 1: Increase Motivation for Task Resumption In one idea, the participants described a gamification approach to keep the learner aware of the disruptive effect of interruptions. In particular, for situations in which interruptions are avoidable (e.g., when the user receives a text message that could be reviewed later), increasing the motivation to keep learning can be helpful. The participants sketched a possible interface design, which includes a tree growing at the lower right corner of the screen indicating interruption-free learning time and will produce an increasing number of fruits, the longer a person resists an interruption.

If a distraction, e.g., a notification pop-up, occurs, the phone notifies users and if they are to react to this notification right away, the tree will shrink, resulting in a loss of fruits and points. This idea picks up the idea of a visual representation such as the one presented by Liu et al. [214] or the ForestApp[11], and combines it with a gamification approach. In the work of Liu et al. [214], the user was presented in one condition with a

[11]Forest App: https://www.forestapp.cc, last accessed January 3, 2022

blooming flower if they resisted giving in to distractions, or in a second condition, with a fading flower for staying in a distracting task respectively [214]. Similarly, common MLAs like Duolingo make use of simple gamification events such as earning points for learning showing their feasibility for application in small-screen settings.

Another idea to target tiredness or feeling cold was to include more interactive learning tasks, such as having to take pictures of things for which one was supposed to learn translations (U2). After a short discussion and a readjustment of the initial task, the focus group members kept focusing on task resumption support after the interruption occurred.

Design Idea 2: Easier Comeback In case of longer and more demanding interruptions, the learning application could start with easier questions to get the user reacquainted with the topic. The participants propose to design the tasks in these situations in a short and easy manner and to include tasks which the user already answered correctly before the interruption occurred.

Design Idea 3: Adaptive Learning Modes The participants recommended implementing different learning modes within the application that would adapt the content, structure, presentation, and task resumption support. For example, a *home* learning mode adapts the MLA to a quiet environment in which the user can focus on the learning task. Thus, it would show the user tasks with high complexity and difficulty. However, a *commute* mode would expect interruptions and therefore, rather focus on shorter units or repetitions of prior tasks. Within these modes, the learning content could furthermore be adapted to the environment (e.g., when learning a new language in the train, the app could ask for translations for 'seat' and 'stop'). Using this method, interruptions due to train-related issues such as the announcement of the next stop would cause less of a distraction from the learning environment and the content the user is learning. Differentiating between usage contexts should be a feasible technique due to increased quality of sensors built into mobile devices.

Design Idea 4: Reminder One member of the focus group suggested sending reminders to the learner to resume the learning task after an interruption. In particular, reminders can support users in situations where interruptions are unavoidable (e.g., changing trains). This reminder could be explicit, but also very subtle through a simple vibration to refresh the memory and remind of the ongoing learning task. Similar reminders have also been used by McDaniel et al. [232] and Smith et al. [312]: simple visual, auditory, or tactile cues that keep users aware of background tasks they could or should return to. In an application domain, the Duolingo app issues daily notifications to get learners back on track if they do not continue their language practice.

Design Idea 5: Mnemonic Cues A solution proposed by E1 and E2 is to present the user an image at the moment of an interruption. This image would then be shown again as a mnemonic cue when the user continues learning. However, it is difficult to select an appropriate visual stimulus and if it has to be related to the learning content. Moreover, this does not work if the interruption lag is too short. Similarly, embedding the learning content in a storytelling frame would make it possible to use the method of loci for recovering context [138].

Design Idea 6: Summary – "What happened so far?" E1 and E2 suggested the presentation of automatically generated summaries, for example, in a textual form show immediately before an interruption. In addition to facilitating task resumption, reading summaries can also improve recall and transfer performance of learning contents [231]. In addition, in the case of videos or photos, visual summaries have been shown to improve recall of events [309]. In our everyday lives, we can encounter summaries when watching a TV series: episodes often feature a recap of relevant events in past episodes. However, generating meaningful summaries remains a challenge, as the relevancy of contents needs to be established or corresponding higher level concepts be identified. Summaries could also be presented as a set of questions for the user to answer upon task resumption as proposed by E3 and E4. Questions can help guide the learners back to the task and furthermore be a tool to adapt the following content according to user's performance on these questions. Asking questions is a common pedagogical method used in school to get students reacquainted with the content of the previous lesson [185] and could be easily realized inside an MLA.

In addition to the function as a resumption cue, retrieval practice through testing has also been shown to improve long-term retention and improve knowledge transfer to new contexts [280]. In particular, testing is beneficial if some time has passed since an item was studied – for example, if the learning task was interrupted.

Design Idea 7: Regaining focus The participants E3 and E4 proposed to include short meditation exercises to regain focus, for example, focused breathing. Although the exercise is not specifically related to the learning task, it can help users focus on the upcoming task again. Research has shown that mindfulness interventions can increase the attention level in subsequent tasks [55] and decrease mind-wandering [242]. Besides, short breathing exercises have been found to improve reading comprehension [68]. These findings suggest that integrating mindfulness components into the interruption recovery process of an MLA could indeed support task performance. Campillo et al. [55] showed that visual and auditory mindfulness-based interventions improved subsequent auditory and visual memory and attention. Thus, this technique is particularly interesting to counteract self-interruptions such as mind-wandering, during which people's focus of attention shifts to internal thoughts. However, the duration of such interventions needs to be evaluated.

7.3.3 Discussion and Limitations

The participants of the focus groups engaged in interesting discussions around interruptions in the mobile learning context and often referred back to their own previous experiences. They quickly came up with a large number of situations where interruptions occur, which shows that there seem to be many sources of distractions. Causes of self-interruptions were mentioned frequently, suggesting that the physical needs and mental state of learners must be taken into account for a seamless learning experience. Self-interruptions were closely followed by device-internal or external social interruptions (e.g., messaging notifications) – situations, where interruptions are unplanned and TRCs are particularly promising.

During the design phase, the participants came up with ideas that build on different aspects of learning and mobile tasks: content-related strategies that sum up previous content or gradually change from prior topics to new content, but also content-independent concentration exercises. Most of the designs applied visual cues, but this was probably induced by the fact that we asked them to draw their ideas and this modality seemed the most obvious. Some of the concepts are similar to what has been used in research or existing applications, for example, the aforementioned summaries and reminders that motivate users to get back on track. Other ideas have not yet been examined in detail, especially in the context of mobile learning, and more in-depth research would be a valuable contribution. This includes situation-aware mnemonic cues, easing back in as in Design Idea 2, and the effect of concentration exercises. Further discussion of the participants' proposals can be found in Section 7.4, where we use them to extract a set of design implications.

Due to the composition of the focus group, the participants we selected reflect only a narrow sample of the overall population of MLA users and experts. The participants who took part in the focus groups are all between 20 and 30 years old and are mainly using language learning applications. However, we do believe that this sample represents a very important age group as they are from a generation that has grown up with technology, but already come to a point of maturity and developed daily routines. Nonetheless, additional focus groups with a larger and more diverse sample would definitely be a benefit for the design of useful TRCs.

To our surprise, the participants of the user focus group showed difficulties staying on topic during the session. Although we explained the use case of task resumption very thoroughly and confirmed their understanding, they often deviated from this original topic and discussed how interruptions could be avoided and managed. This is of course also a central process when dealing with interruptions, it is, however, not per se a tool to support the resumption of a task after an interruption. We consider this observation an indication that users either have difficulties understanding and imagining this specific situation, or would rather prefer a tool to avoid interruptions rather than task resumption support.

Although we did explicitly not restrict the focus group participants to the capabilities of today's technology, the design ideas are of course influenced by technological limitations. Due to the rapid development of the computing power and sensing quality of mobile devices, the design ideas for task resumption support could become more complex in the future. However, until now many of the creative ideas participants came up with come with certain restrictions. In particular, sensing certain situations or interruptions is not yet possible. Furthermore, the process of sketching design ideas with pen and paper could have influenced participant's choice of cue design. It is obviously easier to sketch a visual design rather than, for example, an auditory cue. Nonetheless, we encouraged the participants to at least take notes on their ideas or include other modalities such as sounds through abstractions (e.g., draw icons).

7.4 Design Implications

So far, this chapter described a literature review, which brought up several task resumption concepts in various disciplines. We clustered the results according to various design space categories and evaluated their potential for application in a mobile learning scenario. This section will now evaluate the potential and limitations of individual features and functions for the application in a mobile learning scenario. We derive a set of six design implications for TRCs in mobile learning scenarios to support both researchers and designers.

Design Implication 1: Adapt to the Mobile Device's Qualities and Requirements The TRCs proposed in the design space literature are almost exclusively designed for static desktop computer settings. On the other hand, some existing mobile apps include resumption support features but have not yet been evaluated from a research perspective. We strongly believe that many resumption strategies described in the design space could be translated to mobile settings, i.e., can be adapted to smaller screens and more simple interactions. If TRCs are adapted from stationary settings, it is important to simplify the interface to work with a small screen and to use input and output methods that smartphones provide. One further factor playing in the favor of mobile devices is the fact that resumption cues are often designed such that they require only a limited amount of interaction (cf. Table 7.2), which means they can more easily be integrated into a simple UI. For instance, timeline views, which present the user with a chronological overview of the past actions or created artifacts (as in [148, 241, 258, 308]), could be adapted by simplifying the UI to make sure it does not become too cluttered. Moreover, activity replay is also a possible option – and has already been implemented in many media players, where users can repeat previously played content. In the particular case of mobile learning, the content to be repeated already exists and needs no or almost no modification for generating a replay. Replay in contexts where salient points need to be extracted first is more difficult to implement, but we deem the effort very promising from a task recovery point of view. To avoid overloading the user with

too many information, a details-on-demand design can be implemented, where users can toggle the display of additional information as needed or desired.

Design Implication 2: Make Use of the Interruption Lag Interruptions on mobile devices are manifold. Some of them might be unpredictable such as external interruptions by other people or the environment. It is close to impossible to anticipate them. In contrast, some interruptions can be foreseen and/or delayed, such as internal notifications through the *Attelia* sensing system [250], which detects breakpoints of users' activities during smartphone usage without the use of any additional sensors. TRCs that guide learners during the interruption lag and prepare for upcoming interruptions have shown a high potential for support. The time before an interruption can be used to encode the information to increase the chances of long-term-memory storage as well as for goal encoding ("What was I about to do next?") [5]. Briefly presenting a visual cue in the transitional period before an interruption can already foster mental note-taking and thus, task resumption [67]. More explicitly, the interruption lag could be used to present the user with the current state of the task to remember [6] or to show a tree view of all lesson parts or a timeline of the learner's last actions [258]. Since the use of TRCs during the interruption lag has been shown to have a positive effect on task resumption, we can recommend to facilitate this time if an interruption can be anticipated (even if the anticipation time is short). However, there are contradicting findings regarding the length of the interruption lag, which need further evaluation [6].

Design Implication 3: Leave App Visible During Interruption One of the most thoroughly evaluated type of TRCs is presented during the interruption. If a user is disrupted by a secondary task originating device-internally, such as an incoming notification, it will help to leave the task window at least somewhat visible. Multiple studies (cf. [147, 158, 274]) have shown the positive effect of leaving part of the task screen visible on task resumption time after short interruptions. Showing a secondary task in a corner, or at least showing as much of the primary task as possible can ease task resumption [158]. Implementing a split-screen mode to keep the learning task visible could therefore enhance resumption speed. Since the aforementioned user studies applied either a desktop or a multi-device setting, it has to be further evaluated to what extent the results are transferable to a mobile device. For example, the study of Iqbal and Horvitz [158] found longer resumption times in windows that were less than 25% visible compared to windows that were more than 75% visible. However, on a smaller screen these results might deviate. Further evaluation needs to investigate if even smaller hints such as icons or objects like the Facebook Messenger Chat Heads can achieve similar results in reminding users to resume a learning task.

Design Implication 4: Cue Complexity Should not Exceed Task Complexity In common learning applications, the design of a task creates a unique set of requirements for the user. Many common learning applications only address maintenance rehearsal

processes. For example, language learning apps such as Duolingo only focus on numerous repetitions of vocabulary. These apps commonly neglect the explicit teaching of structures or grammar knowledge [135]. The passive processes of maintenance rehearsal require less focused attention and less cognitive resources than deeper processing mechanisms. Other applications, e.g., for science learning or coding, might require more explicit processing mechanisms and therefore, involve LTM encoding. In contrast to basic maintenance rehearsal, which takes place in the working memory, exercises like this would require deeper processing. So both app and learning content have an effect on the depth of processing required and, thus, influence the negative impact an interruption can have. In conclusion, maintenance rehearsal learning can be supported by simple cues such as summaries or further repetition tasks because learning contents are created to be short and simple (as common in micro-learning). In this situation, interacting with a complex TRC could produce additional cognitive load through the task design, which can hinder central processes of learning such as schema construction. In general, many TRCs of very simplistic design such as visual highlighting [227] or simple audio cues have shown their potential to support the user [312]. However, participants of the focus group discussed to design cues with different levels of complexity and explicitness. For example, participants considered pictures (e.g., screenshots) potentially more effective in helping one recall where a learning task was left off, as opposed to reading summaries. The efficiency of cues for task resumption with different levels of explicitness and complexity needs to be evaluated with regards to the strength and cognitive demand of interruptions. Especially when designing for complex learning tasks, we need to investigate different degrees of cue complexity.

Design Implication 5: Evaluate Different Cue Modalities Related work described in the design space applies a variety of modalities for TRCs, including visual, textual, and auditory. The modality varied according to the interruption context. For example, the evaluation of auditory labels happened in a setting of an urgent and important interruption because user can quickly and easily generate and retrieved auditory cues [358]. The adaptation of modalities to different types of interruptions has not been evaluated in particular and has to be explored in future work. Additionally, we want to highlight the potential of marginally researched modalities such as tangible or haptic cues. Complementary to the translation of cues from desktop settings, mobile devices open up new design opportunities. Input and output methods provide potential for a range of new designs such as haptic cues. Mobile devices are usually equipped with vibration motors and thus, are suitable for haptic cuing. Vibration signals are a popular method for notifications [289] and are discreet. If the device is worn on the body, vibrations can also be noticed in busy, noisy environments as well as when the phone is in the pocket or bag, in contrast to visual or auditory cues. Participants of the focus group suggested using simple vibrations for reminding users of suspended tasks. In theory, vibration patterns could convey a lot more information than they currently tend to do. However, it is important to assure a good learnability of new or complex cues (cf. [88]) to assure that the users can recognize the encoded meanings if it exceeds

simple reminders. Therefore, for task resumption, we recommend using haptic cues preferably for implicit cuing. Moreover, the combination of multiple modalities as well as their use for a broader range of purposes need to be further explored. For the use in learning, it is however important to align the task and task resumption design with the multimedia principles defined by Mayer [230]. He states the importance of designing instructional messages during learning in the light of the workings of the human mind and claims that following these principles will lead more likely to meaningful learning. Even though Mayer supports multi-codality, meaning the use of different modalities (e.g., words and pictures), he also highlights the negative effect of superfluous or redundant materials for learning [229, 230]. Thus, when designing for learning tasks, the use of different modalities can foster the learning process as long as they are carefully used.

Design Implication 6: Evaluate the Cue in Different Situations For the evaluation of any new cue design it is important to consider a broad range of possible usage scenarios. As we described in Chapter 3, mobile learning can take place in a variety of situations and surroundings. These situations vary in their basic characteristics such as the time that can be spent on learning right now, or the noise of the environment. Many of the TRCs described in our analysis only test their designs in a very narrow usage scenario and often perform very controlled laboratory evaluations. However, the effectiveness of TRCs strongly depends on the task and setting they are presented in as well as the characteristics of the interrupting task itself. In the focus group discussion, we collected a variety of possible interruptions that can occur during learning when mobile. We recommend to evaluate TRCs in regards to these common interruptions.

7.5 Chapter Summary

In this chapter, we took a deep dive into literature on memory cues to support task resumption from various domains. We searched digital libraries and carefully selected a set of 30 publications that contain 35 designs and/or evaluation of (prototypical) TRCs. We used this literature set to generate a design space spanning multiple dimensions, namely modalities, expressiveness, purpose of the cue, level of attention required, timing of cue presentation, and interactivity of the cue. With regards to our research question **RQ4** we contribute potentially promising cue designs for task resumption support in mobile learning scenarios.

We emphasize that research on TRCs in mobile and uncontrolled environments is still sparse and the effectiveness of memory cues might be strongly influenced by the specific interruption and situation of the user. We see great potential for future work to evaluate promising cue designs in-the-wild to understand their helpfulness for supporting learners even better. Therefore, the next step in this domain is derive concrete designs for TRCs based on the theoretical recommendations made in this chapter.

8

Memory Cues in Mobile Learning Applications

The previous chapter presented a detailed literature analysis on task resumption cues (TRCs) in different application domains. Looking into research gaps and promising cue designs, Chapter 7 outlined design suggestions for potential TRCs for the application in mobile learning scenarios. To investigate these theoretical recommendations in regards to their ability to support users in resuming interrupted learning tasks, this chapter presents two evaluations.

In Section 8.2, we propose a set of concrete designs for memory cues. These designs were implemented in a mobile language learning (MLL) application to test their effectiveness for supporting learning tasks in a controlled lab-based user study. We report on the results of this study and present a revised set of memory cues. To compensate for the limitations of a laboratory experiment when it comes to evaluating a mobile application, a follow-up study presented in Section 8.3 examined the TRC concept in the wild. We followed the recommendation of Lazar et al. [206] and combined these two methodologies to generate complementary insights.

Overall, this chapter provides two detailed evaluations of TRCs in different scenarios. It discusses the implications of the studies and outlines potential challenges and opportunities for the use of TRCs in mobile learning applications.

Parts of this chapter are published as follows:

- Schneegass, C., Füseschi, V., Konevych, V. & Draxler, F. (2021). Investigating the Use of Task Resumption Cues to Support Learning in Interruption-Prone Environments. In Multimodal Technol. Interact. 2022, 6, 2.

This section is supported by the Master thesis of Vincent Füseschi (Section 8.2) and the Bachelor thesis of Viktoriia Konevych (Section 8.2), see detailed collaboration statement at the beginning of this thesis.

8.1 Related Work

Chapter 7 extensively discussed applications of TRCs in other domains and distilled recommendations for their application in mobile learning scenarios. In this section, we briefly summarize the central findings from related literature for the use case of mobile language learning.

8.1.1 Promising Features of Task Resumption Cues for Mobile Learning

A central observation from the design space created in Chapter 7 was the thorough evaluation and prevalence of visual presentation regarding the cue **modality**. Both implicit (e.g., highlights) and explicit (e.g., summaries) visual memory cues show great potential to support task resumption in various domains. For the nature of learning tasks, the chapter already outlined that cues could fulfill two main **purposes**. First, as general reminders, TRCs could guide users back to the activity itself. Second, summarizing the content presented before, i.e., aiming at fostering retrospective rehearsal, could help users regain the actual task context after an interruption [329]. Since many interruptions can not be anticipated or detected in advance (cf. Chapter 4), adjusting the cue **timing** to during or after the interruption is the only feasible option. Particularly, since we can not always sufficiently distinguish between a temporary interruption and a permanent task switch (e.g., when the user switches apps on the phone), we consider the provision of task resumption support when re-entering the learning application as most promising. Cues presented during the resumption lag can facilitate context recovery [287] or prospective goal encoding [258]. Depending on whether the cue is presented during or after the interruption, the expected **attention** from the user can range from peripheral to full attention. For explicit cues, a certain attention level is necessary for sufficient processing of the displayed content such as in timeline views or other artifacts (cf. [148, 241, 258, 308]). Lastly, the design space differentiated between three levels of cue **interactivity**. While the majority of the evaluated cue designs did not require the users to interact as they were implicit or peripheral, we consider both the simple and more complex user interactions promising for learning applications. Similar to elaborate rehearsal techniques used in pedagogical approaches [185], triggering users' memory of prior content through questions can help them remember.

8.1.2 Measuring Task Resumption Efficiency

The implementation of TRCs in prior work affected the users' interaction with the respective system. As described extensively in Chapter 7, the majority of evaluations looked into their quality for averting the negative effects of interruptions. These effects can be evaluated using subjective and objective measures. As objective metrics, prior work frequently focused on assessing the resumption time (e.g., [36, 67, 146, 158, 159]), often used synonymously with task completion time. The metric measures the time the user needs to resume the task, which is usually increased due to interruptions. The prolonged resumption time is called *resumption lag* (cf. [163, 328]). The application of task resumption cues is expected to mitigate these effects and thus, their effectiveness can be observed by a reduced resumption lag. Further, the error rate can be an

indicator that the interruption affected the users' performance. Therefore, lower error rates are used to measure TRC effectiveness (cf. [36, 232, 293, 312]).

Since the users' experience of the cue helpfulness can deviate from the effects observed through the objective measures, combining objective and subjective metrics can reveal further insights. Standardized questionnaires such as the System Usability Scale (SUS) [42] or specifically designed survey or interview questions can further assess user experience when interacting with TRCs in mobile learning applications.

8.2 Task Resumption Cues in the Lab

In the first part of this chapter, we present the implementation of a mobile language learning application and the design of TRCs. This application was tested in regards to objective and subjective metrics in a laboratory environment. In such a setting, we were able to control the strength and origin of interruptions and can therefore draw informed conclusions regarding the precise effects of memory cues on learning behavior.

8.2.1 Implementation

To perform the laboratory experiment, we developed an iOS language learning application as basis for the evaluation of the TRCs. In it's structure, design, and language content, we aligned our app with common market applications such as Duolingo[1], Memrise[2], and Babbel[3].

Lesson and Interruption Design

The app included several self-contained vocabulary lessons of 20 questions each, grouped into lessons by themes such food or clothing. Every lesson consisted of two parts and focused on explicitly teaching vocabulary and implicitly teaching simple grammar constructs through sentence building tasks. The first part of each lesson consisted of active and passive recognition tasks [205], in which users had to select the correct L1/L2 translation for a displayed L1/L2 word from a set of alternatives (multiple-choice). Pictures were used to help the initial acquisition of new words. The second task format was to translate a L1/L2 sentence by assembling words through sequential selection from a pool of words. The sentence-building task is considered more difficult as it presents more options and thus, does not allow for mere elimination of incorrect answer options to solve the task. We decided use Polish as L2 language

[1] Duolingo: `https://www.duolingo.com/`, last accessed January 3, 2022

[2] Memrise: `https://www.memrise.com/`, last accessed January 3, 2022

[3] Babbel: `https://babbel.com/`, last accessed January 3, 2022

as it relies on the same base alphabet as German without being too similar, making it neither too difficult nor too easy to learn.

To distract participants from the learning task and test the effect of the TRC designs, we interrupted them several times during the learning task. We chose to use mathematical tasks as interruption as it has been previously used in other studies as interruption source (cf. [76, 77, 255]). Inside the application, the users were shown a series of five double-digit multiplication tasks (see Figure 8.1d) they had to solve before they could continue learning. Such device-internal interruptions are frequent in smartphone usage and have been most common in the literature analysis presented in Chapter 7.

8.2.2 Cue Designs

As Task Resumption Cues, we included four different designs in the learning application. The cues were presented as a full-screen overlay once the user reenters the learning application after an interruption. Since the learning tasks contained beginners-level content, we followed Design Implication 4 from Chapter 7 and designed them so that the cue complexity did not exceed the task complexity.

We chose to include two implicit cue designs, labeled *Half-screen Cue* and *Image Cue*, and two explicit cue designs, labeled *History Cue*, *WordCloud Cue*. While the implicit cues aimed at bringing the user back into the context of the lesson, the explicit cues showed the user actual content they learned before, as explained in more detail below. All cues were of visual modality, as the literature review in Chapter 7 emphasized their effectiveness to support retrospective rehearsal. We further decided to only present the cues in the resumption lag and not in the interruption lag. Due to the fact that we aim to deploy this application concept in the wild later on, we can not expect interruptions to be anticipated in everyday settings. Therefore, the presentation of cues in the interruption lag will not be possible.

History Cue The most explicit memory cue to support the resumption of the learning task was designed to show the user their progress over time. As shown in Figure 8.1c, this cue visualized the last lessons the user performed as well as their solution, thus, providing a sense of context to them. Visualizing progress history has been explored in prior work, for example, in programming settings [258], search tasks [308], and mission command or aircraft tasks, where it has shown its potential to increase accuracy and performance [164, 293].

WordCloud Cue This memory cue displayed priorly learned words in the form of a tag cloud (see Figure 8.1b). It is closely related to the *History Cue* as it is a summary of content learned before the interruption. However, it is less structured since it did not include a temporal component of when the word was presented or learned but

gives a more general overview. The concept of word or tag clouds gained popularity in other application areas for the summarization of text analysis tasks [136] or search results [193].

Half-screen Cue The *Half-screen Cue* left a part of the primary task UI visible while the secondary task was performed (see Figure 8.1d) as recommended in Design Implication 3 of Chapter 7. Thus, in contrast to the other resumption cues, it was shown during, and not after, the interruption. This cue is based on several studies conducted in desktop settings. For example, when the primary task UI remained partially visible, study participants were better at maintaining a spatial representation of the primary task [273], more quickly returned to prior tasks [158], and had a shorter resumption lag [146]. We consider the *Half-screen Cue* an implicit cue because it did not include any additional information about the current topic.

Image Cue In this cue, an image or graphical symbol was shown representing the lesson the user interacted with before an interruption. This form of visual memory cue was already suggested in prior work (cf. [95]). We consider it an implicit cue because it is a very simple reminder hinting the content of the prior lesson. Further, Chen et al. [61] noted that instructions in learning should be targeted to the individual learner. While some benefit more from verbal information (e.g., words in the *WordCloud Cue*), others are better supported with visual information such as images. The image was selected from a pool of images used in the lesson for teaching vocabulary. In the lesson, the image supported the learning of content as proposed in the multimedia principle of Richard Mayers' Cognitive Theory of Multimedia Learning [230].

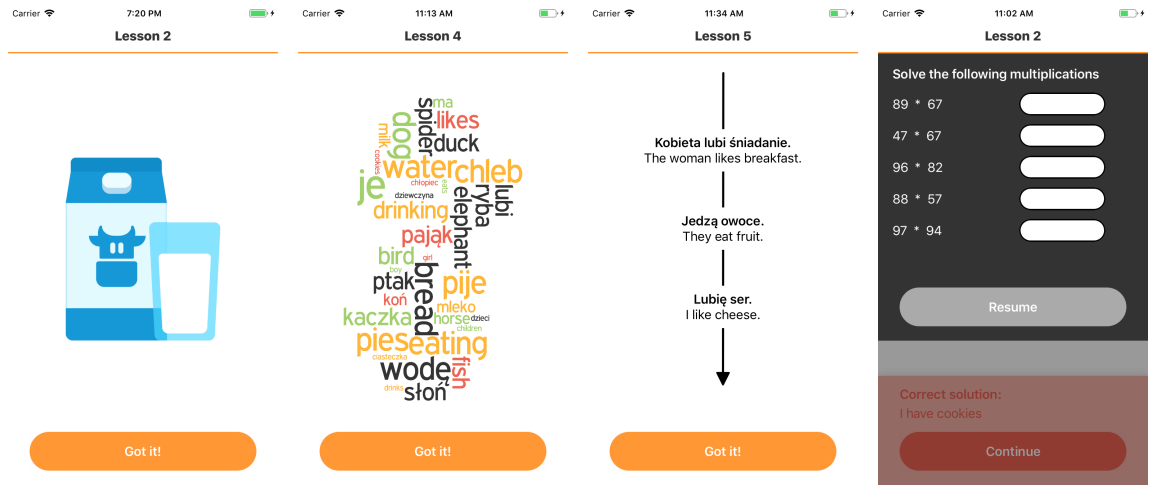
8.2.3 Methodology

Study Design and Apparatus

We performed a within-subject laboratory-based user study investigating the effect of the four different cue types (independent variable) on the participants' task performance. In particular, we measured the error rate as well as the answer duration (dependent variables) after the interruption occurred and the cue was shown. Based on prior work we formulate the following hypothesis:

H1 The error rate and answer duration for tasks following an interruption decrease when the task resumption is guided by a memory cue as compared to the control condition without a cue.

The order of presentation of the cues was counterbalanced over the course of the five content lessons. Each lesson was interrupted once and the interruption was either followed by one of the cues or resumed immediately (no cue condition). Furthermore, we assessed the perceived helpfulness (dependent variable) of the different cues through Likert-scale ratings and a qualitative interview after the study is completed.



(a) Image Cue (b) WordCloud Cue (c) History Cue (d) Half-screen Cue

Figure 8.1: The four task resumption cue designs created for the application in mobile learning, (a) *Image Cue*, (b) *WordCloud Cue*, (c) *History Cue*, and (d) *Half-screen Cue*. The cues are presented when the user resumes the learning task after an interruption to help task resumption.

Procedure

At the beginning of the study, participants were informed about the procedure and asked for consent. Then, they filled in a questionnaire to assess demographics and prior knowledge as well as experience with language learning applications. Next, the participants were given the study task – to complete five lessons in the mobile learning application designed for this study. We also provided pen and paper to help solve the multiplication interruption tasks and logged every interaction with the application to assess answer duration and error rate. As the task time was measured, we asked our participants not to take breaks during the lessons but instead between two lessons. However, they were not encouraged to rush but take as much time as they needed to answer the presented tasks correctly conscientiously. Finally, we conducted post-hoc semi-structured interviews to inquire about the perception of the different cues. The guiding questions of the interview concerned their general opinions of the cues, the cues' helpfulness, and the match of cue types for materials of different complexity levels.

Participants

We invited participants through university mailing lists and social media channels, resulting in a set of 15 participants (8 male, 7 female) ranging between 20 and 33 years of age ($M = 23.7$, $SD = 3.6$). None of the participants had prior knowledge of Polish or any closely related languages such as Czech, Slovac, Russian, or Silesian. The

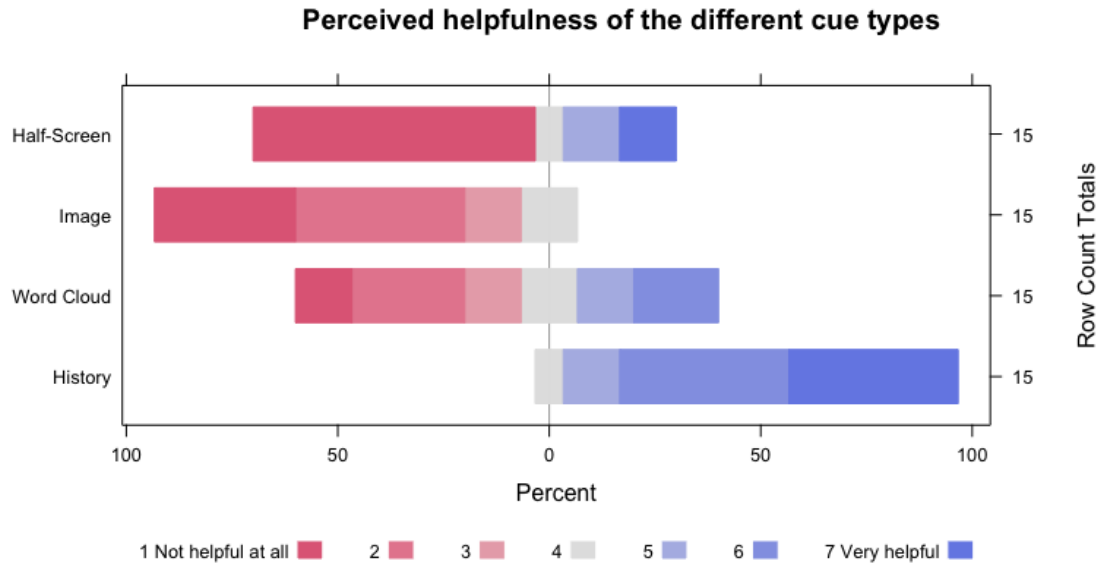


Figure 8.2: Perceived helpfulness of the different cue types.

user study took around 1 hour, and as compensation for the participation, everyone received a 10€ voucher or an equal amount of study credit points.

8.2.4 Results

Our data set included 75 learning sessions (five per participant), 75 interruptions, and the presentation of 60 cues; Each cue type appeared 15 times plus 15 occurrences of the no cue condition). The total number of exercises the participants interacted with is 1485. Post-hoc ratings showed that the lesson content was challenging for the participants, as three participants rated it "very difficult", seven as "somewhat difficult", five as "adequate", and none as "somewhat easy" or "very easy" (7-point Likert scale). Out of 15 participants, 14 would appreciate it if language learning applications displayed such TRCs for everyday usage.

Perceived Helpfulness We asked our participants to rate the helpfulness of the cue types on a 7-point Likert scale from 1 (="Not helpful at all") to 7 (="very helpful"). While the *Half-screen Cue* ($M = 2.53$, $SD = 2.28$), *Image Cue* ($M = 2.07$, $SD = 1$), and *WordCloud Cue* ($M = 3.47$, $SD = 1.75$) were perceived as semi to little helpful, participants valued the *History Cue* ($M = 6.13$, $SD = 0.88$) and stated that it was the most helpful cue by far (cf. Figure 8.2). A (non-parametric) Friedman test showed that there were significant differences between the perceived helpfulness of different cue types ($\chi^2 = 26.1$, $p < .001$; Kendall's $W = 0.37$). Post-hoc Conover comparisons with Bonferroni correction revealed significantly higher helpfulness ratings of the *History*

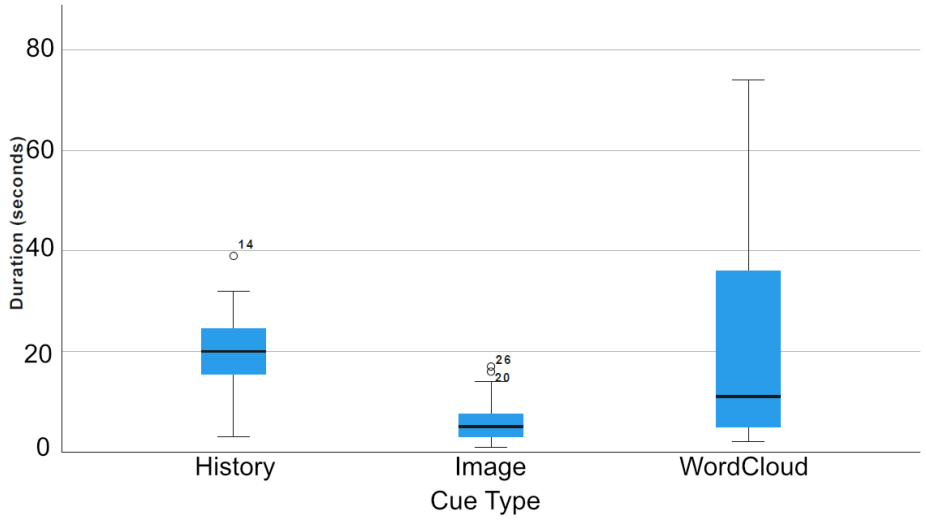


Figure 8.3: The viewing duration in seconds of the three cue types *History Cue*, *Image Cue*, and *WordCloud Cue*.

Cue over the *Half-screen Cue* ($t = 4.45$, $p < .001$), the *History Cue* over the *Image Cue* ($t = 4.30$, $p < .001$), and the *History Cue* over the *WordCloud Cue* ($t = 2.62$, $p < .05$).

Post-hoc Interviews During the interviews, participants confirmed that they liked the *History Cue* most. However, most participants did not notice that the questions displayed there were the last three before the interruption, but assumed the cue displayed a random set of questions. One participant stated that they preferred the *WordCloud Cue* over the *History Cue* as it presents more details. As the *History Cue* can technically present more complex information than the *WordCloud Cue*, participants suggested to use it for presenting grammar knowledge and the *WordCloud Cue* for vocabulary.

For the *Image Cue*, four participants reported not noticing the cue or not looking at it at all. In particular, one of the *Image Cue* images, a Polish flag (horizontal white and red stripe), was not recognized by two of our participants. One of these participants thought it might be an indicator of correct and incorrect questions. Out of the 15 participants, only five noticed the *Half-screen Cue*, of whom three thought it might be a bug in the UI. Overall, all participants considered the implementation of TRCs to be a very helpful feature for mobile learning apps. However, they emphasize that the actual helpfulness depends on the cue design.

Cue View Duration The viewing duration of the different cues varied among the three cue types and among participants (as the *Half-screen Cue* was implicitly embedded in the interrupting task, this cue has no viewing duration). Figure 8.3 shows that the *WordCloud Cue* was examined by the users with the greatest diversity in duration, between 2 and 74 seconds, with an average viewing time of 11 seconds ($SD = 20.67$). In

Table 8.1: Mean response time (and standard deviation) and mean correctness rate (and standard deviation) with different cues for the first or all exercises after an interruption

	No Cue	**Half-Screen**	**Image**	**WordCloud**	**History**
Response time 1st exercise (s)	19.5 (6.2)	22.5 (13.3)	30.6 (29.7)	23.2 (13.7)	25.4 (21.6)
Response time all exercise (s)	16.1 (7.8)	18.3 (11.0)	19.1 (15.9)	18.7 (12.4)	20.8 (16.8)
Correctness rate 1st exercise (s)	0.40 (0.51)	0.47 (0.51)	0.20 (0.41)	0.27 (0.46)	0.20 (0.41)
Correctness rate all exercise (s)	0.49 (0.34)	0.51 (0.32)	0.48 (0.29)	0.49 (0.29)	0.53 (0.34)

comparison, the *History Cue*'s viewing duration shows less variety among participants, ranging from 3 to 39 seconds but is higher on average ($M = 19.8$, $SD = 9.34$). Lastly, the *Image Cue* is viewed for the shortest duration between 1 and 17 seconds ($M = 6.27$, $SD = 5.12$). We found a significant difference in means between the three types (one-way ANOVA[4], $F = 5.82$, $p < .01$) and post-hoc comparisons with Bonferroni correction revealed significant differences between *History Cue* and *Image Cue* ($p > .05$) and *WordCloud Cue* and *Image Cue* ($p > .01$).

Task Completion Time On average across all exercises, participants needed 18.01 seconds ($SD = 13.42$) to complete one learning task. We find that the response time for the first question after an interruption ($N = 75$) differed significantly from the questions before the interruption ($N = 900$) and all subsequent questions ($N = 510$)(repeated-measures ANOVA, $F = 9.0$, $p < .01$, see Figure 8.5a). Compared to the average response time across all exercises, the average response time for the first question after an interruption was 24.24 seconds (cf. Table 8.1). Post-hoc comparisons with Bonferroni correction showed that the task completion time for tasks before an interruption was lower as compared to the first task after an interruption ($t = -3.6$, $p < .01$). Similarly, the task completion time for the first task after an interruption was higher than for compared to all other following tasks ($t = 3.7$, $p < .01$). In other words, the first task after an interruption took users significantly longer to answer than any other task.

In regards to our hypothesis **H1**, which stated that the presentation of TRCs can reduce the task completion time after an interruption, our analysis remains inconclusive. A repeated-measures ANOVA revealed no significant differences ($p > .05$) when comparing the cue conditions with the no cue condition. In fact, the overall response time

[4] A Komogorov-Smirnov test indicated a violation of the normality assumption ($p > .05$). However, due to the robustness of ANOVAs in regard to this violation, we continued with this analysis.

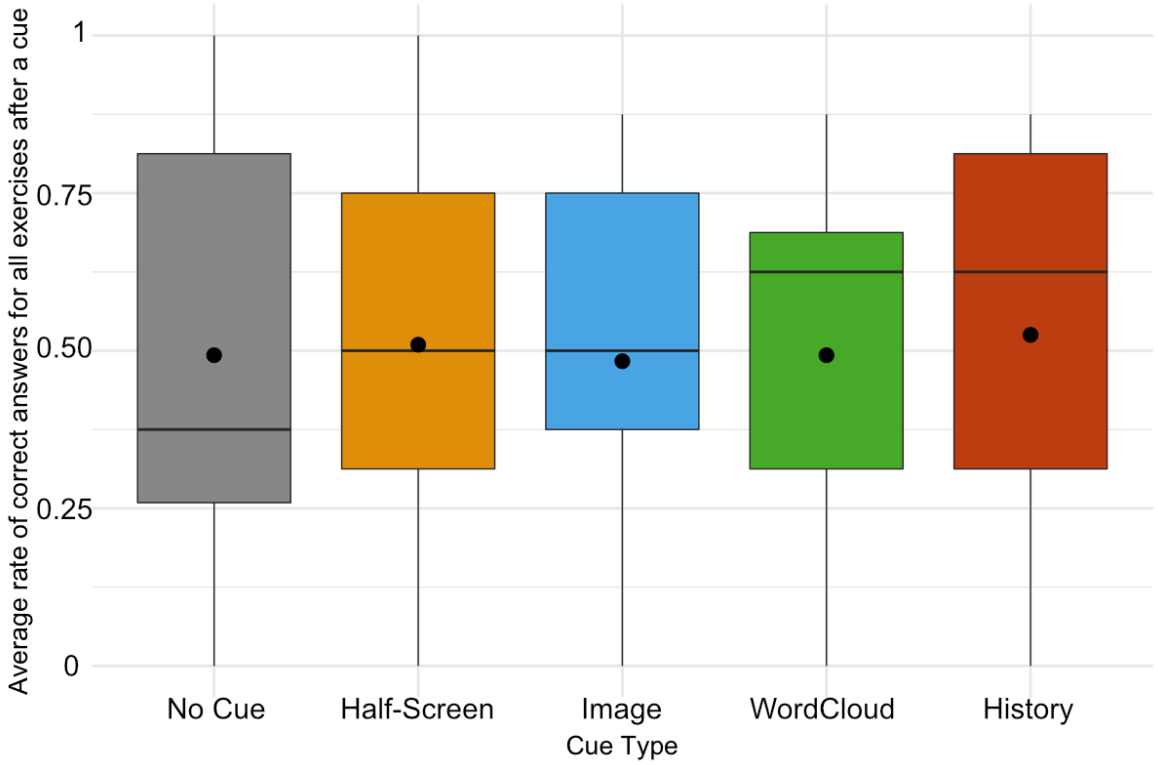
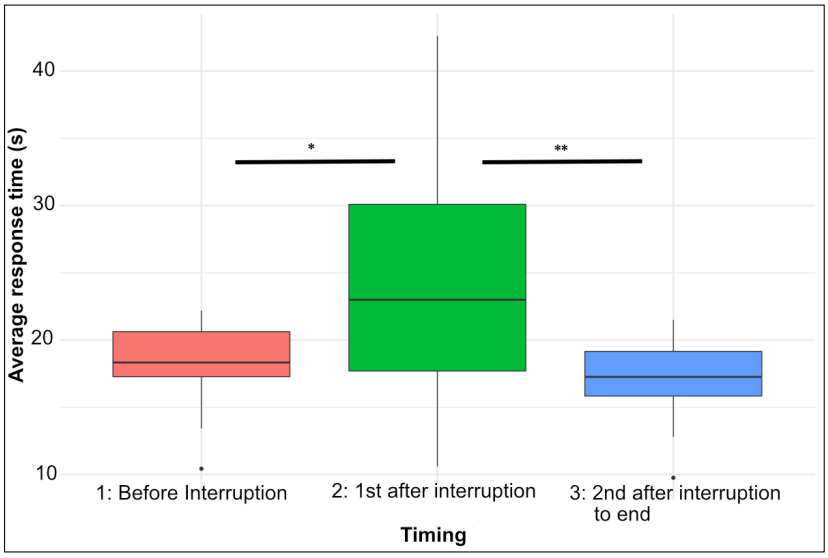


Figure 8.4: The average correctness of all answers after an interruption in regards to the four cue designs compared to the no cue condition (gray left). The horizontal black line indicates the Median, the black dot the Mean.

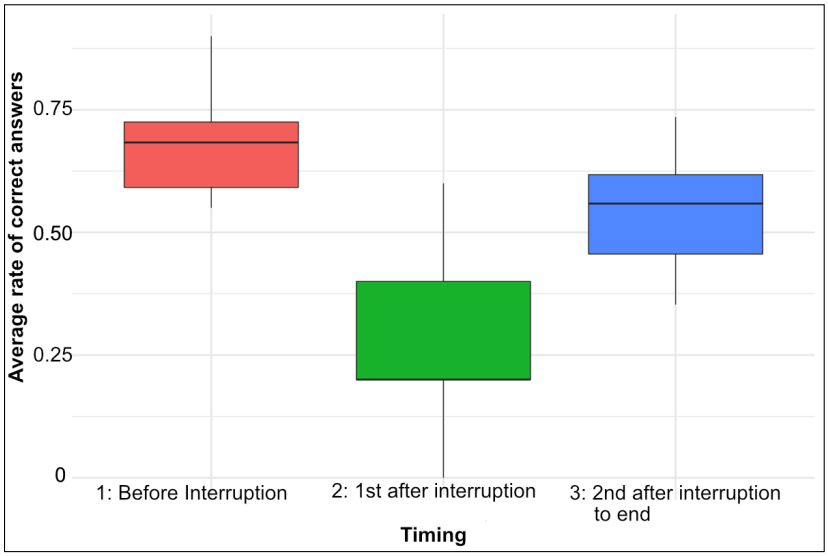
for all exercises following the interruption was lowest in the no cue condition ($M = 16.1$ seconds, $SD = 7.8$), followed by the *Half-screen Cue*, *WordCloud Cue*, *Image Cue*, and *History Cue* (cf. Table 8.1).

Error Rate In line with the results from the task completion time analysis, we found that the interruptions significantly affected the correctness of the first task after the interruption occurred (repeated-measures ANOVA, $F = 69.3$, $p < .001$; see Figure 8.5b). Post-hoc comparisons with Bonferroni correction revealed significantly higher error rates than for the first task solved after an interruption compared to all tasks before ($t = 11.74$, $p < .001$) and all following tasks ($t = -6.53$, $p < .001$). Furthermore, we find higher error rates in tasks solved after an interruption (excluding the first task solved immediately after the interruption) compared to tasks solved before ($t = 5.18$, $p < .001$). I.e., participants made more errors in the exercises of the second lesson part, after they were interrupted.

Regarding the influence of our different cue types (either as four individual cases or as a cue/no-cue comparison), a Chi-square (χ^2) analysis did not show a significant effect of cue type on correctness (yes|no) after the interruption ($p < .05$; for descriptive



(a) Answer Duration (Response Time)



(b) Correctness Rate

Figure 8.5: Effects of interruptions on (a) answer duration and (b) correctness rate.

visualization see Figure 8.4). Similarly, we also did not find an effect of the cue types on the correctness of the first task after an interruption.

8.2.5 Discussion

Limitations

While this laboratory experiment aimed to assess the basic effects of interruptions on mobile learning performance, participants stated in the interviews that the situation felt very artificial. According to their perception, experiencing only one interruption type in an otherwise controlled setting did not reflect everyday situations. Therefore, they did not expect to be strongly influenced by the interruption and felt as if they could still recall the latest learning session. Future work is needed to assess the effect of task resumption cues with diverse interruptions in the wild to reflect real-world usage.

The data furthermore shows a difference in the difficulty of the presented learning lessons. Lesson 1 showed higher error rates and longer duration, which the interruptions or cues cannot explain as their presentation was counterbalanced to enforce randomization. We suspect that this perceived difficulty is due to Polish being a new language for our participants. We recommend increasing the number of learning tasks in future evaluations to overcome differences in task difficulty.

Moreover, we observed a limitation in the implementation of the *Half-screen Cue*. In this cue design, the interrupting multiplication task UI overlays around 60% of the screen. While the lower third of the screen left the learning app visible, the remaining space was overlaid by the keyboard once participants started entering the results of the multiplication tasks. Thus, the *Half-screen Cue* was only visible for a certain amount of time and not during the whole interruption. Together with the fact that several participants did not notice the cue, the results in regards to the *Half-screen Cue* should be viewed with a grain of salt.

Effects of Interruptions In evaluating participants' performance across the different tasks, we observed that the interruptions affected their performance. Especially for the first task after an interruption, we recorded longer task completion times and higher error rates independent from any task resumption cue. For the remaining tasks after the interruptions, the performance improved again. Yet, participants expressed their doubt about the effect of the interruptions. In the interviews, they stated that the interruptions felt artificial and had no problem remembering the content of the learning session before the interruption. From the results of our study, we assume that while the interruptions were short and contained, they still (implicitly) affected our users' focus. We assume that real-world interruptions that take the users out of the context of the learning activity (mentally and physically) and potentially last significantly longer than our experimental interruptions will have a greater negative impact on the learners' performance. For future evaluations, we suggest including a greater variety of interruptions or choose a field study setting with a natural environment altogether.

Objective vs. Subjective Helpfulness of Cues The quantitative analysis of our study results showed that the effect of the TRCs on task completion time and error rate was

limited. Compared to the no-cue condition, the presentation of any of the four cue designs led to longer response times and no improvement of correctness in the exercises after the interruption. Nonetheless, all participants stated that the implementation of TRCs in mobile learning apps would be a helpful feature. While the *WordCloud Cue* was considered helpful for lessons introducing many new words to the vocabulary base, the participants could imagine the *History Cue* to be particularly helpful for grammar lessons (i.e., summarizing the prior lessons' rules). We hypothesize, aligned with the expectations expressed in Chapter 7, that the perceived helpfulness of the TRCs depends on the alignment of task complexity and resumption cue complexity. It is also possible that the effect of resumption cues is stronger for long-term retention than for immediate recall.

Participants Favor Explicit over Implicit Cues Looking at the subjective helpfulness ratings, we observe that explicit cues appear to be more helpful in supporting participants resuming the learning tasks. Especially the *History Cue* was acknowledged as being helpful, while the *Image Cue* received the lowest helpfulness rating of all four cue designs. The viewing duration of the different cues indicates that our participants examined the content of the *WordCloud Cue* (on average eleven but up to 74 seconds) and *History Cue* (on average 19.8 seconds) in detail. Participants noted that the *WordCloud Cue* could be more helpful if the words were arranged according to a certain logic. While the visualization in this evaluation contained words of the prior lessons, colors and sizes could be adapted according to certain criteria such as correctness in participants' answers, frequency of occurrence in the lesson, or importance for the language in general. The *Image Cue* was viewed only briefly, and participants stated in the interviews that they did not perceive this type as very helpful. In the overall subjective helpfulness ratings, the *WordCloud Cue* and especially the *History Cue* ranked higher when compared to the *Half-screen Cue* and *Image Cue*, suggesting participants' preference for explicit cues over implicit. Nonetheless, because many participants did not perceive the implicit cues as actual memory cues, we suggest that future work takes a deeper look into the effectiveness of implicit cues with revised designs.

8.2.6 Summary

In this user study, we evaluated four designs of memory cues to support users in resuming learning tasks on mobile devices after interruptions. We implemented a language learning application that contained two implicit and two explicit cue designs that focus on retrospective rehearsal - presenting the user with information of what they were doing before an interruption to restore the task context. Our evaluation revealed that while the presentation of the cues had no significant effect on objective performance measures (task completion time and error rate), the users still perceived the cues as helpful and would appreciate them in a mobile learning app. This evaluation provides first insights into the implementation of memory cues as a feature in mobile

learning applications and we believe that a follow-up evaluation needs to investigate the cues in an in-the-wild setting further. In particular, the greater variety of interruptions that happen during learning activities in everyday settings (cf. Chapter 3 and 4) was not sufficiently represented in this laboratory evaluation and can potentially impact the necessity for resumption cues.

8.3 Task Resumption Cues in the Wild

In Chapter 3 we observed a great variety in usage situations of mobile learning applications. These usage situations are characterized by a similarly diverse set of external and internal stimuli that potentially distract or interrupt the user. When learning at home, users might be interrupted by kids playing loudly in the background. In contrast, when learning on a commute on public transportation, the learners might have to use the mobile learning app standing in a crowded group while listening to the announcement to avoid missing their stop. Chapter 4 outlines that interruptions can be characterized among others according to their source, duration, anticipation, urgency, modality, and complexity.

The study presented in the first half of this chapter evaluated TRCs in a scenario that only covered one specific interruption format: based in a lab environment, with device-internal interruptions of similar length, semi-anticipated, medium complexity, and similar modality. While the first study revealed critical first insights into users' perception of the TRCs general helpfulness for mobile learning applications, we could not prove objective influences on the learning performance. A main limitation of the study was that users reported not feeling very interrupted. Therefore, in this second half of the chapter, we present a follow-up evaluation of TRCs in a field study. We revised the cues according to the first evaluation's feedback and implemented them in a mobile learning application for users to learn in the wild in their everyday lives. This study procedure creates a more natural user behavior, including diverse usage situations and interruptions. We build on the findings of the prior section and investigate the effect of our TRC designs on the learners' error rate, task completion time, and subjective perception of usability and helpfulness.

8.3.1 Implementation

We decided to implement an Android application for the in-the-wild user study to reach a bigger user group. In this follow-up project, we iterated on the design of the TRCs. We embedded them in a similar mobile language learning android application we developed called *CzechWizard*, which teaches Czech at the beginners' level with a focus on vocabulary and short sentences (published as a closed test in the Google Play Store). We decided to use Czech for the same reasons as Polish before – it relies on the

same base alphabet as German without being too similar, making it neither difficult nor easy to learn. Further, Czech is not commonly taught in schools, thus, making it possible to find participants with no knowledge of the language. Using a second language, we aim to diversify our results, simultaneously iterating on the lessons to tackle the problem from the lab evaluation, where participants reported tasks as too difficult.

The app provided a content overview with lessons on topics such as gender, the use of simple verbs such as "be", or plural use. Each lesson again contained multiple "blocks" including content on the topic of the lesson. We derived several tasks, such as (1) relating vocabulary to images, (2) translating words using multiple-choice answer formats (with and without providing a visual representation of the vocabulary items), and (3) sentence building by selecting the correct words from a list of options. The blocks in each lesson were building on each other and increased in difficulty. Further, words learned in the early sessions were later used to build sentences. The app validated the correctness of the users' answers and presented visual feedback using the commonly applied color scheme of green (correct) and red (incorrect). In case of incorrect answers in multiple-choice answer formats, the correct answer from the set was additionally highlighted in green.

For the later analysis of the data, we logged all user interactions with the application. In particular, we focused on the task completion time and the error rate. As outlined in Chapter 4, the diversity in interruptions users experiences in their daily life - together with current technical limitations in recognizing external interruptions - make automated interruption detection challenging. The majority of interruptions detected in Chapter 4 originated in the device (i.e., notifications, predominantly messaging apps) or in the users themselves (i.e., switching to a different task without external stimulus). Therefore, in this application, we decided to focus on Android's *onStop()* and *onDestroy()* events, caused when the app is moved to the background and no longer visible, the screen is turned off, or the application is terminated.

8.3.2 Revised Cue Designs

Similar to the cue set of the first evaluation in this chapter, the revised cue designs include two implicit and two explicit TRCs (cf. Figure 8.6). Since we cannot guarantee that the *Half-screen Cue* would work reliably for all potentially interrupting applications on devices with different operating systems, we do not include this design in our further evaluation. Instead, we follow Chapter 7's Design Implication 5 and explore a new modality for an implicit cue, a tactile vibration pattern. Further, we extended the *Image Cue* from a single image participants did not consider very helpful to a more subtle but also more pervasive color and icon scheme. The *WordCloud Cue* remains mainly the same with minor adaptations in the display of the words. Lastly, we changed the *History Cue* to not contain an overview of prior lessons but to include

interactive tasks on the content of the prior lessons. In the following, we will outline the cue designs and the implemented changes compared to the earlier cue versions in more detail.

Vibro-tactile Cue Figure 8.6a depicts the idea behind the *Vibro-tactile Cue*, an implicit cue aiming to create a subtle association between the lesson and a vibration pattern. Whenever a user enters a lesson, the device issues a vibration. This type of subtle tactile feedback has been recommended as unobtrusive alternative to other modalities as it requires less attention and cognitive capacities. The vibration mechanism aims at creating an association between the stimulus and learning activity and directs the users' focal attention (cf. [151, 152, 199, 262]).

Color and Icon Cue The second implicit cue design is the *Color and Icon Cue* (see Figure 8.6b). This cue shows an icon and color specifically associated with the lesson's theme the user is currently working on. The design extends the *Image Cue* presented in the first study. Like this cue, an image is chosen from the previous lesson and supplemented by a color. Groups of similar lessons received colors of a related color palette. This cue is inspired by the work of Yatid and Takatsuka [356] who used colors for categorical associations to the context.

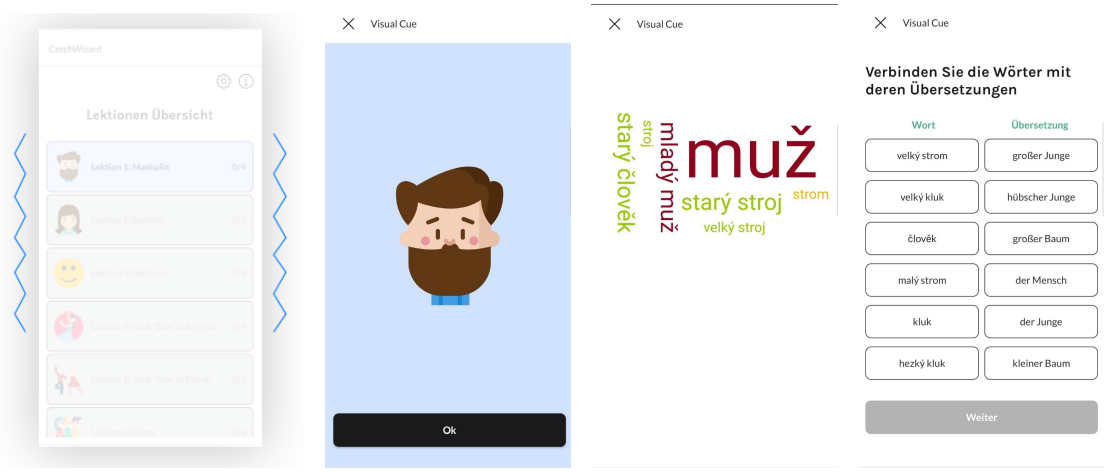
WordCloud Cue The *WordCloud Cue* remains similar to the design presented in the first part of this chapter (see Figure Figure 8.6c). The explicit TRC generates an overview of words learned in the prior lessons. In contrast to the earlier design, the size of the words in the word cloud now indicate the learning process. Words displayed in a larger font have been answered correctly more often than words in smaller font. Prior work of Ardissono et al. [15] showed that such type of a visualization model could help users quickly find the relevant information when accessing a large amount of data. Further, for vocabulary learning, the cloud represents a summary of learning content.

Interactive Test Cue While the *History Cue* received positive feedback in the first study, we decided to further improve the cue by adding interactivity. Since the *WordCloud Cue* already presents a passive visual summary of prior content, this second explicit cue, coined *Interactive Test Cue*, asks users to answer short questions (see Figure 8.6d). In particular, a screen presents six L1 words and their L2 translations, asking the user to match them by selection. Similar to the *WordCloud Cue*, the recognition tasks are generated from the vocabulary of the last lesson the user learned.

8.3.3 Methodology

Study Design and Apparatus

To test the four resumption cue designs (independent variable), we embedded them in a mobile learning application compatible with Android 8 or higher. The cues were



(a) Vibro-tactile Cue (b) Color Cue (c) WordCloud Cue (d) Interaction Cue

Figure 8.6: The four revised designs for task resumption cues for mobile language learning, (a) *Vibro-tactile Cue*, (b) *Color and Icon Cue*, (c) *WordCloud Cue*, and (d) *Interactive Test Cue*.

presented in a counterbalanced order independent from the current task to create a within-subject study design. Further, we included a no-cue condition to work as a baseline. After the interruption and the cue display, we measured the error rate and answer duration (dependent variables) for the first task the user performs. We postulate the following hypothesis:

H_1a Showing a cue after an interruption leads to lower task completion times and error rates as compared to showing no cue in the first five tasks after the interruption occurred.

Procedure

After their recruitment, we informed the participants via email about the study procedure. They received an information sheet outlining LMU's data protection policies and provided informed consent for the study. A first online questionnaire assessed general demographic information and the participants' language proficiency, in particular, regarding Slavic languages. At the end of the survey, participants were guided through the installation process of the application available as a closed test to download through the Google Play Store. We encouraged the participants to use the app as naturally as possible in their everyday life while logging their usage interactions with it, including but not limited to learning session and task duration, and correctness of the tasks. After using the application for the study duration, a second questionnaire asked for the usability of the application in general and feedback on the TRC designs. Additionally to the survey questions we designed specifically for this evaluation, we deployed a standardized questionnaire to assess usability and user experience of the

application, the System Usability Scale (SUS) introduced by Brooke in 1996 [42]. The questionnaire includes ten items using a five-point Likert scale. The aggregated scores range from 0 to 100, with values over 70 attesting an acceptable level of usability [25]. Specifically, we aim to assess the overall learning app usability to rule out the potential effects of usability problems on the users' TRC perception.

Sample

We recruited participants through our university's mailing list as well as other communication and social media channels. In total, our sample size was 17 (twelve identifying as female, four as male, and one as non-binary) took part. Their age ranged from 19-29 ($M = 24$, $SD = 3.17$), and they reported no experience with Slavic languages. Ten participants held an A-Level diploma or equivalent, four a bachelor's degree, and three a master's degree or higher. Fourteen were currently enrolled in a study program; three were employed in full-time jobs[5]. Every participant received a 20 Euro voucher for an online shop or study credit points as compensation for their participation.

8.3.4 Results

We will first outline the subjective assessment of usability and perceived helpfulness before taking an in-depth look into the quantitative measures, task completion time and error rate.

Questionnaires Before reporting the experiences with the TRCs, participants completed an adapted version of the SUS questionnaire [42]. By using the SUS, we aimed to gather data about usability issues with our application in general that might have biased the participants' rating of the TRC feature. The SUS scores ranged from 57.5 to 97.5, with a mean of 82.21 ($SD = 12.21$). Since 70 marks the threshold for acceptable usability, our application's rating of 82 can be considered excellent (cf. [25]), and we do not expect any influences of the usability on participants' cue assessment.

To confirm the appearance of the cue, the questionnaire first asked participants if they had actually noticed the cue by showing pictures of them. In total, five people stated to have not noticed the vibration pattern or at least did not notice them as a TRC. Furthermore, each one person did not notice the *Color and Icon Cue* and *WordCloud Cue*, and two people failed to notice the *Interactive Test Cue*, or at least to identify it as a task resumption support feature.

Overall, participants ranked the helpfulness of the *Color and Icon Cue* and *Interactive Test Cue* above average ($M_{\text{color}} = 4.35$, $SD_{\text{color}} = 1.94$; $M_{\text{test}} = 5.59$, $SD_{\text{test}} = 1.29$, 7-point Likert scale from 1=*not at all* to 7=*very helpful*), while the *Vibro-tactile Cue* and

[5] We conducted this study during the COVID-19 pandemic. Thus, the results may be limited in generalizability as users' routines, mobility, and smartphone usage can deviate due to lockdown and work-from-home phases. We will discuss this issue in the section on limitations.

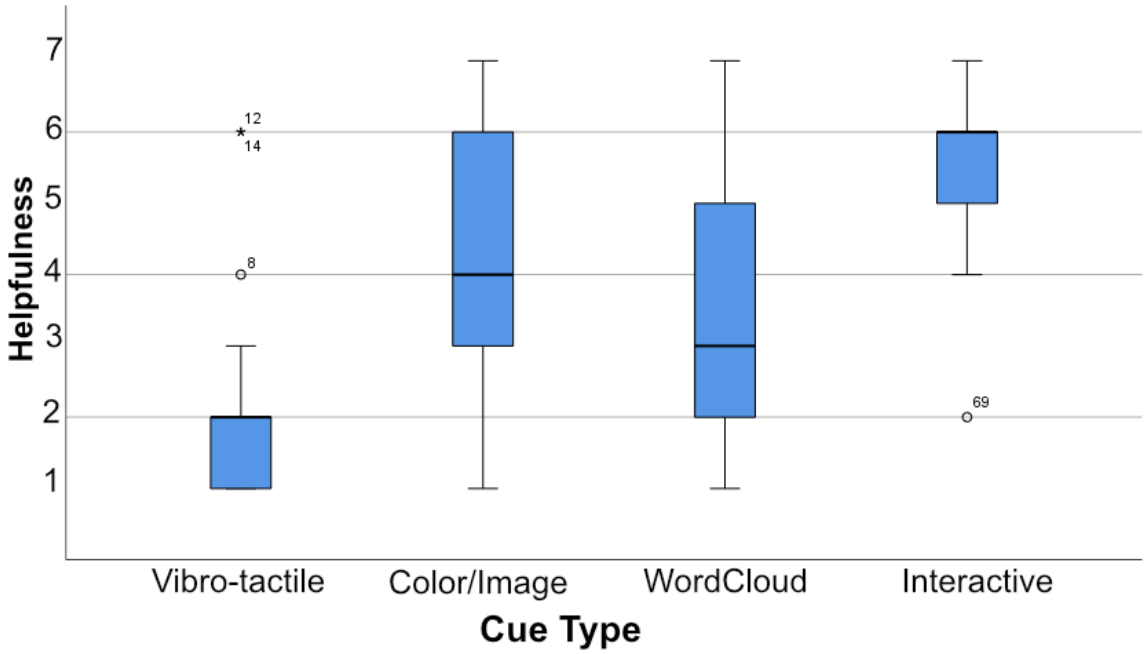


Figure 8.7: Perceived helpfulness of the different cue types on a scale from 1=*not at all* to 7=*very helpful*.

WordCloud Cue received lower ratings ($M_{\text{vibro}} = 2.18$, $SD_{\text{vibro}} = 1.62$; $M_{\text{cloud}} = 3.35$, $SD_{\text{cloud}} = 2.03$, cf. Figure 8.7).

We further asked participants if they would have preferred the presentation of the cues at a different point in time as well as to state feedback for the different cues individually. Participants often did not recognize the vibration as a TRC, thus considering the *Vibro-tactile Cue* useless. Four participants reported they could imagine it being an indicator for correct or incorrect answers. One person mentioned it could help to use it in the middle of a session to keep the user attentive. Two participants stated here that they would rather prefer not to use the cue at all. As suggestions for improvement, they noted to increase the frequency and improve the timing. In contrast, P13 praised the *Vibro-tactile Cue* for being least intrusive, stating it "[...] *made the lesson stand out in a subtle way.*"

None of the participants had any suggestions for improvement for the timing of the *Color and Icon Cue*. It was perceived as a helpful reminder of the last learned lesson (P4, P6, P7, P9, P11, P13, P17). Participants suggested improving the choice of icons (P1) or add a keyword (P3), especially if the last lesson is further in the past and the memory fades.

For the *WordCloud Cue* the opinions were very diverse. Two participants suggested removing it altogether; one person mentioned they wanted to see such cues more frequently throughout the app, while one user perceived the timing of this cue as "arbi-

trary". Three participants considered the cloud as a nice overview (P3, P10, P11) and helpful to track the individual progress (P12). Referring to words that were shown less frequently, P5 mentioned that "[...] *their smaller size caught my attention and I think it helped me to memorize them better.*" In contrast, P13 perceived this cue as "*demotivating*", as they reflect on low performance.

The *Interactive Test Cue* received the most positive ratings and feedback of all cues. Participants described the cue in a very positive way, with all but two considering it useful and helpful (all but P7 & P17 who did not perceive the cue). While two participants considered it a "[...] *a little long*" (P3), P8 reported that it was badly timed when the interruption occurred at the beginning of a learning session. As the cue asked for translations of words of the specific lesson, it would then include words that are unknown to the user at this point. Thus, users suggested using this cue rather as a quick repetition after a lesson, even without an interruption.

As an overall conclusion, the participants perceived the TRCs in general as helpful to recover from interruptions during mobile learning ($M = 4.06$, $SD = 1.21$; 7-point Likert scale from 1=*not at all* to 7=*very helpful*) and would like to see such cues implemented in mobile learning apps (15 participants in favor, two oppose). Two participants further stated that such TRCs as evaluated in this app could also be beneficial for reading longer digital texts (e.g., online articles or e-books) (P1, P7).

Sessions and Interruptions The usage of the *CzechWizard* varied among our participants. For the number of tasks answered, we observed a range between 169 and 821 tasks during this study ($M = 519.44$, $SD = 236.85$). During the learning sessions, we recorded on average 13.19 interruptions per participant ($SD = 6.23$). The wide range of recorded interruptions, between three and 29, confirms the diversity of usage behavior we addressed in Chapter 3. The experience sampling questionnaire meant to assess the learning context was majorly dismissed by the participants and only answered in 163 cases. Of these 163, 87 represent learning at home, 36 on public transportation, 22 on the go, 15 at work, and three at the university.

Cue Viewing and Interaction We cleaned our data set of aborted cues, learning sessions that did not exceed 10 seconds, and cue viewing durations that exceeded five minutes. Our final data set recorded 8276 solved learning tasks ($N = 16$) and 209 detected interruptions. The four TRC designs were counter-balanced, due to the removal of certain instances we included 44 *Vibro-tactile Cue*, 46 *Color and Icon Cue*, 42 *WordCloud Cue*, and 36 *Interactive Test Cue*. In 41 cases, no cue was shown as a control condition.

Figure 8.8 depicts the viewing duration of the three cue types. As the duration of the *Vibro-tactile Cue* was fixed in the app, it is not included here. The interaction time with the *Interactive Test Cue* that asked users to match six words and their translations exceeded those of the other cues. In particular, the interaction time with the *Interactive Test Cue* in seconds ($M = 44.88$, $SD = 19.37$) was on average five times

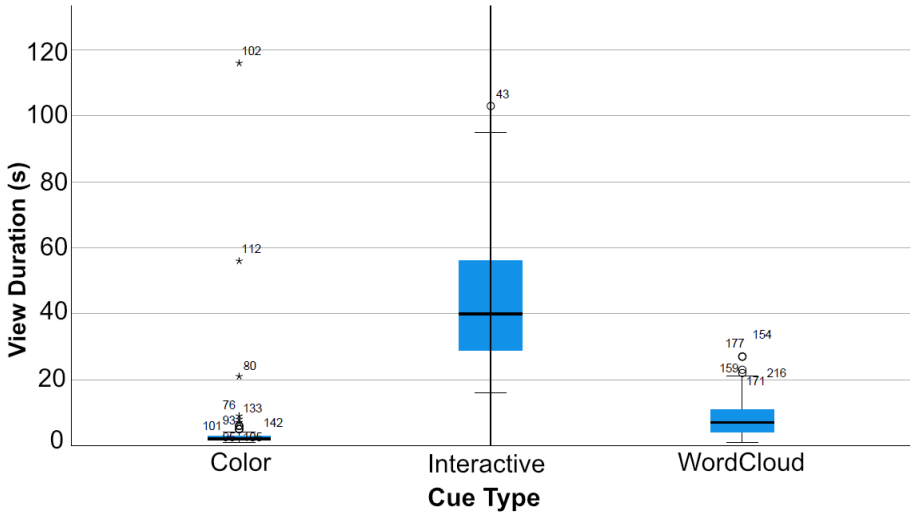


Figure 8.8: The viewing duration in seconds of the three cue types *Color and Icon Cue*, *Interactive Test Cue*, and *WordCloud Cue*.

as high as the viewing duration of the *WordCloud Cue* ($M = 8.48$, $SD = 6.51$) and even ten times as high as the viewing duration of the *Color and Icon Cue* ($M = 4.71$, $SD = 13.36$). Looking at the *Interactive Test Cue* in detail, users aborted or quit the tasks in 32 out of 90 cases. They did so either immediately or after answering a subset of the questions. Out of the completed set of 58 cues, 35 were answered entirely correctly. Overall, the median error rate during the cue tasks remained zero ($max = 6$). Note that all tasks had to be solved correctly in order to continue with the lesson.

Task Completion Time In the lab-based user study of the first part of this chapter, we compared the task completion time before and after an interruption, i.e., before and after a resumption cue was presented. In this field study, the majority of interruptions led to the termination of the current learning session. As a result, the TRC was often shown immediately at the beginning of the next learning session. Therefore, we will not compare the tasks before and after the interruption but only take the first tasks after each resumption cue is shown into the consideration for the analysis ($N = 1004$). Figure 8.9 visualizes the descriptive statistics of the task completion times. The highest average task completion time can be observed after the presentation of the *WordCloud Cue* ($M = 6.91\,\text{s}$, $SD = 3.97\,\text{s}$), the lowest in the no cue control condition ($M = 5.59\,\text{s}$, $SD = 1.99\,\text{s}$).

Due to the removal of outlier values and the cleaning of the data set, we performed a computational imputation for missing values. The missing values were replaced by averages across participants. A repeated-measures ANOVA across participants showed no significant difference ($p > .05$) among the four different cue types and the no-cue condition (averaged per participant). In other words, users needed almost the same

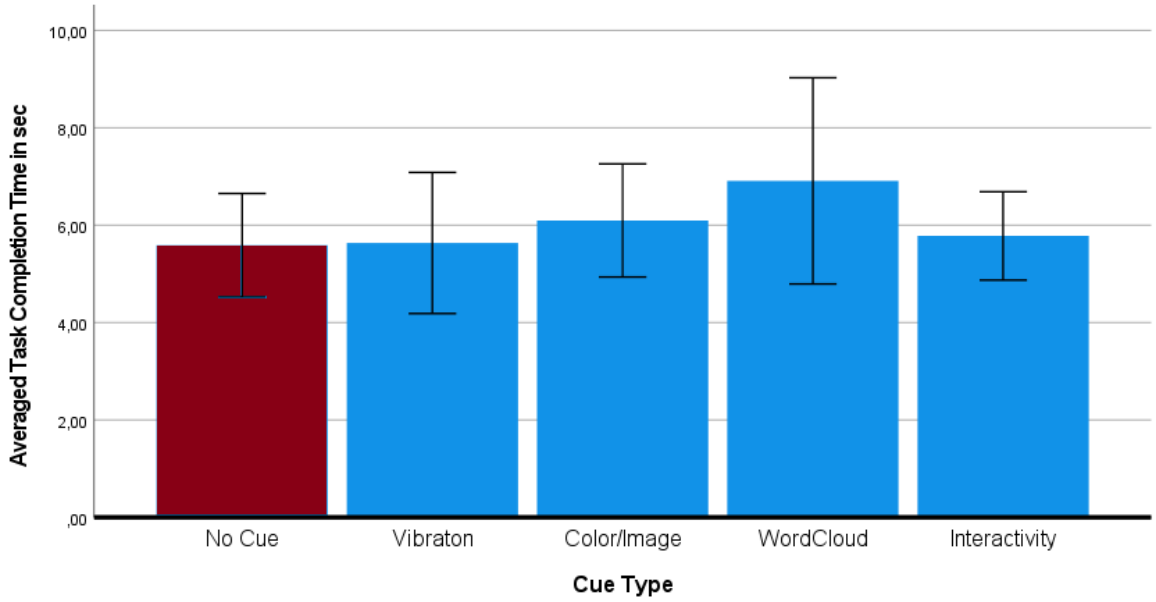


Figure 8.9: The task completion time for the first five tasks after an interruption in seconds (excl. cue view duration). The completion times are averaged per participant and visualized to compare the four cue designs with the no cue condition.

amount of time to solve a learning task after an interruption, no matter if and which TRC was shown.

Error Rate The learning content in the application was at a beginners' level, and the number of errors made by our participants turned out to be fairly low. Out of 8276 recorded learning tasks, only 110 were incorrectly answered, resulting in an error rate of less than 1%. When we limit the responses to the first five tasks after a cue was presented ($N = 1004$), only 36 were incorrect. We consider this sample insufficient for statistical comparison among five test conditions and will refrain from drawing conclusions on the effectiveness of TRCs designs on correctness. We will discuss the feasibility of the applied metrics for measuring the effectiveness of TRCs in the limitation section.

8.3.5 Limitations

The evaluation presented above revealed several limitations in our application and also our study design.

Firstly, the motivation to engage in the learning sessions varied greatly among our participants. We encouraged all study participants to interact with the app frequently but also in the same way they would with any other learning app. We observed similar differences as already reported in Chapter 3 and 4. Some participants learned multiple

times per day, while others only had very few learning sessions. Due to this difference, the number of interruptions, sessions, and breaks varied (min 3, max 29). Without a sufficient number of interruptions, the generalizability of our results concerning the helpfulness of the TRCs is limited.

Secondly, due to the lack of complexity in the learning content and the users' good overall performance, the explanatory power of our metrics task completion time and error rate about the effectiveness of the TRCs is limited. For future work, we recommend evaluating memory cues in a setting with more complex content.

In addition, we observed several methodological constraints during our user study. One critical issue is to define the difference between an interruption and a new learning session. In the laboratory settings, interruptions were fixed in terms of severity and duration. In users' everyday environment, the only indicator for interruption severity we can gather is the duration between two learning sessions. For the current analysis, we considered every break between learning sessions, whether five minutes or five days, an interruption. However, the memory decay after five days will be significantly worse than after five minutes [294]. Thus, adapting the presentation of TRCs to the severity of the interruption might be good to consider for future work.

8.3.6 Discussion

Divided Opinion on Cue Designs

While the majority of participants agreed on the helpfulness of the *Interactive Test Cue*, their opinions were divided on the other cue designs. In particular, both the *WordCloud Cue* and *Color and Icon Cue* received very positive and very poor feedback. Giving users the option to adapt the granularity or content of the cues could help increase their acceptance. Further, deploying the different cues in regard to the users' demands could improve their acceptance. As stated in the limitation section, the severity of the interruptions and the complexity of the content influences the need for memory support. In short, easy or frequent learning sessions users could be supported by implicit tactile or visual cues. At the same time, they would potentially benefit more from explicit and complex cues in long, difficult, or irregular learning sessions. This hypothesis needs to be further investigated in future work to determine these dependencies.

Problems and Opportunities for Using Tactile Memory Cues

Since the *Vibro-tactile Cue* was very subtle and implicit, it was often not perceived by our participants. Some expected that the vibration was a feedback mechanism that implies correct or incorrect answers. While this was not the case, this idea presents the intriguing opportunity to cue certain content specifically. For example, a vibration cue during the corrective feedback presentation of a word that has been incorrectly

answered multiple times in a row could grab the users' attention. By replaying the *Vibro-tactile Cue* the next time the same question is posed, we can guide the users' attention and increase their caution and focus when answering this specific task. Thus, we can ensure deeper processing and potentially foster more elaborate rehearsal, leading to better encoding into their LTM.

Task Resumption beyond Micro-Learning

According to our quantitative metrics, we could not observe any effect of the TRCs on the learning performance. The course was generally designed to include simple content aimed at teaching beginners' level Czech to avoid differences among participants due to existing language proficiency. The very low overall error rate of less than 1% indicated that users had no difficulties remembering the app's vocabulary and answering the tasks. We suspect that while most participants appreciated the TRCs, in particular, the *Color and Icon Cue* and *Interactive Test Cue*, the helpfulness of the cues would be greater for content of greater complexity. The beginners' level Czech course we designed for our experiment had only a few content units building upon each other. Recalling prior content was, therefore, not necessarily required for progressing in the application.

8.4 Chapter Summary

This chapter presents two user studies evaluating task resumption cues in mobile learning applications. In a first study, we implemented an iOS application that included four first designs for potential memory cues (*Half-screen Cue, Image Cue, History Cue, and WordCloud Cue*). We could not observe a significant effect of task resumption cues on learning performance in the laboratory evaluation. However, participants' subjective overall impression favored including memory cues as a feature for task resumption support in mobile learning apps. The follow-up study took the implementation out of the lab and into the wild. We substituted the *Half-screen Cue* with an *Interactive Test Cue*, which turned out to become the favorite design of our participants. The study again showed no objective differences in learning performance if a cue is shown but clear subjective preferences.

Concerning our research question **RQ4** – which asked if we can use memory cues to mitigate the adverse effects of interruptions in everyday mobile language learning – our answer remains inconclusive. Our findings were able to show that learners react positively to the provision of task resumption cues and consider them helpful for learning. Yet, the quantifiable effects of the TRCs when implemented in common learning applications can not be conclusively confirmed. However, based on our results, we still see great potential for such memory supplements to change how we learn on mobile devices. From the need to break the content down into individual micro-learning units, TRCs, if appropriately designed, can create a seamless connection between individual

learning units. Instead of multiple self-contained micro-learning sessions, learners could engage in more complex topics that require longer time and deeper engagement. We will investigate users' experience with task resumption cues regarding other knowledge formats in future work.

9

Outlook - Memory Cues Beyond Language Learning

To design effective task resumption cues (TRCs) for a mobile learning application, the previous Chapters 7 and 8 emphasized the need to adapt the complexity of the cue to the task at hand. Since learning a language at beginners' level only covers the lower end of the content complexity spectrum, this chapter aims to extend the evaluations of the prior chapters using content of different complexity levels.

We present the development of an Android application teaching JAVA programming ("*LearnJava*"), which includes low-complexity recall tasks (i.e., naming data types) but also high-complexity creation tasks (i.e., writing JAVA code). We integrated the tasks in a matrix representing three different knowledge types and four complexity levels aligned with Bloom's taxonomy [35]. We implemented four revised TRC designs based on the results of the prior chapters, namely (1) a current lesson reminder, (2) a word-cloud, (3) a lesson history overview, and (4) an interactive question cue. Each of these cues was designed to target one of the combinations of knowledge types and complexity levels.

By investigating the helpfulness of memory cues for content of higher complexity, we take a first step toward the generalization of the TRC approach to other learning domains. We will discuss the implications of our results for future application of cues in other (learning) applications. Particularly, we present an outlook on how the use of TRCs could change the way we design micro-learning applications for learning on mobile devices.

> This section is supported by the Bachelor thesis of Miriam Halsner, see detailed collaboration statement at the beginning of this thesis.

9.1 Related Work

For the differentiation of task complexity levels in learning, this section will discuss models for characterizing knowledge types as well as cognitive processes. We already introduced the basic processes of human memory and the different knowledge and memory types in Chapter 2.1 (see specifically Figure 2.3). In this section, we will now map different learning tasks as they could occur in a learning application to these knowledge types.

9.1.1 Knowledge Types and Dimensions

In terms of long-term memory knowledge types, Tulving and Schacter [332] distinguishes among others **factual knowledge** (i.e., basic, isolated information), **conceptual knowledge** (i.e., organized knowledge, relationship between elements and processes), **procedural knowledge** (i.e., subject-specific skills and methods), and **meta-cognitive knowledge** (i.e., cognitive processes, awareness of own cognition). In this work, we will use these knowledge types as one indicator for defining content complexity levels. Specifically, the term "complex" is defined as something made out of complicated or interrelated parts, sometimes not fully disclosed[1]. We argue that the complexity of knowledge represented in these four categories increases from isolated factual information to complex conceptual and procedural knowledge. Since meta-cognitive knowledge is a type of knowledge concerning the user and less the content, this type will be excluded.

9.1.2 Cognitive Processing Dimensions

For the categorization of the learning process, Bloom et al. [35] described six phases, from simple to complex information processing, namely knowledge, comprehension, application, analysis, synthesis, and evaluation (see Figure 9.1). He outlined the phases in the commonly known pyramid form, thus, stating the increased level of processing and the frequency distribution in which they occur in learning. Those phases were later revised and renamed, resulting in the following six categories: remember (i.e., recalling a fact), understand, apply, analyze, evaluate, and create (i.e., generate or produce new work). We will use these processing dimensions as a second indicator for task complexity. As it is difficult to draw hard lines between the six individual categories, we narrow them down to four categories according to the level of cognitive outcome. We differentiate between two categories of lower-level cognitive outcome, namely **remember** (recognition of correct answers), and **understand** (recall of correct answers), and each one category of medium and high cognitive outcome, namely **analyze** and **create** (see Table 9.1).

9.2 Concept and Implementation

We designed and implemented the *LearnJava* Android application to teach JAVA on a beginner's level. The app consisted of a sign-in screen, a main view, and eight consecutive learning content sections. The sign-in required users to come up with a (non-personal) nickname used as identification to store with interaction data and

[1] For definition of the term "complex" see `https://www.merriam-webster.com/dictionary/complex`, last accessed January 3, 2022

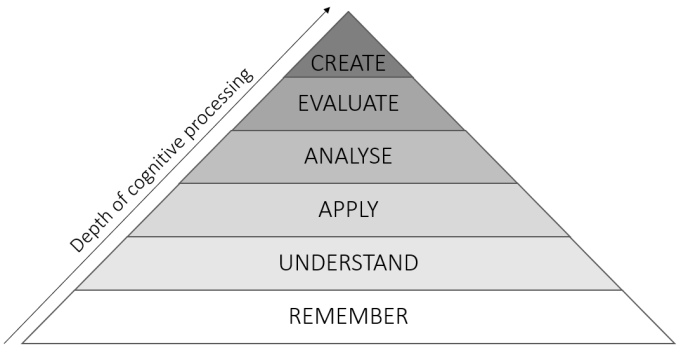


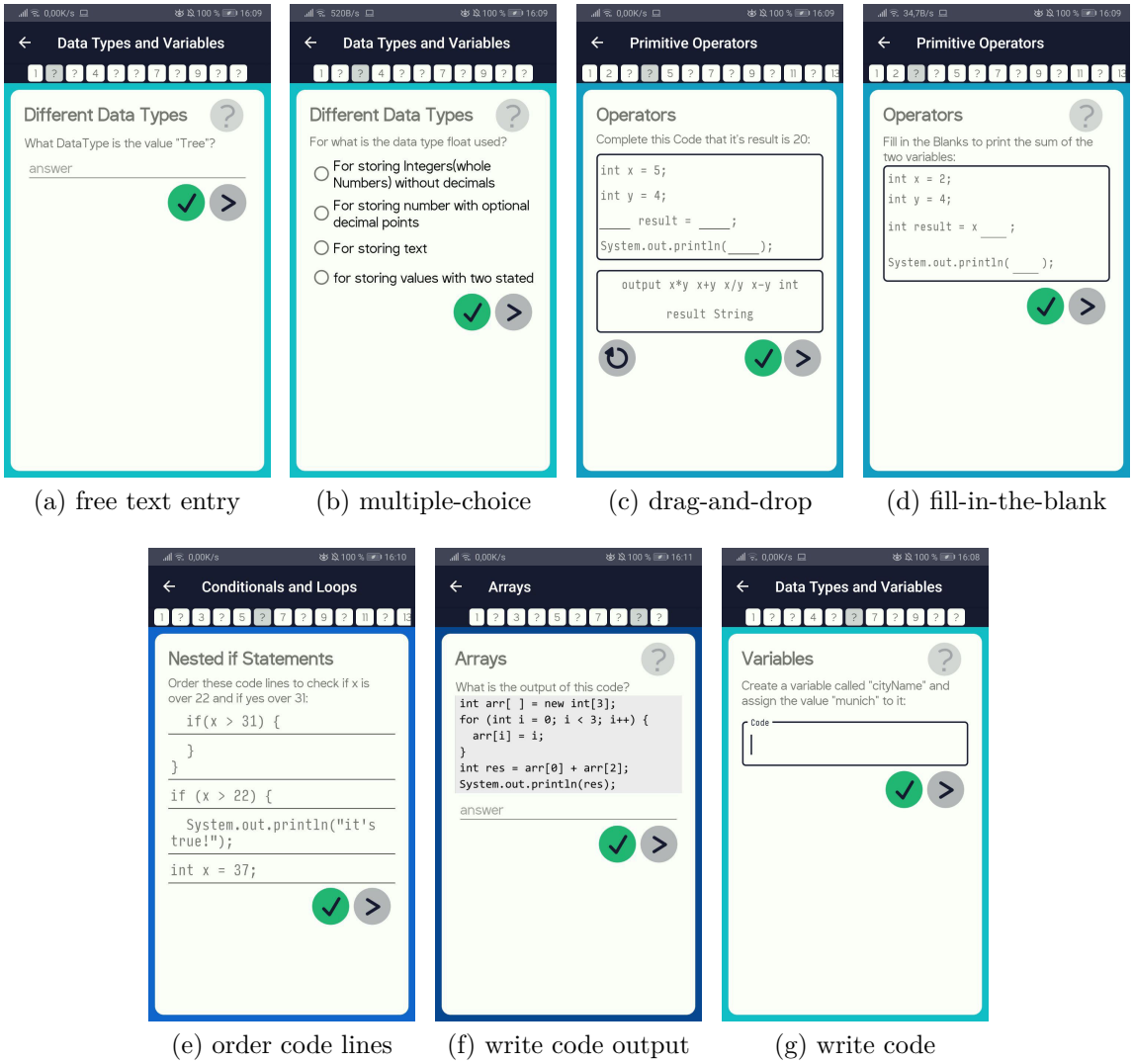
Figure 9.1: The cognitive processing dimensions according to the initial definition by Bloom et al. [35].

map those with answers given in our user study's survey questionnaires. The sign-in is only required once and is then saved for further interactions with the app. The main view displays an overview of the eight sections dealing with the following topics: (1) Tutorial and Introduction, (2) Hello World and Comments, (3) Data Types and Variables, (4) Primitive Operators, (5) Functions, (6) Conditionals and Loops, (7) Arrays, and (8) Classes and Objects (Figure 9.2 shows excerpts of those sections). The app uses Google Firebase[2] to store the users' progress and logging data, in particular, information regarding learning events, session ends, and cue occurrences.

9.2.1 Selection of Learning Content

Currently available mobile micro-learning applications on the market teach primarily languages, in particular, vocabulary knowledge. There are still very few applications teaching more complex content, such as STEM knowledge. This is different in the research domain, as Zydney and Warner [363] report on a variety of applications tested for learning topics such as biology, physics, geology, etc. However, these apps convey basic scientific concepts, often supplemented by multiple-choice tests, and rarely require knowledge construction or synthesis. Prior work suggests using micro-learning for teaching a programming language, as the flexible use of the apps can encourage students to learn at their own pace [311]. Further, teaching programming through micro-learning requires learners to understand the theoretical concepts of the programming language and apply this knowledge by autonomously writing code to solve problems. Skalka and Drlík [311] emphasize that comprehending program code and producing program code are two very different skills. Thus, we consider programming a good example of a topic that requires the teaching of different complexity levels, from recalling basic facts to autonomously creating complex code.

[2] Google Firebase: `https://firebase.google.com/`, last accessed January 3, 2022



(a) free text entry (b) multiple-choice (c) drag-and-drop (d) fill-in-the-blank

(e) order code lines (f) write code output (g) write code

Figure 9.2: The seven different exercise types (a-g) utilized for the *LearnJava* application. Content can be loaded dynamically to the exercise templates.

9.2.2 Task Complexity Measures

To be able to test our assumptions regarding the effects of aligning task complexity and TRC complexity (cf. Chapter 7, Design Impplication 4), we need measures to classify complexity to begin with. As described above, we resort to established models for knowledge and learning objectives to categorize complexity. Prior work suggested the use of Bloom's taxonomy (or its later revised version) for designing a curriculum for programming classes [20, 165, 222]. To categorize our *LearnJava* app's learning content, we draw on Bloom's (1) knowledge types and dimensions, and (2) the cognitive process dimensions (both introduced in the Bloom's taxonomy and later extended and

revised [8, 35, 184]). While the original framework included only the cognitive process dimensions, the knowledge dimensions were added in the revision, resulting in a matrix-like taxonomy table combining both frameworks (see Table 9.1).

Table 9.1: We grouped the different tasks into Complexity Levels (CL) and sorted them according to the knowledge types (rows) they cover and the targeted phase of Bloom's taxonomy (columns).

	Remember	**Understand**	**Analyse**	**Create**
Factual	CL1: Multiple-Choice, Text Answer			
Conceptual		CL2: Drag-and-Drop, Fill-in-the-Blanks	CL3: Output, Order	
Procedural				CL4: Write Code

9.2.3 Lessons and Exercises

Each content lesson comprises a section providing theoretical explanations of the concept and one or more suitable interactive exercises. Those exercises assess the understanding of the theoretical knowledge, and the user receives corrective feedback. In case of a wrong answer, users are encouraged to try again. In addition, there is an option to go back to the content unit if necessary. Only with a correct answer, the next lesson is unlocked. We applied seven different exercise types, namely (a) a free text answer format (Figure 9.2a), (b) multiple-choice questions (Figure 9.2b), (c) drag-and-drop tasks (Figure 9.2c), (d) fill-in-the-blank statements (Figure 9.2d), (e) ordering code snippets (Figure 9.2e), (f) writing code snippet output (Figure 9.2f), and (g) writing code lines autonomously (Figure 9.2g). We designed all exercise types to be modular so that content can be dynamically loaded with content from each session.

9.2.4 Task Resumption Cues

After an interruption occurred and when the user navigates back to the *LearnJava* app, the app continues in the last session the user processed. One out of four TRCs is then presented as an overlay spanning almost the entire screen. The trigger for the cue display could be either a restart of the app after the screen was turned off or an app switch. The users can view the cues as long as they like and continue by clicking a button labeled "Got it" to confirm they perceived the cue. All cue designs are

counterbalanced to match each cue with different task and content complexity levels. Based on the insights gathered in Chapter 7 and 8, we iterated on four cue designs (see Figure 9.3. Due to the positive feedback we received on explicit cue designs, we decided to remove the implicit cues for this evaluation and focus on explicit cues.

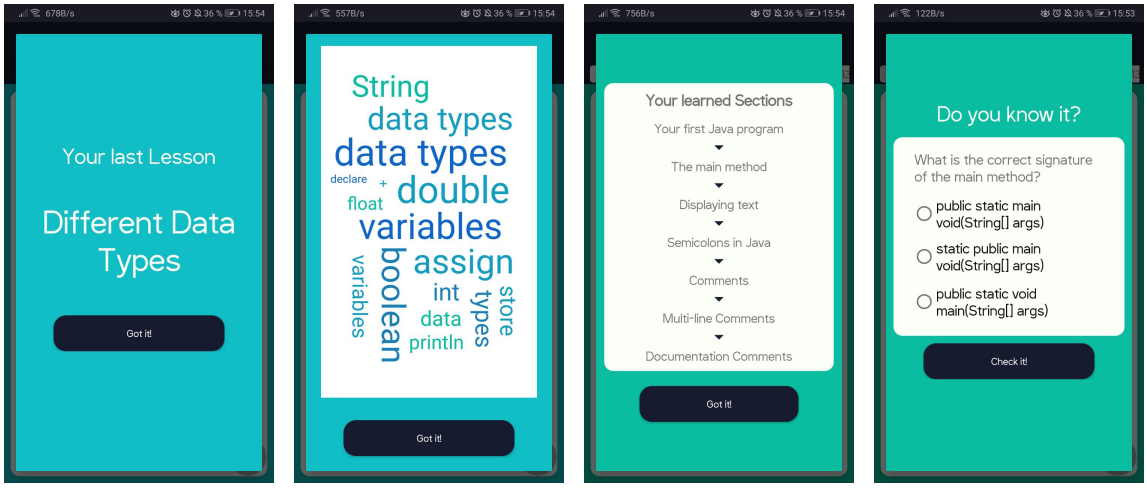
Lesson Name Cue The *Lesson Name Cue* displays the last lesson the user was working on before the interruption occurred (see Figure 9.3a). It is similar to the Color and Icon Cuein that it is supposed to remind the user of the last lesson. However, the screen now shows the expressive name of the lesson rather than an icon. Due to its implicitness and limited content, we rank this cue low in complexity fitting to CL1.

WordCloud Cue The *WordCloud Cue* visualizes domain words in the form of a tag cloud similar to the *WordCloud Cue* in Chapter 8, Section 8.3 (see Figure 9.3b). The cue represents prior content in a more detailed manner than the *Lesson Name Cue* but presents the words out of context. Thus, we assume a medium complexity level for this cue. We expect it to be a good fit for factual recall tasks (CL1) and tasks confirming the users' understanding (CL2).

History Cue The *History Cue* design is aligned with the positively evaluated *History Cue* from Chapter 8 Section 8.2 and gives an overview of the lessons the learner previously solved (see Figure 9.3c). The cue contains information on prior lessons and reminds users of their progress in the *LearnJava* app. It visualizes broader concepts instead of individual words (cf. *WordCloud Cue*) and aims at triggering users' knowledge about the relationship among them. Therefore, we expect a good fit between the *History Cue* and the complexity levels focusing on conceptual knowledge, i.e., CL2 and CL3.

Question Cue Lastly, the *Question Cue* is an adapted version of the *itc* created in Chapter 8 Section 8.3. Both cues present an interactive task aiming to facilitate elaborate rehearsal of the content users learned (see Figure 9.3d). In contrast to the *Interactive Test Cue*, which presents L1-L2 words asking the user to connect the correct translation pairs, the *Question Cue* poses a multiple-choice question. We chose this format as it is more flexible for the generation of questions with different content complexity levels. To adapt the *Question Cue* content to the current session, the app accesses a library that stores the lesson history and a set of keywords of the current session (for the *WordCloud Cue*). Further, the *Question Cue* randomly selects one multiple-choice question from a pool of questions stored for each lesson and gives corrective feedback (correct|incorrect) on the users' input.

Since the cue itself is structured as a multiple-choice task, we expect it to be a good fit for CL1. However, multiple-choice questions can be adapted to check the understanding of conceptual and procedural knowledge. Thus, we also expect the cue to be a good fit for the three other complexity levels CL2, CL3, and CL4.



(a) Lesson Name Cue (b) WordCloud Cue (c) History Cue (d) Question Cue

Figure 9.3: The our different task resumption cues (a-d) implemented in the *LearnJava* application.

9.3 Methodology

9.3.1 Study Design

Our two-week user study followed a within-subject design, as all participants encountered all TRCs and all different task complexity levels (independent variables) throughout the study. The quantitative metrics used in Chapter 8 to determine the effect of the TRCs on mitigating the adverse effects of interruptions (i.e., task completion time and error rates) were not fully applicable to this study setup. The *LearnJava* lessons comprised a theoretical part during which the users had to read and understand the learning content and a second part posing interactive exercises. Interruptions could occur at any point during the lesson. If the app was closed, the user would start again at the beginning of the lesson they interacted with last. Therefore, starting again with the theoretical part. Measuring the error rate in the exercises at the end of the lesson would therefore not provide significant insights into the effects of the TRCs. Similarly, the task completion time is highly affected by the lesson and the users' reading speed, making this metric not feasible either. Consequently, this study focused on the subjective reports of users' on their experience with the TRCs for the different content complexity levels. We pose the following hypothesis:

H_1 Aligning the complexity level of the task resumption cues with the complexity level of the interrupted task increases the perceived helpfulness of the cues.

Additional questionnaires provide subjective insights into more facets of the users' experience with our application and the TRCs beyond mere helpfulness.

9.3.2 Procedure

After the initial recruitment, we provided participants with information on the study procedure, LMU's data protection regulations (aligned with the European GDPR), an installation guide, and the link to a first questionnaire. After the participants provided informed consent, they were able to download the *LearnJava* application from the Google Play Store (submitted as a beta-test version for our participants only). The first questionnaire, including demographic information as well as assessing prior knowledge, experience, and motivation, is later linked to the app usage data via a nickname the users entered[3]. The study ran for two weeks. We encouraged participants to use the application whenever and wherever they saw fit and at their own pace. However, we explicitly stressed that the application's lessons are fairly short and include different tasks; thus, the study would benefit from the participants' interaction with multiple lessons. Although the app supported usage in a mobile context, we did not explicitly recommend a specific usage setting[4]. At the end of the user study, we asked participants to fill in a post-questionnaire to gather additional feedback on the users' experience, particularly the usability of the app and the helpfulness of the TRCs. For taking part in our user study, we rewarded participants with a 20 Euro Voucher for an online shop or an equivalent amount of study credit points.

9.3.3 Sample

We recruited 14 participants via university mailing lists, social media, and other university communication channels (two identifying as male, twelve as female). Their age ranged from 21 to 54 ($M = 25.36$, $SD = 8.56$), and all except for one were enrolled as students. Of those, three were already holding a bachelor's degree and two a master's degree. The fields of study varied and included fields of expertise such as pharmacy, physics, design, psychology, medicine, computer linguistics, engineering, and phonetics. All of our participants owned an Android phone with an Android version of 7 or higher. We primarily recruited study participants with no programming experience (8) or very basic experience (5, stating to have used languages such as C++, Python, or R before). However, one participant stated moderate experience with Processing, Delphi, PHP, JavaScript, and HTML (self-reported). To control for the influence of prior knowledge

[3] We strongly emphasized the need for non-personalized nicknames due to anonymization policies. All nickname choices complied with this policy.

[4] This study was conducted during the lockdown phase of the COVID-19 pandemic. Thus, the results may be limited in generalizability as users' routines, mobility, and smartphone usage can deviate due to lockdown and work-from-home phases. We will discuss this issue in the section on limitations.

on the interaction with our application, we further asked our participants to answer five open-ended questions targeting JAVA basic knowledge (explaining data types, arrays, interpreting a code snippet, etc.). In total, 78.57% (55/70 answers) of those questions were answered with "I don't know", while another 24.29% (17/70) were either incorrect or imprecise. Thus, based on these questions, we do not expect our participants' prior knowledge to influence the study results significantly. In regards to experiences with mobile learning applications, ten of our participants reported that they had used them before, naming the language learning apps Duolingo (7), both Duolingo and Babbel (1), Memrise (1), and a Piano learning application (1). When asked when they would favorably use an application to learn programming, the greatest consensus among our participants arose for usage at home (12/14), in public transport(13/14), or in idle waiting situations.

9.3.4 Results

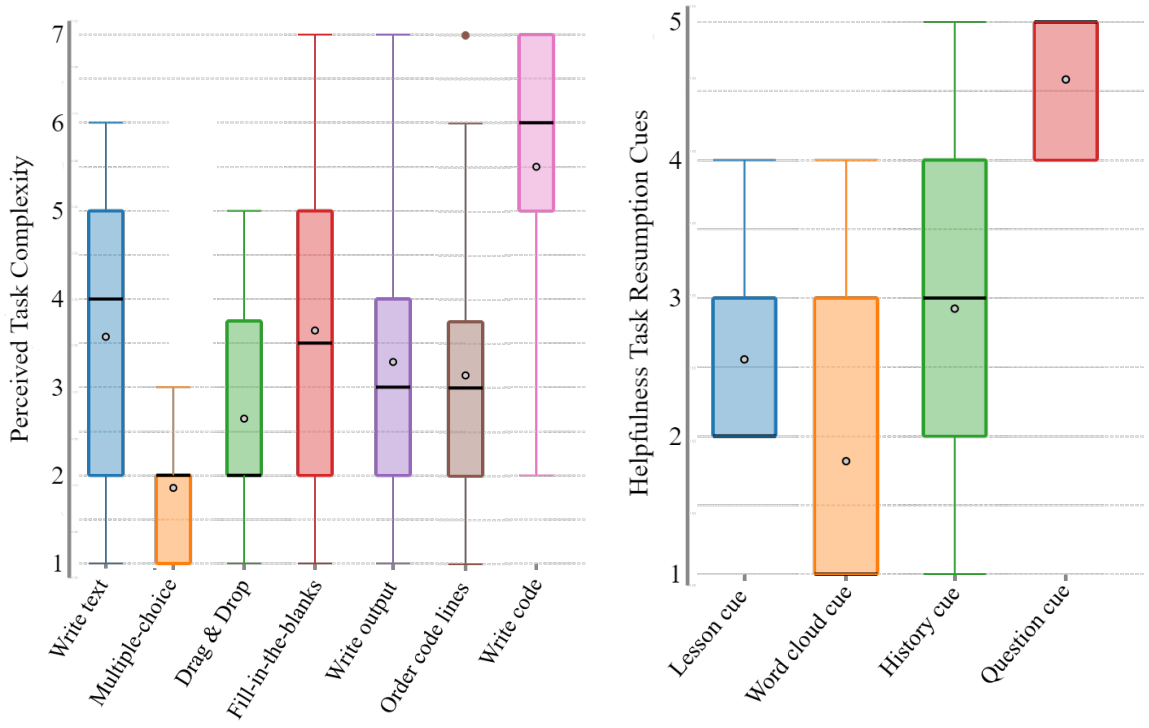
Due to the small sample size of this user study and the high number of different combinations of content complexity and cue type, we will refrain from a quantitative evaluation and significance testing of the collected data.

Difficulty and Complexity Confirmation

To confirm that the content present in our *LearnJava* app was neither too easy nor too hard, we asked participants to express the perceived difficulty of the tasks. *LearnJava* achieved a medium level of difficulty with participants rating it on average with 2.79 (1="very easy", 5="very difficult", $SD = 0.77$). Additionally, we asked the participants to confirm our complexity assessment of the individual tasks. Figure 9.4a shows that while the multiple-choice and drag-and-drop tasks were rated as having low complexity, the write text task was considered as more complex than estimated. Similarly, the fill-in-the-blank task showed higher perceived complexity than the write output and order code lines task.

Overall Cue Assessment

For the subjective assessment of the TRCs, we presented our participants with twelve statements, some of them with opposing views, to rate their compliance on a scale from 1 ="'strongly disagree" to 7="strongly agree" (see Figure 9.5). The questionnaire aimed at the perception of TRCs in general and a summative assessment across all four cue designs. According to the results, the TRCs supported participants' learning experience ($M = 4.21, SD = 1.01$), were a nice repetition of the learned content ($M = 4.29$, $SD = 1.28$), and were not perceived as disruptive ($M = 1.93$, $SD = 0.96$). In regards to the frequency of presentation, participants perceived it as neither too high ($M = 1.93$, $SD = 1.22$), nor too low ($M = 2.93$, $SD = 1.67$). In the following, we will



(a) Complexity

(b) Helpfulness

Figure 9.4: (a) The perceived task complexity of the different task types on a Likert-scale from 1="very simple" to 7="very complex". (b) Participants' overall helpfulness rating of the four task resumption cue designs on a scale from 1="not at all helpful" to 5="very helpful".

take an in-depth look into our participants' assessment of the individual cue designs, specifically in regards to their helpfulness in the learning app context.

Helpfulness, Favorites, and Open Comments

Before asking in-depth questions on the usefulness of the cues and participants' attitudes towards them, we presented a screenshot of each cue and asked participants if they remember seeing it. The cue that was perceived least frequent was the *Lesson Name Cue*, as five participants answered no (2 maybe; *WordCloud Cue*: four no; *History Cue*: one no, one maybe; *Question Cue*: two no). We excluded all further ratings of cues that participants stated not to have seen at all ("no" rating).

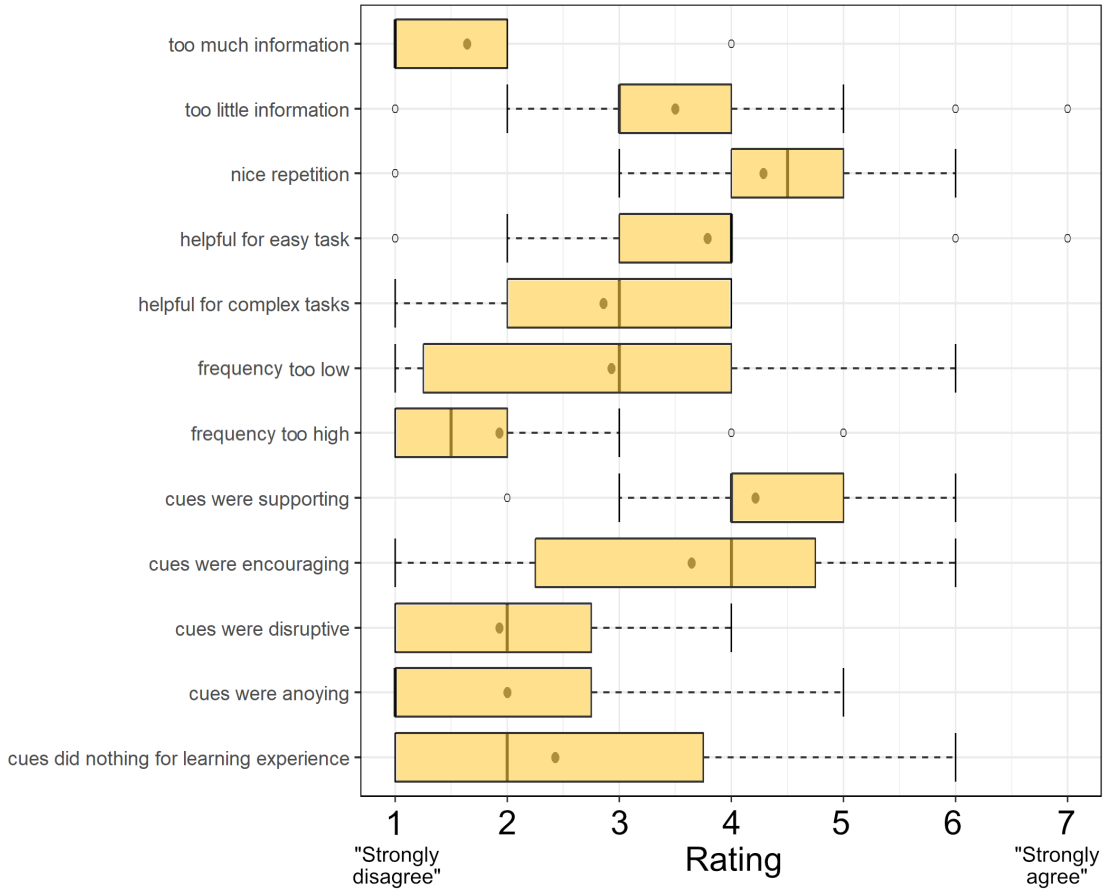


Figure 9.5: Rating of cues in general according to twelve statements.

Afterward, the participants rated the cues' helpfulness for restoring the memory of the previously learned lesson (1="not at all" to 5="very helpful") as summarized in Figure 9.4b. In these ratings, the *WordCloud Cue* received the lowest average helpfulness rating ($M = 1.82$, $SD = 1.11$), while the participants rated the *Question Cue* most helpful ($M = 4.58$, $SD = 0.49$).

We asked our participants to pick their most and least favorite cues to confirm the helpfulness assessment. Eleven participants rated the *Question Cue* as their most favorite, arguing that this cue "*makes you actively think about* [...] *what you've learned*" (P4). Further, it is described as helping to verify the correct understanding of the concept (P3, P7, P10), helping to discover knowledge gaps (P13), and just being a nice repetition (P1) of the priorly learned content. P11 stressed that the *Question Cue* was engaging, saying "*made me actually try to remember the past lessons*".

Of the three participants who did not select the *Question Cue* as their favorite, two named the *History Cue*, explaining that they appreciate having an overview of what they have learned so far (P8, P14). Further, participants stated the cue helped them

order their thoughts (P3), present a nice repetition (P4, P10, P12), and gives them motivation by showing what they have accomplished already (P1). P8 noted that the overview could have been shorter as the screen was crowded by the amounts of past lessons the cue shows.

One further participant preferred the *WordCloud Cue*, describing it as "*nice for visual learners*" (P2). In the open comment section, participants further liked about the *WordCloud Cue* that "*it was easy to notice* [...] *which things I didn't remember that well*" (P10). P3 considered it more adequate for language learning as compared to programming learning. Similarly, P9 noted that "*this cue was helpful at first when I didn't know these terms but progressively it became less useful*". The sentiment was shared by P2 and P6, who did not perceive the cue as useful.

In comparison, ten participants rated the *WordCloud Cue* as their least favorite, arguing that the cue "[...] *seems too chaotic*" (P4), is confusing (P1, P2, P5, P12), and has a low information content (P3, P10).

The *Lesson Name Cue* was disliked by two participants, reporting to either not understand the cue (P2) or to the cue being "[...] *unnecessary due to low information content*" (P10). While many participants did not like the cue, P4 noted in the open comment section to consider it a good indicator to remember where in the app one left off.

Each one participant selected the *Question Cue* and *History Cue* as their least favorite. For the *Question Cue* the participant did not state a reason. For the *History Cue*, P11 reported not feeling engaged enough to read through the history of the prior lessons.

Match Cue and Tasks Complexity

In Table 9.1 we defined the complexity of the learning tasks implemented in the *LearnJava* app. We hypothesized that an alignment of the cue complexity with the task complexity would increase the perceived helpfulness of the cues. We asked our participants for each cue if it was a good fit in terms of helpfulness for each of the exercise types using a simple yes|no answer format. Figure 9.6 visualizes the positive responses, in other words, which cues participants considered a good fit for the individual tasks.

In general, we observe that the *Question Cue* and *History Cue* are considered a good fit for many exercises types. Half of the participants or more consider them helpful for text answers, multiple-choice questions, drag-and-drop tasks, and ordering code. The *WordCloud Cue* and *Lesson Name Cue* received lower helpfulness ratings, which goes in line with participants' overall ratings of these two cues. Especially for the Complexity Level "Analyze", containing Order Code and Write Output tasks, the *WordCloud Cue* and *Lesson Name Cue* were not considered an adequate fit.

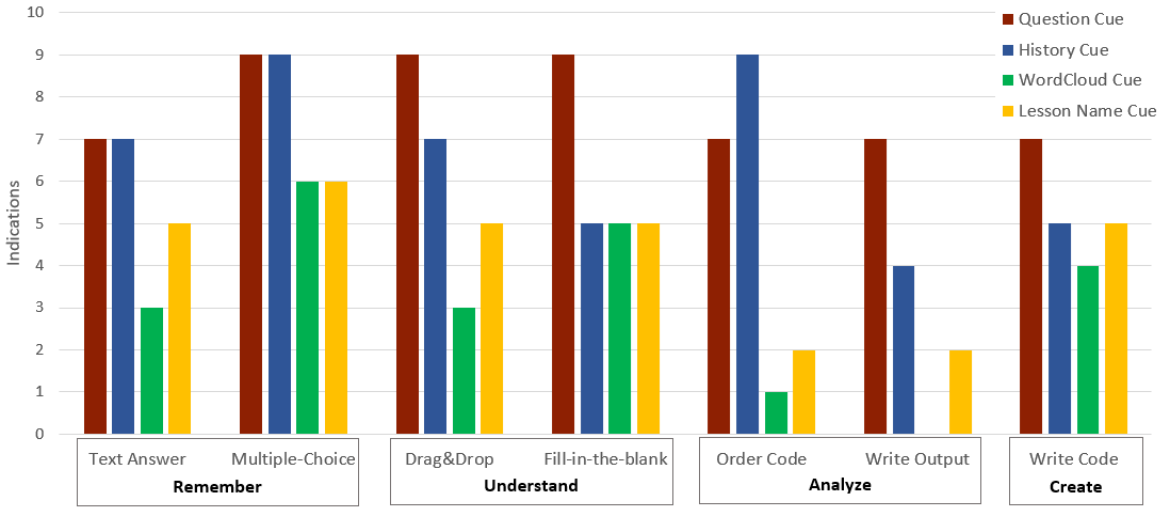


Figure 9.6: This figure visualizes the quality of fit for each cue with the specific task types in regard to their helpfulness. The bars represent a summative view of all positive ("yes") responses across all participants ($N = 14$).

Cue Fit for Lesson Format

To better understand which cue would fit each type of lesson format, we asked participants for their subjective impressions. Particularly, we asked them two questions, if the cue designs would be a good fit for (1) easy and difficult sessions and (2) short and long sessions, as two metrics to categorize lesson formats. The questions were phrased expecting a yes|no answer but further included the options "both" and "neither". Figure 9.7 displays the summative overview of all positive ("yes") responses, indicating the cue would be a good fit for such a format. In line with the overall helpfulness ratings, the *Question Cue* was considered the best fit for all lesson formats. The *History Cue* was rated as a particularly good fit for longer sessions, containing both easy and difficult content, and a less good fit for shorter formats. In contrast, the *Lesson Name Cue* was perceived as a better fit for lessons with easy content. The opinions on the *WordCloud Cue* were diverse; some participants consider it a good fit for short and easy sessions, while others see it as more helpful in long and difficult learning lessons.

9.3.5 Limitations

The presented evaluation is a first outlook on how to generalize the concept of task resumption cues to content beyond language learning. The topic of learning programming is one potential use case that contains simple factual knowledge parts and requires learners to understand concepts and methods (i.e., conceptual and procedural knowledge). Our tasks were designed to cover certain complexity levels (cf. Table 9.1). However, not all of our assessments turned out to exactly represent the complexity we

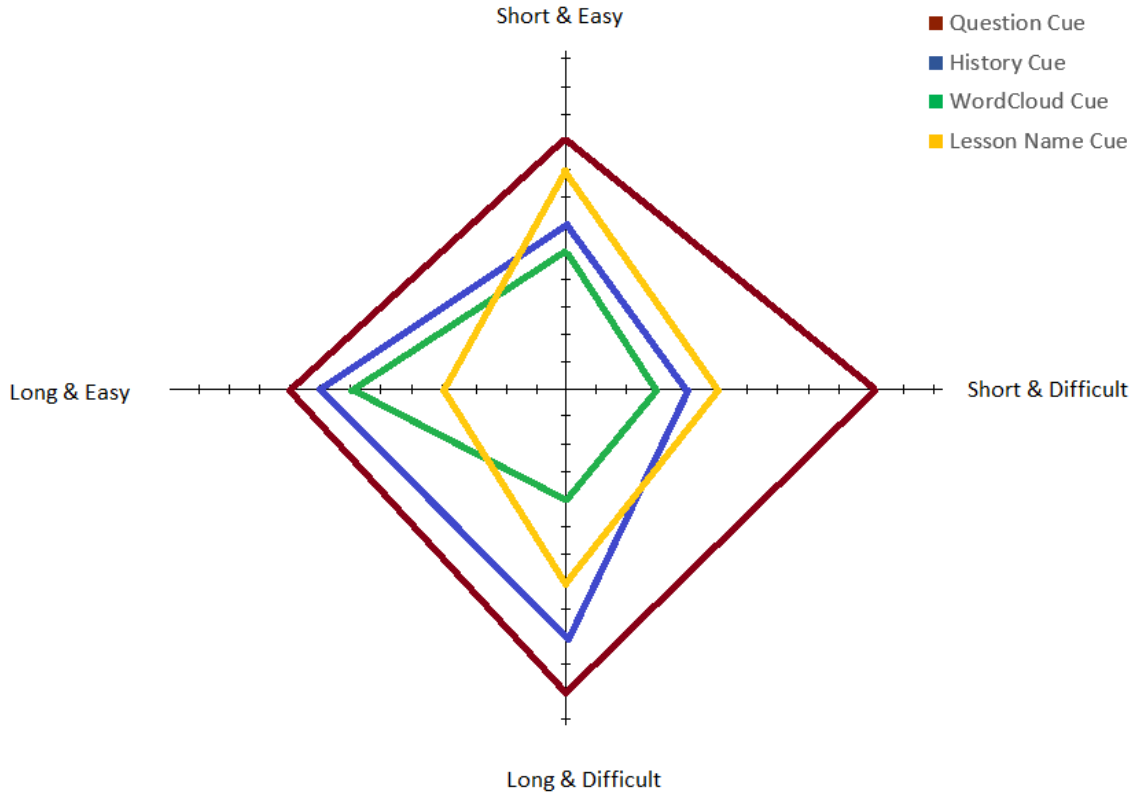


Figure 9.7: Participants' description of which cue would fit to which task type (red *Question Cue*, blue *History Cue*, green *WordCloud Cue*, orange *Lesson Name Cue*. The figure depicts a summative view of all positive ("yes") responses across all participants, where one mark on each axis represents one "yes" vote.

aimed for in the users' perception. For example, while we considered "Write Text" exercises as low in complexity, the free recall of even simple terms appears to be perceived as more complex than a multiple-choice question that assesses conceptual knowledge. Other topics might require different tasks format or use a subset of the ones explored in this study.

Further, a general limitation is that we can not assure that our participants were intrinsically motivated to learn Java and, therefore, actively engaged in the interaction with the app. As this is central for in-depth content processing, we propose that future work evaluate the app as a complementary feature for a university programming course.

9.3.6 Discussion

Transfering the memory cue concept to a topic beyond language learning provides a new perspective on the idea. In Chapter 8, the learning tasks majorly concerned

the recognition or recall of individual words (i.e., factual knowledge). Thus, due to the narrow perspective on knowledge processing applied in this setup, the results' generalizability was limited. In the *LearnJava* app, we aimed to include a variety of complexity levels and more diverse knowledge types.

Match of Complexity Levels

When analyzing our results in regard to our main hypothesis – stating that aligning the complexity level of the TRC with the complexity level of the interrupted task increases the perceived helpfulness of the cues – the data revealed no clear indication that gives us reason to confirm our hypothesis at this point. Rather than seeing a linear relationship between the cue and task complexity, we observe a stronger influence of users' overall helpfulness rating on our results. In other words, the users seemed to favor the *Question Cue* and *History Cue* over the *WordCloud Cue* and *Lesson Name Cue*. Yet, the evaluation indicated that the perceived helpfulness of the cues is related to the task and lesson they are applied in. For instance, the *WordCloud Cue* and *Lesson Name Cue* were not considered helpful at all when applied in "Analyze" tasks (cf. Figure 9.6). Further, the *History Cue* appeared to better fit longer lessons (cf. Figure 9.7) and especially Multiple-Choice and Order tasks (cf. Figure 9.6).

9.4 Chapter Summary

This chapter represents an outlook on applying task resumption cues beyond the domain of language learning. We built the *LearnJava* mobile application teaching basic JAVA knowledge and including four revised task resumption cue designs. By exploring users' subjective experience of the cues with different levels of content complexity, we gained interesting insights into their fit to certain exercise types. A central observation was the positive feedback we received for the *Interactive Test Cue*, independent from the task or content it was presented in. Nevertheless, all designs of our TRCs are simple forms of memory cue aimed at triggering the recall of priorly encountered information from the LTM. Especially when dealing with complex content, further ideas proposed in the design implications of Chapter 7 should be considered. For example, showing the memory cue in the interruption lag (i.e., on closing the app) similar to the work of Clifford and Altmann [67] could help to encode the learning goal of the respective session. Other ideas such as asking the learner to write a quick note summarizing the content they just learned and presenting this sentence at the beginning of the next learning session could be helpful. Considering research question **RQ4**, we can again refer to the positive feedback we received to the idea of using memory cues as task resumption support in general. Future work is needed to confirm the effect on the actual learning process using quantifiable metrics.

V

Conclusion and Future Work

10
Conclusion and Future Work

This thesis investigated users' everyday mobile language learning (MLL) behavior and explored concepts and applications for the seamless integration of second-language acquisition into peoples' lives. We followed a user-centered design process in which we first assessed common usage situations of MLL apps and investigated the prevalence of interruptions and distractions. Further, we explored two primary pathways of supporting learning in everyday settings through the evaluations of individual research probes: (1) embedding learning into everyday technology interactions (such as smartphone authentication and media consumption) and (2) mitigating the adverse effects of interruptions by supporting users' task resumption support.

Before we outline our contributions in regard to the research questions stated in Chapter 1 and reflect on current limitations (Section 10.2) and future research directions (Section 10.3), we want to revisit the scenario of Anna. Chapter 1 introduced Anna as a 20-year-old girl planning to learn Portuguese for a summer vacation in Portugal. Anna encountered several problems in her interaction with the mobile learning application that represented challenges in current mobile learning practices. In Part II of this thesis, we extended our knowledge on the usage of mobile learning applications and derived challenges and opportunities, resulting in **Contribution 1:** Characterizing Mobile Learning and Interruptions in Everyday Settings.

We now imagine that the concepts designed and evaluated in the research probes of this thesis would be integrated into a future mobile learning app (in this example called *'EmbeddedLearningApp'*). Anna's experience of learning a language would look very different. The following text outlines how Anna's challenges would be tackled and how future learning with the *EmbeddedLearningApp* could look like:

> SCENARIO
>
> In preparation for her trip to Portugal, Anna downloads the mobile language learning application *EmbeddedLearningApp* to her smartphone. She sits down on the couch and starts the app at the beginner's level. Anna spends another two hours answering multiple-choice questions on basic phrases, animals, and clothing items.
>
> She picks up her phone during her lunch break. The *EmbeddedLearningApp* reminds Anna that she left off at the lesson concerning 'Animals' last night after learning about clothing and basic phrases (History Task Resumption Cue, cf. Chapter 8): Translate the word 'Cão'. Anna tries to remember: Since I recently learned about animals,

this must mean 'Dog'. Correct! – **Contribution 4:** Mitigating the Effects of Interruptions and Infrequent Practice

The days go by and Anna is unusually busy with work. Even though Anna did not have much time to invest in practicing Portuguese, the *EmbeddedLearningApp* presents Anna with one learning task at each smartphone authentication during the day (cf. Chapter 5). Particularly, the app improves her retention by repeating priorly learned Portuguese vocabulary, so Anna does not diminish her progress due to infrequent practice. – **Contribution 2:** Exploring Opportunities for Embedded Mobile Language Learning

One weekend, Anna is not motivated to engage in learning on her phone but decides to watch a cooking show with Portuguese subtitles on Netflix instead. She puts on her mobile EEG headset and listens to the host: 'Estamos preparando pastel de nata polvilhado com canela.' [We are preparing egg custard tarts dusted with cinnamon.] By analyzing Anna's brain response, the *EmbeddedLearningApp* recognizes that she is unfamiliar with the word 'polvilhado'. As she must be interested in the content, the word and its translation are added to Anna's vocabulary list for her to learn (cf. Chapter 6). – **Contribution 3:** Implicit Personalization of Learning Content According to Interest and Comprehension Levels

The weeks go by and while Anna engages in one or two focused learning sessions with the *EmbeddedLearningApp* per week to learn Portuguese grammar, she keeps repeating vocabulary on her phone's lockscreen.

Finally, Anna arrives in Portugal, and, sitting in a small café on Porto's Rua de São João Novo, can now say - "Um café e uma pastel de nata polvihado com canela, por favor". Anna is happy that she is able to order her food and drinks, and that *EmbeddedLearningApp* supported her frequent engagement with the language and adapted to her content of choice.

10.1 Summary of Research Contributions

This section will summarize the insights and challenges revealed by this thesis concerning the five research questions (RQs). We reflect on the implications of this work along with its four major contributions: (1) *Characterizing Mobile Learning and Interruptions in Everyday Settings*, (2) *Exploring Opportunities for Embedded Mobile Language Learning*, (3) *Implicit Personalization of Learning Content According to Interest and Comprehension Levels*, and (4) *Mitigating the Effects of Interruptions and Infrequent Practice*.

10.1.1 Characterizing Mobile Learning and Interruptions in Everyday Settings

We summarize our assessment of how people engage with MLL apps in their everyday lives. In Chapter 3, we performed an online survey ($N = 74$) in which users reported on their most common learning situations. Besides contextual factors such as environment and company, we inquired about learners' habits of planning learning activities, frequency of learning, and the occurrence of multitasking. We ultimately derived three design recommendations based on our findings. Additionally, Chapter 4 presented a field study ($N = 12$) investigating the occurrence of interruptions and their effects on the learning sessions. For this purpose, we combined log data with in-situ reports collected through experience sampling questionnaires to gain complementary insights. We concluded this part of the thesis by presenting ideas for mitigating the effects of interruptions in everyday settings. These evaluations addressed the following research questions:

RQ1a: How do people use mobile learning applications in everyday settings?

RQ1b: How do interruptions affect mobile learning in everyday settings?

Ubiquity of Mobile Devices Results in Diversity in Usage The omnipresence of mobile technology in our lives creates the opportunity to learn anytime and anywhere. Prior research has already shown that users appreciate the chance to fill gaps in their daily schedule with short learning sessions (cf. [85, 91]). Chapter 3 aimed at finding out more about the characteristics of those learning situations. Participants' reported usage situations of mobile learning applications revealed that their interaction with the apps frequently occurs at home, despite the opportunities mobile devices entail. However, our study showed that learning at home in itself is very diverse. The combination of time of day, company, duration, and device type used for learning creates individual patterns of learning situations that come with specific opportunities and requirements. Situations that users plan routinely with low stress, no company, and long duration can provide the chance to deeply engage in a topic and focus on the acquisition of difficult or complex topics. In contrast, short and spontaneous learning sessions on the sofa in the evening while being surrounded by family members might be a better fit for short and easy repetition tasks to foster retention of previously learned information.

Interruptions are Part of Learning in Everyday Life Settings The evaluation of usage situations in Chapter 3 already showed that learning in everyday settings is not as controlled as in formal settings. Learning in public places, with company, or while watching TV, can lead to less focused learning and distractions. In Chapter 4, we aimed at shedding further light on the prevalence of interruptions and confirmed that they occur at least in one out of three learning sessions. Especially when multiple

interruptions occur, users reported losing focus, and the chances for terminating the learning app increased. The frequency of interruptions (i.e., 276 interruptions in 327 learning sessions), which originated from the mobile device, the environment, or the users themselves, showed us that they are a part of everyday life. Therefore, we conclude that it is of central importance that mobile learning applications factor in interruptions as a frequent everyday event.

Interruptions Cannot be Avoided, but their Effects can be Mitigated In Chapter 4, the participants noted that at least 20% of the interruptions they experienced could have been ignored or postponed. Still, they followed the distracting stimulus and interrupted their learning session, even though notification management systems (e.g., [111, 157, 251, 264]) can help with low importance interruptions and when interruptions originate in the mobile device. We propose that instead of focusing on avoiding interruptions, which is often not possible in everyday mobile learning situations, applications should rather focus on supporting a seamless and immediate resumption of the learning activity afterwards.

10.1.2 Exploring Opportunities for Embedded Mobile Language Learning

We interact with our smartphones so frequently during the day that it becomes a feasible tool to nudge users to engage with certain tasks. Previous work has shown that displaying tasks on the lockscreen (i.e., learning tasks [81] or nutrition tracking [167]) can increase task exposure. Through the integration of tasks into the authentication action, we have the opportunity to create less intrusive tasks and increase the frequency of interaction [362]. In Chapter 5, we transferred this concept to the presentation of learning tasks. We derived twelve different designs to seamlessly integrate learning tasks into different authentication mechanisms, and evaluated them regarding their usability ($N = 10$). We implemented one of the prototypes as a mobile learning application (*UnlockApp*). We performed a field evaluation ($N = 30$) comparing users' behavior and experience with the *UnlockApp* to a notification-based learning app and a standard learning app. The central research question behind this approach was:

> **RQ2:** How can the integration of learning tasks into the smartphone authentication process foster frequent engagement?

Embedding Learning Tasks in Smartphone Authentication Increases Engagement The comparative evaluation showed that the *UnlockApp* and *NotificationApp* designs, compared to the *StandardApp*, led to approximately 20-40% more interaction with the learning content. Furthermore, the presentation of vocabulary tasks after the authentication with the *UnlockApp* led to a more spread-out exposure of tasks across the

day. While the interaction with the learning content in this approach is short and limited regarding the depth of processing, the embedded learning concept can foster high repetition counts. In particular, the frequent interactions allow for good integration of a spaced repetition schedule for teaching declarative knowledge (e.g., vocabulary learning along with the Leitner Index [123]) and thereby foster retention [168].

Offering Different Levels of Intrusiveness Addresses User Needs and Preferences The initial usability evaluation showed that participants appreciate the idea of being nudged to learn more frequently through the embedded authentication learning approach. However, they stressed the need for a quick and simple interaction. Our comparative evaluation confirmed this user expectation. The study revealed that for continuous usage, our participants would favor the *NotificationApp* over the *UnlockApp* due to its lower intrusiveness. We observed a division among our participants; while some appreciated being nudged toward more frequent interaction, others perceived the constant learning impulses as intrusive and annoying. This division could be an indicator for different levels of self-regulation capabilities among participants and their different needs. Self-driven learners who can set aside time to learn might profit less from an application that proposes frequent prompts, whereas learners with lower self-regulation skills might benefit more. Especially considering the latter case, we conclude that embedding learning tasks into common smartphone interactions is a promising concept for fostering frequent learning interaction with comparably low intrusiveness.

10.1.3 Implicit Personalization of Learning Content According to Interest and Comprehension Levels

In Chapter 6, we explored embedding foreign-language comprehension assessment into digital reading and listening. Media content such as foreign-language texts or subtitled movies can support effective learning [134, 360]. To tailor the learning process to users' language proficiency, we propose using Electroencephalography (EEG) as an implicit method for detecting unknown vocabulary. In a first laboratory user study ($N = 10$), we investigated the use of event-related potentials to detect differences in users' processing of known and unknown words when reading foreign-language texts. A second user study ($N = 16$) extended the approach to also encompass auditory language comprehension and reduced the number of electrodes needed for the analysis. We aligned our evaluation with the following research question:

RQ3: How can we utilize users' everyday reading or listening activities to generate personalized language learning content?

EEG Can Be Utilized to Reliably Detect Unknown Vocabulary In both user studies, we used differences in ERP amplitudes as the metrics to detect word-based incomprehension. The first evaluation found significant differences between amplitudes of known and unknown words. Based on these initial results, we went one step further and trained a person-dependent classifier in the second study. This classifier showed that we can recognize unknown vocabulary during reading and listening with sufficient accuracy, i.e., 87.13% in reading and 82.64% during listening. While current EEG is not yet ready for language comprehension assessment in the wild, we argue that first studies on EEG in real-life situations paint a promising picture for future applications (cf. [34, 82, 343]).

EEG Analysis Enables Personalization Beyond Explicit Input We investigated EEG as a technology for implicitly detecting and extracting words that users could not translate while watching (foreign-language) movies with subtitles or listening to podcasts of their choice. This approach would enable us to generate learning content personalized to users' *interests* and also their specific *proficiency levels*. Both factors are known to have a significant impact on learners' motivation. Learning with personalized content (e.g., media content matching personal interests) can increase intrinsic motivation [70], which is considered a key factor for learning success [84, 161]. While learning applications could ask users to define their interests and proficiency levels explicitly, explicit or self-ascribed personalization is time-consuming, potentially inaccurate, and limited in scope.

10.1.4 Mitigating the Effects of Interruptions and Infrequent Practice

In the evaluations of Chapters 3 and 4, we observed the prevalence of interruptions and the risks of infrequent engagement with mobile learning applications in everyday settings. Due to the often unpredictable nature of interruptions, we suggested focusing on supporting the resumption of the learning task after the interruption through memory cues. In Chapter 7, we performed an extensive literature survey on memory cues across domains and derived six design implications. Based on these implications, Chapter 8 introduced four designs for task resumption cues in MLL. Two user studies, one in a lab environment and one in the wild, confirmed the negative effects of interruptions and evaluated the cue designs regarding performance measures and user experience. We further presented an outlook in Chapter 9 on how task resumption cues could be generalized beyond the use case of MLL. The main research question behind these research probes was:

> **RQ4:** How can we use memory cues to mitigate the negative effects of interruptions in everyday mobile language learning?

Task Resumption Cues Support Reflection Across all evaluations, most study participants appreciated the implementation of task resumption cues in the learning applications. While we could not show quantifiable effects of the cues on learning performance, users' subjective assessment of the cues was undoubtedly positive. Besides the fact that the cues were a helpful summary, our participants further noted that the cues made them think more thoroughly about what they had learned and gave them an opportunity to reflect on their understanding of the content. However, certain cue designs, such as the *Interactive Test Cue* and *History Cue*, received more positive feedback throughout the studies than other cue designs and should therefore be favored for further evaluations.

Users Favor Explicit and Interactive Cues Regarding the perception of different memory cue designs, a central criterion for distinction is how immediately they trigger prior memories. An implicit cue, which presented a lesson name as a trigger, required our learners to think back and try to retrieve the content of that lesson from their long-term memory actively. In contrast, explicit cues such as the *Interactive Test Cue* presented a question from among previously answered ones, showing the correct solution as one of several potential answer options. Having only to recognize the correct answer from the set, this kind of cue created a more direct link to their prior memories. Overall, cues containing more information were favored over less detailed or implicit cues. Adapting the cue granularity in respect to user needs, for example, by considering the severity of the interruption or length of learning break, could further increase the perceived helpfulness of the cues.

Task Resumption Cues Connect Micro-Learning Sessions The results of our evaluations indicate that memory cues have the potential to support task resumption after interruptions or learning breaks. Previous work on memory cues has been shown to support users in providing task context [252] or spatial context (e.g., [56, 62]), summarizing content (e.g., [309]), or reminding them of their goals (e.g., [258]). By applying the concept of memory cues, particularly task resumption cues, to the domain of mobile learning, we envision moving beyond the limited scope of micro-learning. Specifically, applications do not have to be limited to micro-content units if we can support users by closing the gap between two learning sessions with a "memory bridge". Our subjects stated that cues such as the *WordCloud Cue* are good reminders and summaries and made them actively think about prior lessons.

10.2 Reflections and Limitations

This section will reflect on current and future research of embedded MLL. We will describe challenges of mobile learning and mobile learning evaluations and propose recommendations to help researchers and designers of MLL apps to overcome them.

Generalizability beyond Language Learning

In Chapters 3 and 4, we asked our participants to report on the applications they use for learning. The great majority indicated that they use mobile applications for learning a language. Only a handful stated that they use general flashcard-based learning applications to learn other content. We infer that at least in our sample, mobile learning seems to revolve predominantly around language learning, and other learning apps are used rarely or temporally, for example, to prepare for university exams.

We believe that this presents opportunities for greater generalizability beyond language learning. For example, we can imagine that the *UnlockApp* implemented in Chapter 5, which presents one multiple-choice learning task with every smartphone authentication, could easily be generalized to micro-learning in other topics. Since this concept aims to increase users' exposure to content in short but frequent interactions, its main benefit lies in increasing repetition and strengthening retention rather than knowledge acquisition. Every information chunk that can be used could be translated into a question in a pop-up window with a multiple-choice answer format. The most feasible content for tests with recognition and short recall tasks would be declarative "know-what" knowledge, including factual and conceptual knowledge. For example, the approach could be beneficial for teaching STEM topics such as definitions of biological terminology, abbreviations of chemical elements, or other topics such as mapping historical events to years or stating capitals of the world's countries. Further, early-stage research projects have explored how to support learning with user-selected content on mobile devices. By extending Personal Knowledge Management Applications (PKMAs) with a learning feature, highlighted sections of news articles or notes on podcasts could be turned into questions posed to the user to repeat the respective content [306].

Nonetheless, as already discussed in Chapter 5, we propose to view micro-interaction, such as through the *UnlockApp* concept, as a complementary feature. Particularly when applying it for complex content beyond vocabulary training, the strength of embedded mobile learning lies in supporting the frequent repetition of all content one needs to learn by heart. Embedded learning could extend a fully-featured mobile or desktop application that focuses on long and intense learning sessions on-demand and allows for in-depth processing of complex topics. Beyond the mere generalization of our concepts to other learning domains, we believe that our concepts can contribute strongly to the feasibility of extending the micro-learning paradigm beyond language learning. Considering the concept of task resumption cues we outlined in the literature analysis of Chapter 7 the various application domains of cues. While we see particular helpfulness of memory cues for infrequent or interrupted learning sessions, our evaluations indicate that the helpfulness of task resumption cues might increase with higher content complexity.

Unpredictability of Cognitive Processes

All processes related to learning, particularly attention, memory encoding and recall, are susceptible to factors such as attention diversion, stress, or environmental noise. Other factors like arousal (cf. the Yerkes-Dodson law [357]) or motivation to learn are strongly associated with learning success [263]. In all our laboratory studies, which (implicitly) assessed cognitive processes, i.e., Chapter 6 (comprehension), and Chapter 8 Section 8.2 (memory recall), we aimed at creating a quiet and stress-free environment for our participants to isolate potential confounding variables. Further, each study included at least 10 participants to compensate for individual differences and followed a carefully designed study procedure. After the first laboratory evaluations, we tested some of our research probes in the wild (cf. Chapter 5 Section 5.4 and Chapter 8 Section 8.3), thereby investigating users' experience interacting with the applications in their natural everyday settings.

Furthermore, cueing prior memories depends on solid encoding of the information in the first place. If a learner is distracted during practice or lacks motivation for in-depth engagement, the information might only get superficially encoded. Its memory will decay after a few days making it more challenging to recall [294]. The learning performance and the helpfulness of task resumption cues will significantly decrease as they rely on the encoding. One promising approach to ensure the proper encoding of learned information is to use the interruption lag, which is the time between noticing an upcoming interruption and switching to this secondary task. This recommendation (already stated in Chapter 7 Design Implication 2) is not limited to research settings but can also foster the helpfulness of resumption cues, and therefore the learning performance in the actual MLL app. If an interruption can be anticipated, the resumption lag can be used for encoding the knowledge and user's goals ("What was I doing before?" and "What was I about to do?" [5]). Users could be encouraged to create short summaries or notes before exiting the application, creating their own memory cues for the later resumption.

On a more general level, we advise researchers and practitioners to pay careful attention when evaluating cognitive processes during learning. We recommend to (1) select participants who are intrinsically motivated to engage with the learning content to ensure natural behavior and proper information encoding and (2) apply complementary metrics for the evaluation of cognitive processes. Specifically, we propose using both objective, quantifiable measures if applicable (e.g., error rates or task completion time) and supplement them with subjective qualitative measures (e.g., ratings or user-reported insights). While cognitive processes will never be entirely predictable and remain prone to being affected by various internal and external factors, this combination of metrics can provide a well-rounded view of users' learning experience.

Long-term Assessment of Learning

We performed the evaluations in this thesis in the lab and the field. Especially when studying learning interactions in users' everyday life settings, we can observe almost natural behavior. However, all evaluations were limited in their duration to a period of two to four weeks, so even though we can draw initial conclusions and implications based on the collected data, the generalizability remains limited. Social expectations or pressure can be extrinsic motivators (i.e., using an application during a study to receive the compensation), and new technologies can result in biases caused by novelty effects. If technology is the factor that motivates the learner, the motivation is not persistent for long-term engagement. Once accustomed to novel learning technologies, users' motivation and willingness to use technology-based learning can decline [174, 187]. Intrinsic motivation (i.e., learning out of personal interest or fun) is better suited as an incentive and considered a key to ensuring perseverance and long-term learning success [84].

As of today, research on the long-term usage of mobile learning applications for lifelong learning is sparse. While language learning is not a task that requires frequent and intensive studying over decades, other mobile learning tasks might. For example, a flashcard-based learning app that encompasses topics from one's study program could be a study companion for around three to five years. We argue that we need to move beyond assessing short-term learning outcomes and rather focus on fostering long-term engagement and repetition with the content. This focus becomes particularly relevant if we aim to move beyond micro language learning. Thus, it remains essential to confirm the acceptance and adoption of our proposed concepts in the long term. We are confident that embedded mobile learning, with its focus on improving intrinsic motivation due to seamless and personalized interaction, has the potential to be accepted for lifelong learning.

10.3 Future Work

The work presented in this thesis outlines several starting points for future research. In the following, we will describe how our concepts and approaches can be generalized, extended, and transferred to other domains and where additional research is needed beyond this work's scope. We discuss future work opportunities for the short-term (1 year), mid-term (5 years), and long-term (10+ years).

In Chapters 5 to 8, we probed our research ideas with the use case of language learning. We argue that the concepts explored in this thesis are generalizable to other learning domains. **Short-term** future work is needed to explore the concepts' potential and evaluate their feasibility beyond language learning. Examples include, but are not limited to, the use of the *UnlockApp* and the application of task resumption cues for learning STEM topics. The learning content could be pre-defined or generated based

on user-selected content or interests. Moreover, concepts such as task resumption cues could even be applied beyond the learning domain to any task performed on mobile devices in everyday settings. As the prevalence of interruptions is not restricted to learning applications, users could benefit from memory cues when reading a book, writing an email, searching for a new apartment, or shopping for groceries online. To successfully implement such cues, future work is further needed to improve the detection of interruptions and the process for estimating their duration and severity. These factors are essential for designing targeted resumption cues. To improve the recognition of interruptions, additional physiological sensing approaches such as proposed by D'Mello et al. [92] and Steil et al. [318] could be utilized to complement device-internal detection and forecasts.

Therefore, we see great potential for research in the **mid-term** future for investigating the application of memory cues to help learn more complex topics in mobile and interruption-prone environments. Considering the changes in educational paradigms (some being facilitated by the COVID-19 pandemic), we argue that it will become increasingly important to fit learning more seamlessly into our everyday lives. Particularly, we see the need to adapt the way we teach complex content not to require long and static interaction but to what can be performed in shorter sessions. Integrating task resumption cues into mobile learning can connect multiple independent microlearning sessions, making learning complex content such as STEM topics more feasible on mobile devices. When we think about other devices, we further think that some of our ideas can benefit learning at scale, such as in Massive Open Online Courses (MOOCs). Schools and universities often provide the same learning content to everyone as they have to serve a huge number of students at the same time. To foster lifelong learning, it is necessary to target users' interests and allow personalized learning even in MOOCs. There are first research insights into how personalized learning can be integrated into MOOCs to support enabling learners to set their own learning objectives and goals [282, 283]. By including methods of implicit attention and comprehension assessment more broadly, such as eye-tracking or other physiological sensing approaches [169], we can better adapt learning content and assessments to learners' estimated proficiency. Including students' interaction with digital media content such as podcasts, documentaries, or news articles could further help tailor the content even more to their interests, thus increasing their motivation.

To inform the **long-term** research perspective, we have to look at how users' interaction with technology will change. Smartphones are currently the most prevalent computing device in our society. For some activities in our daily lives we have already moved beyond mobile interaction toward a more ubiquitous interaction paradigm. We use smart voice assistants at home to answer our questions and play our favorite music and we use smart watches to receive text messages and track our fitness workouts. Even though smart glasses and garments are not widely adopted yet, they could become part of our daily interaction, as they have reached the consumer market. The latest research already argues that for the future of Human-Computer Interaction, we

need to consider a trend toward Human-Computer Integration, where users and technology are interwoven [243]. Technologies that augment and enhance the human, such as memory augmentation tools or implicit sensing mechanisms (e.g., EEG as in Chapter 6), follow this trend and we can envision technology unobtrusively becoming more embedded in our everyday lives. With the increasing reliability and steady improvement of physiological sensing technologies, we see great potential for EEG to make comprehension assessment in the wild possible in the future. Prior work has shown that brain response analysis can become unobtrusive and take place outside laboratory environments by embedding EEG electrodes in caps and glasses [34, 82, 343]. Thus, we believe that using EEG to extract comprehension problems during digital reading and listening might be extended to any everyday perception of text or speech in the future.

Regarding learning technologies, even the adoption of smartphones as a medium is still ongoing, as we see that many learners still rely on analog methods (e.g., books) and laptops, particularly for longer learning sessions (cf. Chapter 3). While today's technology has the potential to revolutionize the way we learn, technological adoption in education is slow. When we look back at how learning at schools and universities looked 50 years ago and today, the similarities by far exceed the changes. While students still sit in university lecture halls listening to a professor teaching in front of the class for hours, we already have the technology at hand that could help us move toward learning that is tailored much more perfectly to every individual.

10.4 Concluding Remark

This thesis investigates how we can support peoples' learning with mobile devices in everyday settings. It targets fundamental challenges that occur when taking learning activities from the formal context of a classroom into an informal and uncontrollable environment. Due to the ubiquity of mobile devices, learning can occur in a great diversity of situations, requiring designers and developers to account for the inherent risk of interruptions and distractions. Yet, mobile learning also inherits the opportunity to be seamlessly embedded into users' everyday live, creating a learning experience tailored to personal needs, preferences, and proficiency levels. As future technologies will keep changing the way we learn, future research will need to continue to reevaluate users' learning experiences to exploit the full potential technology has to offer.

Acronyms

ANOVA Analysis of Variance. 63

BCI Brain-computer Interface. 128

EEG Electroencephalography. 17

EOG Electrooculogram. 116

ESL English as Second Language. 105

ESM Experience Sampling Method. 9

ESQ Experience Sampling Questionnaire. 56

GDPR General Data Protection Regulations. 17

HCI Human-Computer Interaction. 8

LAIRA Learning Activity and Interruption Recognition Application. 56

LTM Long-term Memory. 22

MLL Mobile Language Learning. 3

PIN Personal Identification Number. 75

RFID Radio-frequency identification. 148

RSVP Rapid Serial Visualization Presentation. 11

STM Short-term Memory. 22

SUS System Usability Scale. 184

SVM Support Vector Machine. 125

TRC Task Resumption Cue. 135

TTS Text-to-Speech. 117

UCD User-centered Design. 8

UI User Interface. 32

UX User Experience. 64

WPM Words per Minute. 107

Glossary

Embedded Learning

The term refers to learning with embedded technologies or technology as a medium to enable interactions integrated into users' everyday lives..

Implicit Personalization

Tailoring learning content, sequence, or assessment to the end-user without the need for their active input or awareness of the personalization.

Interruption

An event or action that requires users to shift their attentional focus away from the primary task toward a secondary task. The duration of an interruption can range from seconds up to days, which could then also be called a learning "break". The interruption ends when the user resumes the primary task (cf. [329]).

Lifelong Learning

Describes the continuous expansion of our knowledge beyond the scope of formal education lasting over our whole lifetime.

Memory Cue

A memory cue is a trigger to help users recall information from their long-term memory. A cue can be anything connected to the information, e.g., a word, image, sound, or smell, and helps retrieve the respective knowledge (cf. [141]).

Micro-Learning

A learning approach that focuses on presenting *micro-content* units in *micro-interactions* to help users learn without information overload. Micro-Learning is particularly used in designing mobile learning applications to adapt to the device's limited screen size and interaction possibilities [44].

Mobile Learning

In this thesis, we define mobile learning as learning that takes place with or through mobile devices, independent from the learners' current situation and context .

Mobile Learning Situation

Any situation in which mobile learning takes place in users' everyday lives. Particularly, we consider a *situation* a combination of characteristics that describe the specific setting in which the learning activity is performed, including but not limited to users' physical, temporal, task, social, or technical context (cf. [166]).

Personalized Learning System

A system or application that adapts the learning content, sequence, assessment, or other factors to create a learning experience tailored to the individual end-user.

Seamless Learning

The paradigm of seamless learning can be defined as being able to switch with ease between different (technology-supported) learning contexts or situations (cf. [58]).

Task Resumption Support

(Technological or technology-supported) Features and methods designed to aid users in resuming a task after an interruption occurred.

Technology-enhanced Learning (TEL)

Learning that uses technology as a medium, or that is supported by technology.

Ubiquitous Learning

This concept is a specification of mobile learning, specifically focusing on learning with ubiquitous technologies or technology as a medium to enable ubiquitous interactions..

User Experience (UX)

The term UX encompasses the experience of the user when interacting with a product or service and can include factors such as utility, ease of use, and efficiency.

User-Centered Design (UCD)

A set of processes and methods for creating designs that consider or include end-users (cf. [1]).

List of Figures

List of Tables

Bibliography

[1] Abras, C., Maloney-Krichmar, D., and Preece, J. (2004). User-Centered Design. In Brainbridge, W. S., editor, *Berkshire Encyclopedia of Human-Computer Interaction, Volume 2*, volume 37, pages 445–456. Berkshire, Great Barrington, MA.

[2] Adams, D. (1979). *The Hitchhiker's Guide to the Galaxy*. Pan Books.

[3] Adler, R. F. and Benbunan-Fich, R. (2013). Self-interruptions in discretionary multitasking. *Computers in Human Behavior*, 29(4):1441–1449.

[4] Aljohani, N. R., Davis, H. C., and Loke, S. W. (2012). A comparison between mobile and ubiquitous learning from the perspective of human-computer interaction. *International Journal of Mobile Learning and Organisation*, 6(3–4/2012):218–231, doi:https://doi.org/10.1504/IJMLO.2012.050046.

[5] Altmann, E. M. and Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive science*, 26(1):39–83, doi:https://doi.org/10.1207/s15516709cog2601_2.

[6] Altmann, E. M. and Trafton, J. G. (2004). Task interruption: Resumption lag and the role of cues. Technical report, Michigan State Univ Eest Lansing Dept of Psychology.

[7] Anderson, C., Hübener, I., Seipp, A.-K., Ohly, S., David, K., and Pejovic, V. (2018). A survey of attention management systems in ubiquitous computing environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–27.

[8] Anderson, L. W., Bloom, B. S., et al. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman,.

[9] APA Definition Cue (2021). American Psychology Association Definition of "Cue". Last access: 2019-01-12, `https://dictionary.apa.org/cue`.

[10] APA Definition Learning (2021). American Psychology Association Definition of "Learning". Last access: 2019-01-12, `https://dictionary.apa.org/learning`.

[11] APA Definition Learning and Memory (2021). American Psychology Association Definition of "Learning and Memory". Last access: 2019-01-05, `https://www.apa.org/topics/learning/index.aspx`.

[12] APA Definition Mnemonic (2021). American Psychology Association Definition of "Mnemonic". Last access: 2019-01-12, `https://dictionary.apa.org/mnemonic`.

[13] APA Definition Modality (2021). American Psychology Association Definition of "Modality". Last access: 2019-01-15, https://dictionary.apa.org/modality.

[14] APA Definition Retrieval Cue (2021). American Psychology Association Definition of "Retrieval Cue". Last access: 2019-01-11, https://dictionary.apa.org/retrieval-cue.

[15] Ardissono, L., Bosio, G., and Segnan, M. (2011). A visualization model supporting an efficient context resumption in collaboration environments. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 5–17. Springer.

[16] Ashbrook, D. L. (2010). *Enabling Mobile Microinteractions*. PhD thesis, USA. AAI3414437.

[17] Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes1. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier.

[18] Augereau, O., Fujiyoshi, H., and Kise, K. (2016). Towards an automated estimation of English skill via TOEIC score based on reading analysis. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1285–1290. IEEE, doi:10.1109/ICPR.2016.7899814.

[19] Ausubel, D. P. and Youssef, M. (1965). The effect of spaced repetition on meaningful retention. *The Journal of General Psychology*, 73(1):147–150.

[20] Azuma, M., Coallier, F., and Garbajosa, J. (2003). How to apply the Bloom taxonomy to software engineering. In *Eleventh annual international workshop on software technology and engineering practice*, pages 117–122. IEEE.

[21] Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature reviews neuroscience*, 4(10):829.

[22] Bailey, B. P. and Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior*, 22(4):685–708.

[23] Baillet, S., Mosher, J. C., and Leahy, R. M. (2001). Electromagnetic brain mapping. *IEEE Signal processing magazine*, 18(6):14–30.

[24] Baldauf, M., Khamis, M., Steiner, S., and Thiel, S.-K. (2019). Investigating the User Experience of Smartphone Authentication Schemes-The Role of the Mobile Context. doi:10.24251/HICSS.2019.579.

[25] Bangor, A., Kortum, P., and Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123.

[26] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, doi:10.18637/jss.v067.i01.

[27] Beaudin, J. S., Intille, S. S., Tapia, E. M., Rockinson, R., and Morris, M. E. (2007). Context-sensitive microlearning of foreign language vocabulary on a mobile device. In *European conference on Ambient intelligence*, pages 55–72. Springer.

[28] Berzak, Y., Katz, B., and Levy, R. (2018). Assessing Language Proficiency from Eye Movements in Reading. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1986–1996.

[29] Berzak, Y., Nakamura, C., Flynn, S., and Katz, B. (2017). Predicting Native Language from Gaze. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 541–551.

[30] Bhagavatula, R., Ur, B., Iacovino, K., Kywe, S. M., Cranor, L. F., and Savvides, M. (2015). Biometric authentication on iPhone and Android: Usability, perceptions, and influences on adoption. In *USEC'15: Workshop on Usable Security, 8 February 2015, San Diego, CA: Proceedings*, pages 1–10.

[31] Bialystok, E. (2006). Second-language acquisition and bilingualism at an early age and the impact on early cognitive development. *Encyclopedia on early childhood development*, pages 1–4.

[32] Bialystok, E., Craik, F. I., and Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in cognitive sciences*, 16(4):240–250.

[33] Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K.-R. (2011). Single-trial analysis and classification of ERP components - a tutorial. *NeuroImage*, 56(2):814–825, doi:10.1016/j.neuroimage.2010.06.048.

[34] Bleichner, M. G., Lundbeck, M., Selisky, M., Minow, F., Jäger, M., Emkes, R., Debener, S., and De Vos, M. (2015). Exploring miniaturized EEG electrodes for brain-computer interfaces. An EEG you do not see? *Physiological reports*, 3(4), doi:10.14814/phy2.12362.

[35] Bloom, B. S. et al. (1956). Taxonomy of educational objectives. Vol. 1: Cognitive domain. *New York: McKay*, 20:24.

[36] Borojeni, S. S., Ali, A. E., Heuten, W., and Boll, S. (2016). Peripheral Light Cues for In-Vehicle Task Resumption. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2971485.2971498.

[37] Borovsky, A., Kutas, M., and Elman, J. (2010). Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition*, 116(2):289–296, doi:10.1016/j.cognition.2010.05.004.

[38] Borst, J. P., Buwalda, T. A., van Rijn, H., and Taatgen, N. A. (2013). Avoiding the problem state bottleneck by strategic use of the environment. *Acta Psychologica*, 144(2):373–379.

[39] Borst, J. P., Taatgen, N. A., and van Rijn, H. (2015). What Makes Interruptions Disruptive? A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2971–2980, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2702123.2702156.

[40] Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

[41] Broadbent, J. and Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, 27:1–13.

[42] Brooke, J. (1996). SUS: a "quick and dirty'usability. *Usability evaluation in industry*, page 189.

[43] Brown, A. S., Schilling, H. E., and Hockensmith, M. L. (1999). The negative suggestion effect: Pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology*, 91(4):756.

[44] Bruck, P. A., Motiwalla, L., and Foerster, F. (2012). Mobile Learning with Micro-content: A Framework and Evaluation. *Bled eConference*, 25:527–543.

[45] Brumby, D. P., Cox, A. L., Back, J., and Gould, S. J. (2013). Recovering from an interruption: Investigating speed-accuracy trade-offs in task resumption behavior. *Journal of Experimental Psychology: Applied*, 19(2):95.

[46] Bryan, T., Mathur, S., and Sullivan, K. (1996). The impact of positive mood on learning. *Learning Disability Quarterly*, 19(3):153–162.

[47] Budd, T. W., Barry, R. J., Gordon, E., Rennie, C., and Michie, P. T. (1998). Decrement of the N1 auditory event-related potential with stimulus repetition: habituation vs. refractoriness. *International Journal of Psychophysiology*, 31(1):51–68.

[48] Bulling, A. (2016). Pervasive Attentive User Interfaces. *Computer*, 49(01):94–98, doi:10.1109/MC.2016.32.

[49] Buschke, H. (1984). Cued recall in amnesia. *Journal of Clinical and Experimental Neuropsychology*, 6(4):433–440.

[50] Butler, A. C., Marsh, E. J., Goode, M. K., and Roediger III, H. L. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, 20(7):941–956.

[51] Butler, A. C. and Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & cognition*, 36(3):604–616.

[52] Cades, D. M., Trafton, J. G., and Boehm-Davis, D. A. (2006). Mitigating disruptions: can resuming an interrupted task be trained? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(3):368–371.

[53] Cai, C. J., Guo, P. J., Glass, J., and Miller, R. C. (2014). Wait-Learning: Leveraging Conversational Dead Time for Second Language Education. In *Extended Abstracts Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems*, CHI EA '14, page 2239–2244, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2559206.2581183.

[54] Campbell, J. L., Quincy, C., Osserman, J., and Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, 42(3):294–320.

[55] Campillo, E., Ricarte, J., Ros, L., Nieto, M., and Latorre, J. (2018). Effects of the Visual and Auditory Components of a Brief Mindfulness Intervention on Mood State and on Visual and Auditory Attention and Memory Task Performance. *Current Psychology*, 37(1):357–365.

[56] Cane, J. E., Cauchard, F., and Weger, U. W. (2012). The time-course of recovery from interruption during reading: Eye movement evidence for the role of interruption lag and spatial memory. *The Quarterly journal of experimental psychology*, 65(7):1397–1413.

[57] Carrier, L. M., Rosen, L. D., Cheever, N. A., and Lim, A. F. (2015). Causes, effects, and practicalities of everyday multitasking. *Developmental Review*, 35:64–78.

[58] Chan, T.-W., Roschelle, J., Hsi, S., Kinshuk, Sharples, M., Brown, T., Patton, C., Cherniavsky, J., Pea, R., Norris, C., et al. (2006). One-to-one technology-enhanced learning: An opportunity for global research collaboration. *Research and Practice in Technology Enhanced Learning*, 1(01):3–29.

[59] Chen, C.-H. and Chien, Y.-H. (2007). Effects of RSVP display design on visual performance in accomplishing dual tasks with small screens. *International Journal of Design*, 1(1).

[60] Chen, D. and Vertegaal, R. (2004). Using Mental Load for Managing Interruptions in Physiologically Attentive User Interfaces. In *Extended Abstracts Proceedings of the 2004 CHI Conference on Human Factors in Computing Systems*, CHI EA '04, page 1513–1516, New York, NY, USA. Association for Computing Machinery, doi:10.1145/985921.986103.

[61] Chen, N.-S., Hsieh, S.-W., et al. (2008). Effects of short-term memory and content representation type on mobile language learning. *Language learning & technology*, 12(3):93–113.

[62] Cheng, S., Fan, J., and Dey, A. K. (2018). Smooth gaze: a framework for recovering tasks across devices using eye tracking. *Personal and Ubiquitous Computing*, 22(3):489–501.

[63] Choe, E. K., Lee, B., Kay, M., Pratt, W., and Kientz, J. A. (2015). SleepTight: Low-Burden, Self-Monitoring Technology for Capturing and Reflecting on Sleep Behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 121–132, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2750858.2804266.

[64] Chun, B. A. and Heo, H. J. (2018). The effect of flipped learning on academic performance as an innovative method for overcoming ebbinghaus' forgetting curve. In *Proceedings of the 6th International Conference on Information and Education Technology*, pages 56–60.

[65] Churchill, D. and Hedberg, J. (2008). Learning object design considerations for small-screen handheld devices. *Computers & Education*, 50(3):881–893.

[66] Ciravegna, F. (2019). Creating a never ending background service in Android > 7. *Fabcirablog - Grappling with electronics*. `https://fabcirablog.weebly.com/blog/creating-a-never-ending-background-service-in-android-gt-7`.

[67] Clifford, J. D. and Altmann, E. M. (2004). Managing multiple tasks: Reducing the resumption time of the primary task. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

[68] Clinton, V., Swenseth, M., and Carlson, S. E. (2018). Do Mindful Breathing Exercises Benefit Reading Comprehension? A Brief Report. *Journal of Cognitive Enhancement*, pages 1–6.

[69] Cohen, S. (1980). Aftereffects of stress on human performance and social behavior: a review of research and theory. *Psychological bulletin*, 88(1):82.

[70] Cordova, D. I. and Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of educational psychology*, 88(4):715.

[71] Coulson, S., King, J. W., and Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and cognitive processes*, 13(1):21–58.

[72] Council for Cultural Co-operation. Education Committee. Modern Languages Division, C. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

[73] Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science*, 19(1):51–57.

[74] Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, 14(3):215–235.

[75] Czeisler, C. A. and Gooley, J. (2007). Sleep and circadian rhythms in humans. In *Cold Spring Harbor symposia on quantitative biology*, volume 72, pages 579–597. Cold Spring Harbor Laboratory Press.

[76] Czerwinski, M., Chrisman, S., and Schumacher, B. (1991). Interactive Posters: The effects of warnings and display similarity on interruption in multitasking environments. *SIGCHI Bull.*, 23(4):38–39, doi:10.1145/126729.1056014.

[77] Czerwinski, M., Cutrell, E., and Horvitz, E. (2000). Instant messaging: Effects of relevance and timing. In *People and computers XIV: Proceedings of HCI*, volume 2, pages 71–76.

[78] Davidson, D. J. (2012). Brain Activity During Second Language Processing (ERP). *The Encyclopedia of Applied Linguistics*, doi:10.1002/9781405198431.wbeal0106.

[79] De, M. V. and Debener, S. (2014). Mobile EEG: towards brain activity monitoring during natural action and cognition. *International Journal of Psychology*, 91(1):1–2, doi:10.1016/j.ijpsycho.2013.10.008.

[80] De Witt, C. and Gloerfeld, C. (2018). *Handbuch Mobile Learning*. Springer-Verlag.

[81] Dearman, D. and Truong, K. (2012). *Evaluating the Implicit Acquisition of Second Language Vocabulary Using a Live Wallpaper*, page 1391–1400. Association for Computing Machinery, New York, NY, USA.

[82] Debener, S., Emkes, R., De Vos, M., and Bleichner, M. (2015). Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear. *Scientific reports*, 5:16743, doi:10.1038/srep16743.

[83] Debener, S., Minow, F., Emkes, R., Gandras, K., and De Vos, M. (2012). How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology*, 49(11):1617–1621, doi:10.1111/j.1469-8986.2012.01471.x.

[84] Deci, E. L. and Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. *Perspectives in social psychology*.

[85] Demouy, V., Jones, A., Kan, Q., Kukulska-Hulme, A., and Eardley, A. (2016). Why and How Do Distance Learners Use Mobile Devices for Language Learning? *The EuroCALL Review*, 24(1):10–24.

[86] Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, 79(2):162.

[87] Dingler, T., El Agroudy, P., Le, H. V., Schmidt, A., Niforatos, E., Bexheti, A., and Langheinrich, M. (2016a). Multimedia memory cues for augmenting human memory. *IEEE MultiMedia*, 23(2):4–11.

[88] Dingler, T., Lindsay, J., and Walker, B. N. (2008). Learnabiltiy of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. In *Proceedings of the 14th International Conference on Auditory Display*. International Community for Auditory Display.

[89] Dingler, T. and Pielot, M. (2015). I'll be there for you: Quantifying Attentiveness towards Mobile Messaging. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–5. ACM.

[90] Dingler, T., Rzayev, R., Schwind, V., and Henze, N. (2016b). RSVP on the Go: Implicit Reading Support on Smart Watches through Eye Tracking. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, ISWC '16, page 116–119, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2971763.2971794.

[91] Dingler, T., Weber, D., Pielot, M., Cooper, J., Chang, C.-C., and Henze, N. (2017). Language Learning On-the-Go: Opportune Moments and Design of Mobile Microlearning Sessions. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3098279.3098565.

[92] D'Mello, S., Kopp, K., Bixler, R. E., and Bosch, N. (2016). Attending to Attention: Detecting and Combating Mind Wandering during Computerized Reading. In *Extended Abstracts Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI EA '16, page 1661–1669, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2851581.2892329.

[93] D'Mello, S., Olney, A., Williams, C., and Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, 70(5):377–398.

[94] Draxler, F., Schneegass, C., Lippner, N., and Schmidt, A. (2019a). Exploring Visualizations for Digital Reading Augmentation to Support Grammar Learning. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*, MUM '19, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3365610.3365623.

[95] Draxler, F., Schneegass, C., and Niforatos, E. (2019b). Designing for Task Resumption Support in Mobile Learning. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3338286.3344394.

[96] Draxler, F., Schneegass, C., Safranek, J., and Hußmann, H. (2021). Why did you stop?-Investigating Origins and Effects of Interruptions during Mobile Language Learning. In *Accepted for publication in Proceedings of the Conference Mensch und Computer*. Gesellschaft für Informatik e.V.

[97] Döring, T., Krüger, A., Schmidt, A., and Schöning, J. (2009). Tangible, Embedded, and Reality-Based Interaction. *it - Information Technology*, 51:319–324, doi:10.1524/itit.2009.0558.

[98] Economides, A. A. (2008). Context-aware mobile learning. In *World Summit on Knowledge Society*, pages 213–220. Springer.

[99] Edge, D., Fitchett, S., Whitney, M., and Landay, J. (2012). MemReflex: adaptive flashcards for mobile microlearning. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pages 431–440.

[100] Edge, D., Searle, E., Chiu, K., Zhao, J., and Landay, J. A. (2011). MicroMandarin: Mobile Language Learning in Context. In *Proceedings of the 2011 CHI Conference on*

Human Factors in Computing Systems, CHI '11, page 3169–3178, New York, NY, USA. Association for Computing Machinery, doi:10.1145/1978942.1979413.

[101] EdSurge (2016). Decoding Adaptive. Last access: 2021-08-04, https://www.pearson.com/content/dam/corporate/global/pearson-dot-com/files/innovation/Pearson-Decoding-Adaptive-v5-Web.pdf.

[102] Edwards, M. B. and Gronlund, S. D. (1998). Task interruption and its effects on memory. *Memory*, 6(6):665–687.

[103] Falaki, H., Mahajan, R., Kandula, S., Lymberopoulos, D., Govindan, R., and Estrin, D. (2010). Diversity in smartphone usage. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 179–194.

[104] Fan, S. P., Liberman, Z., Keysar, B., and Kinzler, K. D. (2015). The exposure advantage: Early exposure to a multilingual environment promotes effective communication. *Psychological Science*, 26(7):1090–1097, doi:10.1177%2F0956797615574699.

[105] Fasihuddin, H., Alsolami, S., Alzahrani, S., Alasiri, R., and Sahloli, A. (2018). Smart tutoring system for Arabic sign language using Leap Motion controller. In *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, pages 1–5. IEEE.

[106] Ferreira, D., Goncalves, J., Kostakos, V., Barkhuus, L., and Dey, A. K. (2014). Contextual experience sampling of mobile application micro-usage. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, pages 91–100.

[107] Fischer, G. (1991). Supporting learning on demand with design environments. In *Proceedings of the International Conference on the Learning Sciences*, volume 199, pages 165–172. Citeseer.

[108] Fischer, G. (1998). *Making Learning a Part of Life–Beyond the "Gift Wrapping" Approach of Technology*, pages 435–462. Donat Verlag, Bremen, Germany.

[109] Fischer, G. (2000). Lifelong learning—more than training. *Journal of Interactive Learning Research*, 11(3):265–294.

[110] Fischer, G. (2012). Context-aware systems: the'right'information, at the'right'time, in the'right'place, in the'right'way, to the'right'person. In *Proceedings of the international working conference on advanced visual interfaces*, pages 287–294.

[111] Fischer, J. E., Greenhalgh, C., and Benford, S. (2011). Investigating Episodes of Mobile Phone Activity as Indicators of Opportune Moments to Deliver Notifications. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, page 181–190, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2037373.2037402.

[112] Foehr, U. G. (2006). Media multitasking among American youth: Prevalence, predictors and pairings. *Henry J. Kaiser Family Foundation.*

[113] Fortin, P. E., Huang, Y., and Cooperstock, J. R. (2019). Exploring the Use of Fingerprint Sensor Gestures for Unlock Journaling: A Comparison With Slide-to-X. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3338286.3340135.

[114] Fragoso, V., Gauglitz, S., Zamora, S., Kleban, J., and Turk, M. (2011). TranslatAR: A mobile augmented reality translator. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 497–502. IEEE, doi:10.1109/WACV.2011.5711545.

[115] Franke, J. L., Daniels, J. J., and McFarlane, D. C. (2002). Recovering context after interruption. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 24.

[116] Frey, J., Mühl, C., Lotte, F., and Hachet, M. (2014). Review of the Use of Electroencephalography as an Evaluation Method for Human-Computer Interaction. *Proceedings of the International Conference on Physiological Computing Systems - Volume 1: PhyCS,*, pages 214–223, doi:10.5220/0004708102140223.

[117] Fujii, K. and Rekimoto, J. (2019). SubMe: An Interactive Subtitle System with English Skill Estimation Using Eye Tracking. In *Proceedings of the 10th Augmented Human International Conference 2019*, AH2019, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3311823.3311865.

[118] Gannon, E., He, J., Gao, X., and Chaparro, B. (2016). RSVP Reading on a Smart Watch. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, pages 1130–1134. SAGE Publications Sage CA: Los Angeles, CA, doi:10.1177%2F1541931213601265.

[119] Gardner, R. C. (2001). Integrative motivation and second language acquisition. *Motivation and second language acquisition*, 23(1):1–19.

[120] Geri, N., Winer, A., and Zaks, B. (2017). Challenging the six-minute myth of online video lectures: Can interactivity expand the attention span of learners? *Online Journal of Applied Knowledge Management*, 5(1):101–111.

[121] Gitsaki, C. (1998). Second language acquisition theories: Overview and evaluation. *Journal of communication and international studies*, 4(2):89–98.

[122] Glatz, C., Krupenia, S. S., Bülthoff, H. H., and Chuang, L. L. (2018). Use the Right Sound for the Right Job: Verbal Commands and Auditory Icons for a Task-Management System Favor Different Information Processes in the Brain. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3173574.3174046.

[123] Godwin-Jones, R. (2003). Emerging technologies. *Language Learning & Technology*, 7(2):12–16.

[124] González, V. M. and Mark, G. (2004). *"Constant, Constant, Multi-Tasking Craziness": Managing Multiple Working Spheres*, page 113–120. Association for Computing Machinery, New York, NY, USA.

[125] Graf, S. and Kinshuk (2012). Personalized Learning. In Seel, N. M., editor, *Encyclopedia of the Sciences of Learning*, pages 2592–2594. Springer US, Boston, MA, doi:10.1007/978-1-4419-1428-6_151.

[126] Hagoort, P. (2007). The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):1055–1069, doi:10.1098/rstb.2007.2159.

[127] Hagoort, P., Hald, L., Bastiaansen, M., and Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *science*, 304(5669):438–441, doi:10.1126/science.1095455.

[128] Harbach, M., Von Zezschwitz, E., Fichtner, A., De Luca, A., and Smith, M. (2014). It's a hard lock life: A field study of smartphone (un) locking behavior and risk perception. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 213–230.

[129] Harji, M. B., Woods, P. C., and Alavi, Z. K. (2010). The effect of viewing subtitled videos on vocabulary learning. *Journal of College Teaching & Learning*, 7(9):37–42.

[130] Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage Publications Sage CA: Los Angeles, CA, doi:10.1177%2F154193120605000909.

[131] Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, doi:10.1016/S0166-4115(08)62386-9.

[132] Hassib, M., Pfeiffer, M., Schneegass, S., Rohs, M., and Alt, F. (2017a). *Emotion Actuator: Embodied Emotional Feedback through Electroencephalography and Electrical Muscle Stimulation*, page 6133–6146. Association for Computing Machinery, New York, NY, USA.

[133] Hassib, M., Schneegass, S., Eiglsperger, P., Henze, N., Schmidt, A., and Alt, F. (2017b). *EngageMeter: A System for Implicit Audience Engagement Sensing Using Electroencephalography*, page 5114–5119. Association for Computing Machinery, New York, NY, USA.

[134] Hayati, A. and Mohmedi, F. (2011). The effect of films with and without subtitles on listening comprehension of EFL learners. *British Journal of Educational Technology*, 42(1):181–192, doi:10.1111/j.1467-8535.2009.01004.x.

[135] Heil, C. R., Wu, J. S., Lee, J. J., and Schmidt, T. (2016). A Review of Mobile Language Learning Applications: Trends, Challenges, and Opportunities. In *The EuroCALL Review*, volume 24, pages 32–50. Universitat Politècnica de València, doi:https://doi.org/10.4995/eurocall.2016.6402.

[136] Heimerl, F., Lohmann, S., Lange, S., and Ertl, T. (2014). Word Cloud Explorer: Text Analytics Based on Word Clouds. In *2014 47th Hawaii International Conference on System Sciences*, pages 1833–1842, Waikoloa, HI. IEEE, doi:10.1109/HICSS.2014.231.

[137] Herron, C., Morris, M., Secules, T., and Curtis, L. (1995). A comparison study of the effects of video-based versus text-based instruction in the foreign language classroom. *French Review*, pages 775–795.

[138] Higbee, K. L. (1979). Recent research on visual mnemonics: Historical roots and educational fruits. *Review of Educational Research*, 49(4):611–629.

[139] Higgins, J. M. (1994). *101 creative problem solving techniques: The handbook of new ideas for business*. New Management Publishing Company.

[140] Higgins, J. M. (1996). Innovate or evaporate: creative techniques for strategists. *Long Range Planning*, 29(3):370–380.

[141] Higham, P. A. and Guzel, M. A. (2012). Cueing. In Seel, N. M., editor, *Encyclopedia of the Sciences of Learning*, pages 871–873. Springer US, Boston, MA, USA, doi:10.1007/978-1-4419-1428-6_695.

[142] Hintze, D., Findling, R. D., Scholz, S., and Mayrhofer, R. (2014). Mobile device usage characteristics: The effect of context and form factor on locked and unlocked usage. In *Proceedings of the 12th international conference on advances in mobile computing and multimedia*, pages 105–114.

[143] Hintze, D., Hintze, P., Findling, R. D., and Mayrhofer, R. (2017). A Large-Scale, Long-Term Analysis of Mobile Device Usage Characteristics. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(2), doi:10.1145/3090078.

[144] Ho, C.-Y., Nikolic, M. I., Waters, M. J., and Sarter, N. B. (2004). Not now! Supporting interruption management by indicating the modality and urgency of pending tasks. *Human Factors*, 46(3):399–409.

[145] Ho, J. and Intille, S. S. (2005). Using Context-Aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices. In *Proceedings of the 2005 CHI Conference on Human Factors in Computing Systems*, CHI '05, page 909–918, New York, NY, USA. Association for Computing Machinery, doi:10.1145/1054972.1055100.

[146] Hodgetts, H. M. and Jones, D. M. (2006a). Contextual cues aid recovery from interruption: The role of associative activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5):1120–1132, doi:10.1037/0278-7393.32.5.1120.

[147] Hodgetts, H. M. and Jones, D. M. (2006b). Interruption of the Tower of London task: support for a goal-activation approach. *Journal of Experimental Psychology: General*, 135(1):103, doi:https://doi.org/10.1177%2F154193120304700810.

[148] Hodgetts, H. M., Tremblay, S., Vallières, B. R., and Vachon, F. (2015). Decision support and vulnerability to interruption in a dynamic multitasking environment. *International Journal of Human-Computer Studies*, 79:106–117.

[149] Holcomb, P. J., Coffey, S. A., and Neville, H. J. (1992). Visual and auditory sentence processing: A developmental analysis using event-related brain potentials. *Developmental Neuropsychology*, 8(2-3):203–241, doi:10.1080/87565649209540525.

[150] Holcomb, P. J. and Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and cognitive processes*, 5(4):281–312, doi:10.1080/01690969008407065.

[151] Hopp, P. J., Smith, C., Clegg, B. A., and Heggestad, E. D. (2005). Interruption management: The use of attention-directing tactile cues. *Human Factors*, 47(1):1–11.

[152] Hopp-Levine, P. J., Smith, C., Clegg, B. A., and Heggestad, E. D. (2006). Tactile interruption management: tactile cues as task-switching reminders. *Cognition, Technology & Work*, 8(2):137–145.

[153] Horvitz, E., Apacible, J., and Subramani, M. (2005). Balancing awareness and interruption: Investigation of notification deferral policies. In *International Conference on User Modeling*, pages 433–437. Springer.

[154] Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J., and Yang, J. (2003). Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. In *Proceedings of the 2003 CHI Conference on Human Factors in Computing Systems*, CHI '03, page 257–264, New York, NY, USA. Association for Computing Machinery, doi:10.1145/642611.642657.

[155] Hutt, S., Mills, C., White, S., Donnelly, P. J., and D'Mello, S. K. (2016). The Eyes Have It: Gaze-based Detection of Mind Wandering during Learning with an Intelligent Tutoring System. In *EDM*, pages 86–93.

[156] Iqbal, S. T. and Bailey, B. P. (2005). Investigating the Effectiveness of Mental Workload as a Predictor of Opportune Moments for Interruption. In *Extended Abstracts Proceedings of the 2005 CHI Conference on Human Factors in Computing Systems*, CHI EA '05, page 1489–1492, New York, NY, USA. Association for Computing Machinery, doi:10.1145/1056808.1056948.

[157] Iqbal, S. T. and Bailey, B. P. (2007). Understanding and Developing Models for Detecting and Differentiating Breakpoints during Interactive Tasks. In *Proceedings of the 2007 CHI Conference on Human Factors in Computing Systems*, CHI '07, page 697–706, New York, NY, USA. Association for Computing Machinery, doi:10.1145/1240624.1240732.

[158] Iqbal, S. T. and Horvitz, E. (2007a). Conversations Amidst Computing: A Study of Interruptions and Recovery of Task Activity. In Conati, C., McCoy, K., and Paliouras, G., editors, *User Modeling 2007*, volume 4511, pages 350–354. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-540-73078-1_43.

[159] Iqbal, S. T. and Horvitz, E. (2007b). Disruption and Recovery of Computing Tasks: Field Study, Analysis, and Directions. In *Proceedings of the 2007 CHI Conference on Human Factors in Computing Systems*, CHI '07, page 677–686, New York, NY, USA. Association for Computing Machinery, doi:10.1145/1240624.1240730.

[160] Jasper, H. H. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalogr. Clin. Neurophysiol.*, 10:370–375.

[161] Jensen, M. (2015). Personality Traits, Learning and Academic Achievements. *Journal of Education and Learning*, 4(4):91–118.

[162] Jeuris, S. and Bardram, J. E. (2016). Dedicated workspaces: Faster resumption times and reduced cognitive load in sequential multitasking. *Computers in Human Behavior*, 62:404–414.

[163] Jo, J., Kim, B., and Seo, J. (2015). EyeBookmark: Assisting recovery from interruption during reading. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2963–2966. ACM.

[164] John, M., Smallman, H. S., and Manes, D. I. (2005). Recovery from Interruptions to a Dynamic Monitoring Task: The Beguiling Utility of Instant Replay. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(3):473–477, doi:10.1177/154193120504900355.

[165] Johnson, C. G. and Fuller, U. (2006). Is Bloom's taxonomy appropriate for computer science? In *Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006*, pages 120–123.

[166] Jumisko-Pyykkö, S. and Vainio, T. (2010). Framing the context of use for mobile HCI. *International journal of mobile human computer interaction (IJMHCI)*, 2(4):1–28.

[167] Jung, J., Nour, M., Allman-Farinelli, M., and Kay, J. (2017). Harnessing the" Ambience" of the Mobile-phone Lockscreen for Ultra-lite Logging. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, pages 21–30.

[168] Kang, S. H. (2016). Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences*, 3(1):12–19.

[169] Kar, P., Chattopadhyay, S., and Chakraborty, S. (2020). Gestatten: Estimation of User's Attention in Mobile MOOCs From Eye Gaze and Gaze Gesture Tracking. *Proc. ACM Hum.-Comput. Interact.*, 4(EICS), doi:10.1145/3394974.

[170] Karolus, J., Wozniak, P. W., Chuang, L. L., and Schmidt, A. (2017). *Robust Gaze Features for Enabling Language Proficiency Awareness*, page 2998–3010. Association for Computing Machinery, New York, NY, USA.

[171] Katidioti, I., Borst, J. P., van Vugt, M. K., and Taatgen, N. A. (2016). Interrupt me: External interruptions are less disruptive than self-interruptions. *Computers in Human Behavior*, 63:906–915.

[172] Kay, M., Nelson, G. L., and Hekler, E. B. (2016). Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*,

CHI '16, page 4521–4532, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2858036.2858465.

[173] Kazmer, M. M. (2005). Community-embedded learning. *The Library Quarterly*, 75(2):190–212.

[174] Keller, J. and Suzuki, K. (2004). Learner motivation and e-learning design: A multinationally validated process. *Journal of educational Media*, 29(3):229–239.

[175] Kern, D., Marshall, P., and Schmidt, A. (2010). Gazemarks: Gaze-Based Visual Placeholders to Ease Attention Switching. In *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems*, CHI '10, page 2093–2102, New York, NY, USA. Association for Computing Machinery, doi:10.1145/1753326.1753646.

[176] Khalil, M. K. and Elkhider, I. A. (2016). Applying learning theories and instructional design models for effective instruction. *Advances in physiology education*, 40(2):147–156.

[177] Khamis, M., Alt, F., and Bulling, A. (2018). The Past, Present, and Future of Gaze-Enabled Handheld Mobile Devices: Survey and Lessons Learned. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '18, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3229434.3229452.

[178] Knittel, J. and Dingler, T. (2016). Mining Subtitles for Real-Time Content Generation for Second-Screen Applications. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '16, page 93–103, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2932206.2932217.

[179] Koelle, M., Wolf, K., and Boll, S. (2018). Beyond LED Status Lights-Design Requirements of Privacy Notices for Body-worn Cameras. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 177–187. ACM.

[180] Kolk, H. H., Chwilla, D. J., Van Herten, M., and Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and language*, 85(1):1–36.

[181] Kranz, M., Holleis, P., and Schmidt, A. (2009). Embedded interaction: Interacting with the internet of things. *IEEE internet computing*, 14(2):46–53.

[182] Krashen, S. D. (1981). *Second language acquisition and second language learning*. Pergamon Press Inc.

[183] Krashen, S. D. (1982). Principles and Practice. *Learning*, 46(2):327–69.

[184] Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

[185] Krause, U.-M. and Stark, R. (2006). Vorwissen aktivieren. *Handbuch Lernstrategien*, pages 38–49.

[186] Kreifeldt, J. G. and McCarthy, M. (1981). Interruption as a test of the user-computer interface. *Proceedings of the 17th Annual Conference on Manual Interaction*, pages 655–667.

[187] Krendl, K. A. and Broihier, M. (1992). Student responses to computers: A longitudinal study. *Journal of Educational Computing Research*, 8(2):215–227.

[188] Kroll, J. F. and Dussias, P. E. (2017). The benefits of multilingualism to the personal and professional development of residents of the US. *Foreign Language Annals*, 50(2):248–259.

[189] Kruger, J.-L., Hefer, E., and Matthew, G. (2013). Measuring the impact of subtitles on cognitive load: eye tracking and dynamic audiovisual texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*, pages 62–66. ACM.

[190] Kukulska-Hulme, A. (2006). Mobile language learning now and in the future. In Svensson, P., editor, *In Från vision till praktik: Språkutbildning och Informationsteknik (From vision to practice: language learning and IT)*. Swedish Net University (Nätuniversitetet).

[191] Kukulska-Hulme, A. (2009). Will mobile learning change language learning? *ReCALL*, 21(2):157–165.

[192] Kukulska-Hulme, A. (2012). Language learning defined by time and place: A framework for next generation designs. In Díaz-vera, J. E., editor, *Left to my own devices: Learner autonomy and mobile-assisted language learning*, chapter 1, pages 1–20. Brill.

[193] Kuo, B. Y.-L., Hentrich, T., Good, B. M. ., and Wilkinson, M. D. (2007). Tag clouds for summarizing web search results. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, page 1203, Banff, Alberta, Canada. ACM Press, doi:10.1145/1242572.1242766.

[194] Kutas, M. and Dale, A. (1997). Electrical and magnetic readings of mental functions. *Cognitive neuroscience*, pages 197–242.

[195] Kutas, M. and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in cognitive sciences*, 4(12):463–470, doi:10.1016/S1364-6613(00)01560-6.

[196] Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, doi:10.1126/science.7350657.

[197] Kutas, M. and Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & cognition*, 11(5):539–550.

[198] Kutas, M., Van Petten, C., and Besson, M. (1988). Event-related potential asymmetries during the reading of sentences. *Electroencephalography and clinical neurophysiology*, 69(3):218–233.

[199] Kuznetsov, S., Dey, A. K., and Hudson, S. E. (2009). The effectiveness of haptic cues as an assistive technology for human memory. In *International Conference on Pervasive Computing*, pages 168–175. Springer.

[200] Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26, doi:10.18637/jss.v082.i13.

[201] Lai, C. (2013). A framework for developing self-directed technology use for language learning. *Language Learning & Technology*, 17(2):100–122.

[202] Lai, C. and Gu, M. (2011). Self-regulated out-of-class language learning with technology. *Computer assisted language learning*, 24(4):317–335.

[203] Lai, C. and Zheng, D. (2018). Self-directed use of mobile devices for language learning beyond the classroom. *ReCALL*, 30(3):299–318.

[204] Latorella, K. A. (1998). Effects of modality on interrupted flight deck performance: Implications for data link. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 42, pages 87–91. SAGE Publications Sage CA: Los Angeles, CA.

[205] Laufer, B. and Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language learning*, 54(3):399–436.

[206] Lazar, J., Feng, J. H., and Hochheiser, H. (2017). *Research methods in human-computer interaction*. Morgan Kaufmann.

[207] Ledoux, K. and Gordon, P. C. (2006). Interruption-similarity effects during discourse processing. *Memory*, 14(7):789–803.

[208] Legge, G. E., Mansfield, J. S., and Chung, S. T. (2001). Psychophysics of reading: XX. Linking letter recognition to reading speed in central and peripheral vision. *Vision research*, 41(6):725–743, doi:10.1016/S0042-6989(00)00295-9.

[209] Leiva, L., Böhmer, M., Gehring, S., and Krüger, A. (2012). Back to the App: The Costs of Mobile Application Interruptions. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '12, page 291–294, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2371574.2371617.

[210] Lennon, C. and Burdick, H. (2004). The lexile framework as an approach for reading measurement and success. *electronic publication on www.lexile.com*.

[211] Liberman, Z., Woodward, A. L., Keysar, B., and Kinzler, K. D. (2017). Exposure to multiple languages enhances communication skills in infancy. *Developmental science*, 20(1):e12420, doi:10.1111/desc.12420.

[212] Lindblom, J. and Gündert, J. (2017). Managing mediated interruptions in manufacturing: Selected strategies used for coping with cognitive load. In *Advances in neuroergonomics and cognitive engineering*, pages 389–403. Springer.

[213] Liu, X., Tan, P.-N., Liu, L., and Simske, S. J. (2017). Automated classification of EEG signals for predicting students' cognitive state during learning. In *Proceedings of the International Conference on Web Intelligence*, pages 442–450. ACM.

[214] Liu, Y., Jia, Y., Pan, W., and Pfaff, M. S. (2014). Supporting task resumption using visual feedback. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 767–777. ACM.

[215] Lonsdale, P., Baber, C., Sharples, M., Byrne, W., Arvanitis, T. N., Brundell, P., and Beale, R. (2005). Context awareness for MOBIlearn: creating an engaging learning experience in an art museum. *Mobile learning anytime everywhere*, 115.

[216] Looi, C.-K., Seow, P., Zhang, B., So, H.-J., Chen, W., and Wong, L.-H. (2010). Leveraging mobile technology for sustainable seamless learning: a research agenda. *British journal of educational technology*, 41(2):154–169.

[217] Lotte, F. (2014). A tutorial on EEG signal-processing techniques for mental-state recognition in brain–computer interfaces. In *Guide to Brain-Computer Music Interfacing*, pages 133–161. Springer.

[218] Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., and Yger, F. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, doi:10.1088/1741-2552/aab2f2.

[219] Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., and Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1, doi:10.1088/1741-2560/4/2/R01.

[220] Luzhnica, G., Veas, E., and Pammer, V. (2016). Skin Reading: Encoding Text in a 6-Channel Haptic Display. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, ISWC '16, page 148–155, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2971763.2971769.

[221] Ma, Q., Wang, S., Liu, J., and Li, N. (2018). InteractiveSubtitle: Subtitle Interaction for Language Learning. In *Proceedings of the Sixth International Symposium of Chinese CHI*, ChineseCHI '18, page 116–119, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3202667.3202685.

[222] Machanick, P. (2000). Experience of applying Bloom's Taxonomy in three courses. In *Proc. Southern African Computer Lecturers' Association Conference*, pages 135–144.

[223] Mahfouz, A., Muslukhov, I., and Beznosov, K. (2016). Android users in the wild: Their authentication and usage behavior. *Pervasive and Mobile Computing*, 32:50–61.

[224] Malkin, N., Harbach, M., De Luca, A., and Egelman, S. (2017). The anatomy of smartphone unlocking: Why and how android users around the world lock their phones. *GetMobile: Mobile Computing and Communications*, 20(3):42–46.

[225] Mancero, G., Wong, B., and Loomes, M. (2009). Radio dispatchers' interruption recovery strategies. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7*, pages 113–120. ACM.

[226] Mao, J.-Y., Vredenburg, K., Smith, P. W., and Carey, T. (2005). The state of user-centered design practice. *Communications of the ACM*, 48(3):105–109.

[227] Mariakakis, A., Goel, M., Aumi, M. T. I., Patel, S. N., and Wobbrock, J. O. (2015). SwitchBack: Using focus and saccade tracking to guide users' attention for mobile task resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2953–2962. ACM, doi:https://doi.org/10.1145/2702123.2702539.

[228] Marsh, E. J., Agarwal, P. K., and Roediger III, H. L. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied*, 15(1):1.

[229] Mayer, R. E. (2002). Multimedia learning. In *Psychology of learning and motivation*, volume 41, pages 85–139. Elsevier.

[230] Mayer, R. E. (2005). Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, 43.

[231] Mayer, R. E., Bove, W., Bryman, A., Mars, R., and Tapangco, L. (1996). When less is more: Meaningful learning from visual and verbal summaries of science textbook lessons. *Journal of educational psychology*, 88(1):64.

[232] McDaniel, M. A., Einstein, G. O., Graham, T., and Rall, E. (2004). Delaying execution of intentions: Overcoming the costs of interruptions. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 18(5):533–547.

[233] McLean, R. and Gregg, L. (1967). Effects of induced chunking on temporal aspects of serial recitation. *Journal of experimental psychology*, 74(4p1):455.

[234] Meara, P. (1990). Some notes on the Eurocentres vocabulary tests. *Foreign language comprehension and production*, pages 103–113.

[235] Meara, P. and Jones, G. (1988). Vocabulary size as a placement indicator. *In Grunwell, Pamela, Ed. Applied Lignguistics in Society. 20th Annual Meeting of the British Association for Applied Linguistics.*

[236] Michalko, M. (2014). *Thinkpak: A Brainstorming Card Deck;[a Creative-thinking Toolbox].* Ten Speed Press.

[237] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

[238] Miyata, Y. and Norman, D. A. (1986). Psychological issues in support of multiple activities. In Normal, D. A. & Drapwer, S. W., editor, *User centered system design: New perspectives on human-computer interaction*, chapter 13, pages 265–284. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

[239] Monk, C. A., Trafton, J. G., and Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied*, 14(4):299.

[240] Mori, S., Nishida, H., and Yamada, H. (1999). *Optical character recognition*. John Wiley & Sons, Inc.

[241] Morris, D., Ringel Morris, M., and Venolia, G. (2008). SearchBar: A Search-Centric Web History for Task Resumption and Information Re-Finding. In *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems*, CHI '08, page 1207–1216, New York, NY, USA. Association for Computing Machinery, doi:10.1145/1357054.1357242.

[242] Mrazek, M. D., Smallwood, J., and Schooler, J. W. (2012). Mindfulness and mind-wandering: Finding convergence through opposing constructs. *Emotion*, 12(3):442–448, doi:10.1037/a0026678.

[243] Mueller, F. F., Lopes, P., Strohmeier, P., Ju, W., Seim, C., Weigel, M., Nanayakkara, S., Obrist, M., Li, Z., Delfa, J., Nishida, J., Gerber, E. M., Svanaes, D., Grudin, J., Greuter, S., Kunze, K., Erickson, T., Greenspan, S., Inami, M., Marshall, J., Reiterer, H., Wolf, K., Meyer, J., Schiphorst, T., Wang, D., and Maes, P. (2020). *Next Steps for Human-Computer Integration*, page 1–15. Association for Computing Machinery, New York, NY, USA.

[244] Mueller, J. L. (2005). Electrophysiological correlates of second language processing. *Second Language Research*, 21(2):152–174, doi:10.1191%2F0267658305sr256oa.

[245] Mullen, T. R., Kothe, C. A., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., Jung, T.-P., and Cauwenberghs, G. (2015). Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE Transactions on Biomedical Engineering*, 62(11):2553–2567, doi:10.1109/TBME.2015.2481482.

[246] Näätänen, R. and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24(4):375–425.

[247] Neuman, S. B. and Koskinen, P. (1992). Captioned television as comprehensible input: Effects of incidental word learning from context for language minority students. *Reading research quarterly*, pages 95–106, doi:10.2307/747835.

[248] Niforatos, E., Laporte, M., Bexheti, A., and Langheinrich, M. (2018). Augmenting Memory Recall in Work Meetings: Establishing a Quantifiable Baseline. In *Proceedings of the 9th Augmented Human International Conference*, page 4. ACM.

[249] Nowell, L. S., Norris, J. M., White, D. E., and Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, 16(1):1609406917733847.

[250] Okoshi, T., Nakazawa, J., and Tokuda, H. (2014). Attelia: Sensing User's Attention Status on Smart Phones. In *Proceedings of the 2014 ACM International Joint*

Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, UbiComp '14 Adjunct, page 139–142, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2638728.2638802.

[251] Okoshi, T., Ramos, J., Nozaki, H., Nakazawa, J., Dey, A. K., and Tokuda, H. (2015). Reducing Users' Perceived Mental Effort Due to Interruptive Notifications in Multi-Device Mobile Environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 475–486, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2750858.2807517.

[252] Okundaye, O., Quek, F., Sargunam, S. P., Suhail, M., and Das, R. (2017). Facilitating Context Switching Through Tangible Artifacts. In *Extended Abstracts Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI EA '17, page 1940–1946, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3027063.3053116.

[253] O'Malley, C., Vavoula, G., Glew, J., Taylor, J., Sharples, M., Lefrere, P., Lonsdale, P., Naismith, L., and Waycott, J. (2005). Guidelines for learning/teaching/tutoring in a mobile environment. Public deliverable from the MOBILearn project (D.4.1).

[254] Oulasvirta, A. (2005). *Interrupted Cognition and Design for Non-Disruptiveness: The Skilled Memory Approach*, page 1124–1125. Association for Computing Machinery, New York, NY, USA.

[255] Oulasvirta, A. and Saariluoma, P. (2006). Surviving task interruptions: Investigating the implications of long-term working memory theory. *International Journal of Human-Computer Studies*, 64(10):941–961, doi:10.1016/j.ijhcs.2006.04.006.

[256] Oulasvirta, A., Tamminen, S., Roto, V., and Kuorelahti, J. (2005). Interaction in 4-Second Bursts: The Fragmented Nature of Attentional Resources in Mobile HCI. In *Proceedings of the 2005 CHI Conference on Human Factors in Computing Systems*, CHI '05, page 919–928, New York, NY, USA. Association for Computing Machinery, doi:10.1145/1054972.1055101.

[257] Page, T. (2013). Usability of text input interfaces in smartphones. *Journal of Design Research*, 11(1):39–56.

[258] Parnin, C. and DeLine, R. (2010). *Evaluating Cues for Resuming Interrupted Programming Tasks*, page 93–102. Association for Computing Machinery, New York, NY, USA.

[259] Pearson, P. D., Hiebert, E. H., and Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading research quarterly*, 42(2):282–296, doi:10.1598/RRQ.42.2.4.

[260] Pham, X. L., Pham, T., Nguyen, Q. M., Nguyen, T. H., and Cao, T. T. H. (2018). Chatbot as an intelligent personal assistant for mobile language learning. In *Proceedings of the 2018 2nd International Conference on Education and E-Learning*, pages 16–21.

[261] Pielot, M., Dingler, T., Pedro, J. S., and Oliver, N. (2015). When Attention is Not Scarce - Detecting Boredom from Mobile Phone Usage. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 825–836, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2750858.2804252.

[262] Pielot, M. and Oliveira, R. d. (2013). Peripheral vibro-tactile displays. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, pages 1–10.

[263] Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of educational Psychology*, 95(4):667, doi:10.1037/0022-0663.95.4.667.

[264] Poguntke, R., Schneegass, C., Van der Vekens, L., Rzayev, R., Auda, J., Schneegass, S., and Schmidt, A. (2020). NotiModes: An Investigation of Notification Delay Modes and Their Effects on Smartphone Users. In *Proceedings of the Conference on Mensch Und Computer*, MuC '20, page 415–419, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3404983.3410006.

[265] Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, 118(10):2128–2148, doi:10.1016/j.clinph.2007.04.019.

[266] Prange, S., Mecke, L., Nguyen, A., Khamis, M., and Alt, F. (2020). Don't Use Fingerprint, it's Raining! How People Use and Perceive Context-Aware Selection of Mobile Authentication. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–5.

[267] Pritchard, W. S. (1981). Psychophysiology of P300. *Psychological bulletin*, 89(3):506.

[268] Proakis, J. G. (2001). *Digital signal processing: principles algorithms and applications*. Pearson Education India.

[269] Qiu, L., De Luca, A., Muslukhov, I., and Beznosov, K. (2019). *Towards Understanding the Link Between Age and Smartphone Authentication*, page 1–10. Association for Computing Machinery, New York, NY, USA.

[270] Quinn, E. and Nation, I. S. P. (1974). *Speed reading: A course for learners of English*. Oxford University Press.

[271] Quinn, E., Nation, I. S. P., and Millett, S. (2007). Asian and Pacific speed readings for ESL learners. *English Language Institute Occasional Publication*, 24.

[272] R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[273] Ratwani, R. M., Andrews, A. E., McCurry, M., Trafton, J. G., and Peterson, M. S. (2007). Using Peripheral Processing and Spatial Memory to Facilitate Task Resumption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 51(4):244–248, doi:10.1177/154193120705100421.

[274] Ratwani, R. M. and Trafton, J. G. (2008). Spatial memory guides task resumption. *Visual Cognition*, 16(8):1001–1010.

[275] Rayner, K. and McConkie, G. W. (1976). What guides a reader's eye movements? *Vision research*, 16(8):829–837, doi:10.1016/0042-6989(76)90143-7.

[276] Rayner, K., Slattery, T. J., Drieghe, D., and Liversedge, S. P. (2011). Eye movements and word skipping during reading: effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):514, doi:10.1037/a0020990.

[277] Remijn, G., Hasuo, E., Fujihira, H., and Morimoto, S. (2014). An introduction to the measurement of auditory event-related potentials (ERPs). *Acoustical Science and Technology*, 35:229–242, doi:10.1250/ast.35.229.

[278] Richards, J. C. (2015). The changing face of language learning: Learning beyond the classroom. *RELC Journal*, 46(1):5–22, doi:10.1177/0033688214561621.

[279] Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.

[280] Roediger III, H. L. and Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, 15(1):20–27.

[281] Roediger III, H. L. and Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):1155.

[282] Rohloff, T., Sauer, D., and Meinel, C. (2019). On the Acceptance and Usefulness of Personalized Learning Objectives in MOOCs. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*, L@S '19, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3330430.3333624.

[283] Rohloff, T., Sauer, D., and Meinel, C. (2020). Students' Achievement of Personalized Learning Objectives in MOOCs. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, L@S '20, page 147–156, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3386527.3405918.

[284] Rose, E. (2010). Continuous partial attention: Reconsidering the role of online learning in the age of interruption. *Educational Technology*, 50(4):41–46.

[285] Rosell-Aguilar, F. (2018). Autonomous language learning through a mobile application: a user evaluation of the busuu app. *Computer Assisted Language Learning*, pages 1–28.

[286] Rubin, G. S. and Turano, K. (1992). Reading without saccadic eye movements. *Vision research*, 32(5):895–902, doi:10.1016/0042-6989(92)90032-E.

[287] Rule, A. and Hollan, J. (2016). Thinking in 4D: Preserving and Sharing Mental Context Across Time. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, pages 389–392. ACM.

[288] Sadeghian Borojeni, S., Löcken, A., and Müller, H. (2014). Using Peripheral Cues to Support Task Resumption. In *Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 1–4. ACM, doi:https://doi.org/10.1145/2667239.2667290.

[289] Sahami, A., Holleis, P., Schmidt, A., and Häkkilä, J. (2008). Rich tactile output on mobile devices. In *European Conference on Ambient Intelligence*, pages 210–221. Springer.

[290] Sahami Shirazi, A., Funk, M., Pfleiderer, F., Glück, H., and Schmidt, A. (2012). Mediabrain: Annotating videos based on brain-computer interaction. *Mensch & Computer 2012: interaktiv informiert–allgegenwärtig und allumfassend!?*

[291] Sakunkoo, N. and Sakunkoo, P. (2013). Gliflix: Using movie subtitles for language learning. In *Proceedings of the 26th Symposium on User Interface Software and Technology*. ACM.

[292] Sanches, C. L., Kise, K., and Augereau, O. (2017). Japanese Reading Objective Understanding Estimation by Eye Gaze Analysis. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, page 121–124, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3123024.3123092.

[293] Sasangohar, F., Scott, S. D., and Cummings, M. (2014). Supervisory-level interruption recovery in time-critical control tasks. *Applied Ergonomics*, 45(4):1148–1156, doi:10.1016/j.apergo.2014.02.005.

[294] Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American psychologist*, 54(3):182.

[295] Schmidt, A. (2007). Eingebettete Interaktion — Symbiose von Mensch und Information. In Mattern, F., editor, *Die Informatisierung des Alltags: Leben in smarten Umgebungen*, pages 77–101. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-540-71455-2_6.

[296] Schmidt, A., Beigl, M., and Gellersen, H.-W. (1999). There is more to context than location. *Computers & Graphics*, 23(6):893–901.

[297] Schmidt, C., Collette, F., Cajochen, C., and Peigneux, P. (2007). A time to think: circadian rhythms in human cognition. *Cognitive neuropsychology*, 24(7):755–789.

[298] Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. *Attention and awareness in foreign language learning*, 9:1–63.

[299] Schneegass, C. and Draxler, F. (2020). Cognitive Biases and their Effect on Mobile Learning: The Example of the Continued Influence Bias and Negative Suggestion Effect. In *In Workshop on Detection and Design for Cognitive Biases in People and Computing Systems at the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*.

[300] Schneegass, C. and Draxler, F. (2021). Designing Task Resumption Cues for Interruptions in Mobile Learning Scenarios. In Dingler, T. and Niforatos, E., editors, *Technology-Augmented Perception and Cognition*, pages 125–181. Springer International Publishing, Cham, doi:10.1007/978-3-030-30457-7_5.

[301] Schneegass, C., Füseschi, V., Konevych, V., and Draxler, F. (2022). Investigating the Use of Task Resumption Cues to Support Learning in Interruption-Prone Environments. In *In Multimodal Technol. Interact. 6, 2.*

[302] Schneegass, C., Kosch, T., Baumann, A., Rusu, M., Hassib, M., and Hußmann, H. (2020). BrainCoDe: Electroencephalography-Based Comprehension Detection during Reading and Listening. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3313831.3376707.

[303] Schneegass, C., Kosch, T., Schmidt, A., and Hußmann, H. (2019). Investigating the Potential of EEG for Implicit Detection of Unknown Words for Foreign Language Learning. In *IFIP Conference on Human-Computer Interaction*, pages 293–313. Springer.

[304] Schneegass, C., Sigethy, S., Eiband, M., and Buschek, D. (2021a). Comparing Concepts for Embedding Second Language Vocabulary Acquisition into Everyday Smartphone Interactions. In *Accepted for publication in Proceedings of the Conference Mensch und Computer*. Gesellschaft für Informatik e.V.

[305] Schneegass, C., Terzimehić, N., Nettah, M., and Schneegass, S. (2018). Informing the Design of User-Adaptive Mobile Language Learning Applications. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, MUM 2018, page 233–238, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3282894.3282926.

[306] Schneegass, C., Wojcicki, Y., and Niforatos, E. (2021b). Design for Long-term Memory Augmentation in Personal Knowledge Management Applications. In *12th Augmented Human International Conference*, pages 1–5.

[307] Schoedel, R., Pargent, F., Au, Q., Völkel, S. T., Schuwerk, T., Bühner, M., and Stachl, C. (2020). To Challenge the Morning Lark and the Night Owl: Using Smartphone Sensing Data to Investigate Day–Night Behaviour Patterns. *European Journal of Personality.*

[308] Scott, S. D., Mercier, S., Cummings, M. L., and Wang, E. (2006). Assisting interruption recovery in supervisory control of multiple UAVs. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 699–703. SAGE Publications Sage CA: Los Angeles, CA, doi:10.1177/154193120605000518.

[309] Shekhar, S., Singal, D., Singh, H., Kedia, M., and Shetty, A. (2017). Show and Recall: Learning What Makes Videos Memorable. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2730–2739.

[310] Shirazi, A. S., Hassib, M., Henze, N., Schmidt, A., and Kunze, K. (2014). What's on your mind?: mental task awareness using single electrode brain computer interfaces.

In *Proceedings of the 5th Augmented Human International Conference*, page 45. ACM, doi:10.1145/2582051.2582096.

[311] Skalka, J. and Drlík, M. (2018). Educational model for improving programming skills based on conceptual microlearning framework. In *International Conference on Interactive Collaborative Learning*, pages 923–934. Springer.

[312] Smith, C., Clegg, B. A., Heggestad, E. D., and Hopp-Levine, P. J. (2009). Interruption management: A comparison of auditory and tactile cues for both alerting and orienting. *International Journal of Human-Computer Studies*, 67(9):777–786.

[313] Smith, G. F. (1998). Idea-generation techniques: A formulary of active ingredients. *The Journal of Creative Behavior*, 32(2):107–134.

[314] Sparrow, B., Liu, J., and Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *science*, 333(6043):776–778.

[315] Speier, C., Vessey, I., and Valacich, J. S. (2003). The effects of interruptions, task complexity, and information presentation on computer-supported decision-making performance. *Decision Sciences*, 34(4):771–797.

[316] Stasinopoulos, M. D., editor (2017). *Flexible regression and smoothing: using GAMLSS in R*. CRC Press/Taylor & Francis Group, Boca Raton.

[317] Steel, C. (2012). Fitting learning into life: Language students' perspectives on benefits of using mobile apps. In *ascilite*, pages 875–880.

[318] Steil, J., Müller, P., Sugano, Y., and Bulling, A. (2018). Forecasting user attention during everyday mobile interactions using device-integrated and wearable sensors. In *Proceedings of the 20th international conference on human-computer interaction with mobile devices and services*, pages 1–13. ACM.

[319] Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y., Dong, J., Duffy, V. G., Fang, X., Fidopiastis, C., Fragomeni, G., Fu, L. P., et al. (2019). Seven HCI grand challenges. *International Journal of Human–Computer Interaction*, 35(14):1229–1269.

[320] Stokes, T. A., Welk, A. K., Zielinska, O. A., and Gillan, D. J. (2017). The Oddball Effect and Inattentional Blindness: How Unexpected Events Influence Our Perceptions Of Time. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pages 1753–1757. SAGE Publications Sage CA: Los Angeles, CA.

[321] Sur, S. and Sinha, V. (2009). Event-related potential: An overview. *Industrial psychiatry journal*, 18(1):70, doi:10.4103/0972-6748.57865.

[322] Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schölkopf, B., and Gomez-Rodriguez, M. (2019). Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences*, 116(10):3988–3993.

[323] Tatsuno, S. (1990). *Created in Japan: From imitators to world-class innovators*. Ballinger Publishing Company.

[324] Technische Universität Dresden, I. (2014). Toolbox "Umgang mit Vorwissen in der Lehre". Last access: 2019-03-06, https://bildungsportal.sachsen.de/opal/auth/RepositoryEntry/6931742725/CourseNode/89718347352042/Texte_Toolbox.pdf.

[325] Terzimehić, N., Häuslschmid, R., Hussmann, H., and schraefel, m. (2019). A Review & Analysis of Mindfulness Research in HCI: Framing Current Lines of Research and Future Opportunities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3290605.3300687.

[326] Thaler, L., Schütz, A. C., Goodale, M. A., and Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, 76:31–42, doi:10.1016/j.visres.2012.10.012.

[327] Thompson, C. P. (1982). Memory for unique personal events: The roommate study. *Memory & Cognition*, 10(4):324–332.

[328] Toreini, P., Langner, M., and Maedche, A. (2018). Use of Attentive Information Dashboards to Support Task Resumption in Working Environments. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA '18, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3204493.3208348.

[329] Trafton, J. G., Altmann, E. M., Brock, D. P., and Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, 58(5):583–603, doi:https://doi.org/10.1016/S1071-5819(03)00023-5.

[330] Truong, K. N., Shihipar, T., and Wigdor, D. J. (2014). Slide to X: Unlocking the Potential of Smartphone Unlocking. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems*, CHI '14, page 3635–3644, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2556288.2557044.

[331] Trusty, A. and Truong, K. N. (2011). *Augmenting the Web for Second Language Vocabulary Learning*, page 3179–3188. Association for Computing Machinery, New York, NY, USA.

[332] Tulving, E. and Schacter, D. L. (1990). Priming and human memory systems. *Science*, 247(4940):301–306.

[333] Ullman, M. T. (2001). The declarative/procedural model of lexicon and grammar. *Journal of psycholinguistic research*, 30(1):37–69.

[334] Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, 92(1-2):231–270.

[335] Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Growdon, J. H., Koroshetz, W. J., and Pinker, S. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of cognitive neuroscience*, 9(2):266–276.

[336] van Berkel, N., Ferreira, D., and Kostakos, V. (2017). The Experience Sampling Method on Mobile Devices. *ACM Comput. Surv.*, 50(6), doi:10.1145/3123988.

[337] Van Cauwenberge, A., Schaap, G., and Van Roy, R. (2014). "TV no longer commands our full attention": Effects of second-screen viewing and task relevance on cognitive load and learning from news. *Computers in Human Behavior*, 38:100–109, doi:10.1016/j.chb.2014.05.021.

[338] Van Hell, J. G. and Tokowicz, N. (2010). Event-related brain potentials and second language learning: Syntactic processing in late L2 learners at different L2 proficiency levels. *Second Language Research*, 26(1):43–74, doi:10.1177/0267658309337637.

[339] Vanderplank, R. (1994). Resolving inherent conflicts: Autonomous language learning from popular broadcast television. *Barriers and bridges: Media technology in language learning*, pages 119–133.

[340] Verleger, M. and Pembridge, J. (2018). A pilot study integrating an AI-driven chatbot in an introductory programming course. In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–4. IEEE.

[341] Viberg, O. and Andersson, A. (2019). The role of self-regulation and structuration in mobile learning. *International Journal of Mobile and Blended Learning (IJMBL)*, 11(4):42–58.

[342] von Zezschwitz, E., Dunphy, P., and De Luca, A. (2013). Patterns in the Wild: A Field Study of the Usability of Pattern and Pin-Based Authentication on Mobile Devices. In *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '13, page 261–270, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2493190.2493231.

[343] Vourvopoulos, A., Niforatos, E., and Giannakos, M. (2019). EEGlass: An EEG-Eyeware Prototype for Ubiquitous Brain-Computer Interaction. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, page 647–652, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3341162.3348383.

[344] Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J. N., and Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, 25(1):58–76, doi:10.3758/s13423-017-1323-7.

[345] Wang, C., Wang, Y., Chen, Y., Liu, H., and Liu, J. (2020). User authentication on mobile devices: Approaches, threats and trends. *Computer Networks*, 170:107118.

[346] Weiser, M. (1991). The Computer for the 21 st Century. *Scientific american*, 265(3):94–105.

[347] Wise, R. J. S. and Brownsett, S. L. E. (2009). The Contribution of the Parietal Lobes to Speaking and Writing. *Cerebral Cortex*, 20(3):517–523, doi:10.1093/cercor/bhp120.

[348] Wobbrock, J. O. and Kientz, J. A. (2016). Research contributions in human-computer interaction. *interactions*, 23(3):38–44.

[349] Woelki, D., Oulasvirta, A., Kiefer, J., and Lischke, R. (2008). Practice effects on interruption tolerance in algebraic problem-solving. *Proceedings of the 30th Annual Cognitive Science Conference (CogSci2008)*.

[350] Wood, E. and Bulling, A. (2014). EyeTab: Model-Based Gaze Estimation on Unmodified Tablet Computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '14, page 207–210, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2578153.2578185.

[351] Woods, D. L. (1995). The component structure of the N 1 wave of the human auditory evoked potential. *Electroencephalography and Clinical Neurophysiology-Supplements Only*, 44:102–109.

[352] Xiao, X. and Wang, J. (2017). Undertanding and Detecting Divided Attention in Mobile MOOC Learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 2411–2415, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3025453.3025552.

[353] Yamada, Y., Irie, K., Gushima, K., Ishizawa, F., Sada, M. A., and Nakajima, T. (2017). HoloMoL: Human Memory Augmentation with Mixed-Reality Technologies. In *Proceedings of the 21st International Academic Mindtrek Conference*, Academic-Mindtrek '17, page 235–238, New York, NY, USA. Association for Computing Machinery, doi:10.1145/3131085.3131097.

[354] Yan, T., Chu, D., Ganesan, D., Kansal, A., and Liu, J. (2012). Fast App Launching for Mobile Devices Using Predictive User Context. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, page 113–126, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2307636.2307648.

[355] Yang, F., Heeman, P. A., and Kun, A. L. (2011). An Investigation of Interruptions and Resumptions in Multi-Tasking Dialogues. *Computational Linguistics*, 37(1):75–104, doi:10.1162/coli_a_00036.

[356] Yatid, M. and Takatsuka, M. (2012). Understanding the Effectiveness of Visual Cues to Support Categorical Notification. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, OzCHI '12, page 661–664, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2414536.2414636.

[357] Yerkes, R. M. and Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology*, 18(5):459–482.

[358] Yeung, W. and Li, S. Y. (2016). Prototyping the Machine-Human Dialogues in a Smartphone Voice Call Application With Task Resumption Support. In *Extended Abstracts*

Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI EA '16, page 1788–1793, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2851581.2892464.

[359] Yuksel, B. F., Oleson, K. B., Harrison, L., Peck, E. M., Afergan, D., Chang, R., and Jacob, R. J. (2016). Learn Piano with BACh: An Adaptive Learning Interface That Adjusts Task Difficulty Based on Brain State. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5372–5384, New York, NY, USA. Association for Computing Machinery, doi:10.1145/2858036.2858388.

[360] Zanón, N. T. (2006). Using subtitles to enhance foreign language learning. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras*, (6):4.

[361] Zhang, X., Kosmyna, N., Maes, P., and Rekimoto, J. (2018). Investigating Bodily Responses to Unknown Words: a Focus on Facial Expressions and EEG. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5211–5215. IEEE.

[362] Zhang, X., Pina, L. R., and Fogarty, J. (2016). *Examining Unlock Journaling with Diaries and Reminders for In Situ Self-Report in Health and Wellness*, page 5658–5664. Association for Computing Machinery, New York, NY, USA.

[363] Zydney, J. M. and Warner, Z. (2016). Mobile apps for science learning: Review of research. *Computers & Education*, 94:1–17.

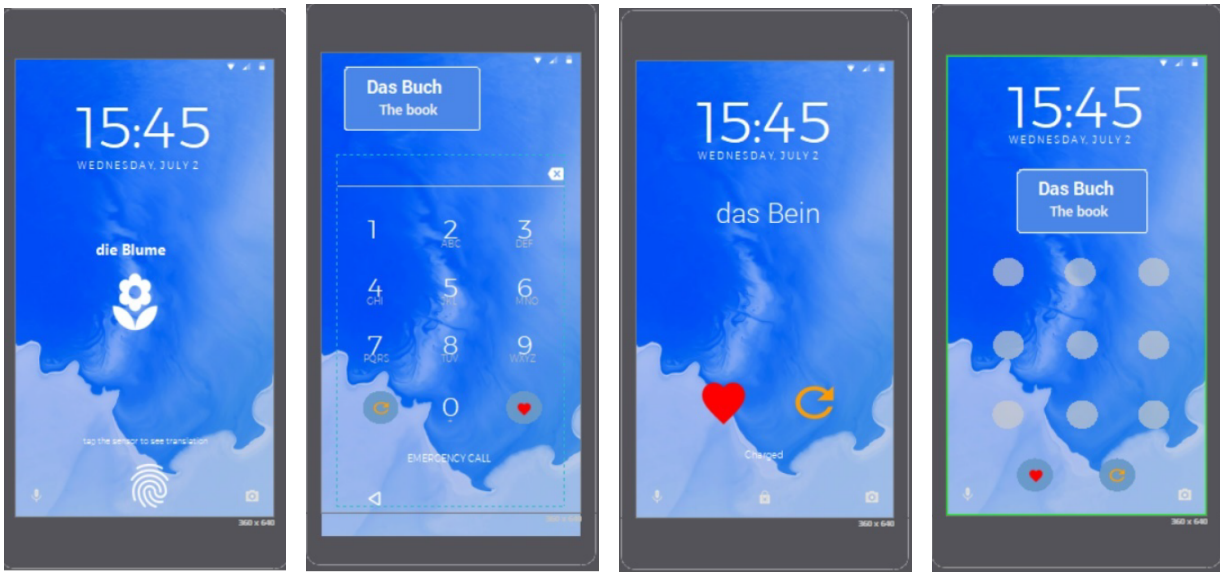
Appendix

A

Additional Materials

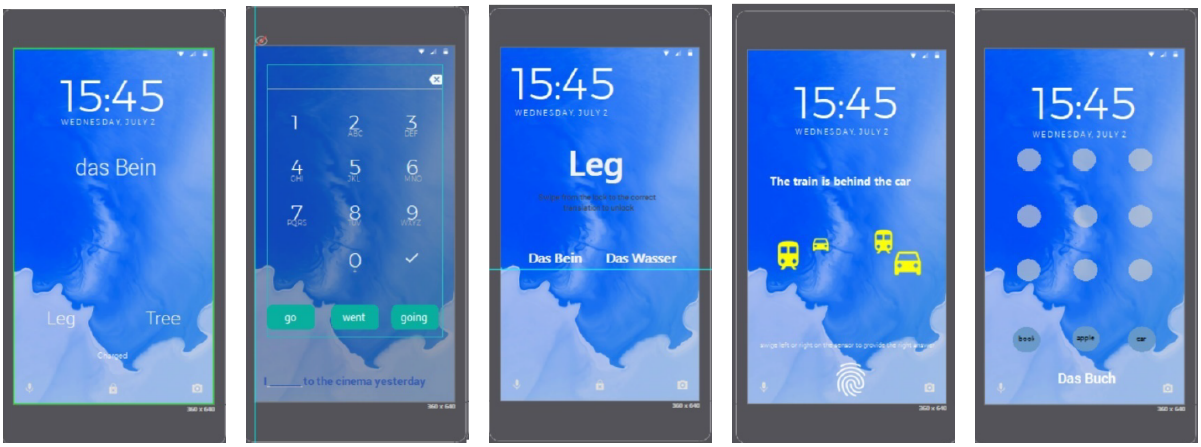
Table A.1: Person χ^2 tests for all variables assessed in the survey on users' common mobile learning situations. Asterisks indicate all statistically significant differences (* <.05| ** <.01| *** <.001, Bonferroni-corrected adjusted significance level), non-significant tests marked "n.s.". For Cramér's V, we consider values <0.2 weak associations, 0.2-0.6 moderate, and >0.6 strong associations. Gray cell text refers to a violation of the χ^2 requirement of having at least a number of five expected occurrences per cell.

	Location	**Time**	**Noise**	**Device**	**Planned**	**Company**	**Setting**	**Activities**	**Stress**	**Frequency**	**Duration**
Location	-	χ^2(35, N=131) = 79.16 *** Cramér's V = .349 ***	χ^2(21, N=131) = 84.18 *** Cramér's V = .465 ***	χ^2(21, N=131) = 61.59 *** Cramér's V = .397 ***	n.s.	χ^2(35, N=131) = 95.69 *** Cramér's V = .384 ***	χ^2(14, N=131) = 141.87 *** Cramér's V = .739 ***	χ^2(49, N=131) = 287.09 *** Cramér's V = .562 ***	χ^2(21, N=131) = 71.90 *** Cramér's V = .429 ***	χ^2(42, N=131) = 149.75 *** Cramér's V = .438 ***	n.s.
Time	χ^2(35, N=131) = 79.16 *** Cramér's V = .349 ***	-	χ^2(15, N=131) = 31 ** Cramér's V = .282 **	n.s.	n.s.	n.s.	χ^2(10, N=131) = 34.44 *** Cramér's V = .364 ***	χ^2(35, N=131) = 65.91 ** Cramér's V = .318 **	χ^2(15, N=131) = 41.06 *** Cramér's V = .324 ***	χ^2(30, N=131) = 80.25 *** Cramér's V = .351 **	n.s.
Noise	χ^2(21, N=131) = 84.18 *** Cramér's V = .465 ***	χ^2(15, N=131) = 31 ** Cramér's V = .282 **	-	χ^2(9, N=131) = 18.41 * Cramér's V = .217 *	n.s.	n.s.	χ^2(6, N=131) = 47.7 *** Cramér's V = .428 ***	χ^2(21, N=131) = 61.27 *** Cramér's V = .396 ***	χ^2(9, N=131) = 28.02 * Cramér's V = .215 *	n.s.	n.s.
Device	χ^2(21, N=131) = 61.59 *** Cramér's V = .397 ***	n.s.	χ^2(9, N=131) = 18.41 * Cramér's V = .217 *	-	n.s.	n.s.	n.s.	χ^2(21, N=131) = 50.79 *** Cramér's V = .361 ***	n.s.	n.s.	$H(3) = 13.66$ $p < .01$ **
Planned	n.s.	n.s.	n.s.	n.s.	-	n.s.	n.s.	n.s.	n.s.	χ^2(12, N=131) = 23.98 * Cramér's V = .304 *	$H(2) = 10.44$ $p < .01$ **
Company	χ^2(35, N=131) = 95.69 *** Cramér's V = .384 ***	n.s.	n.s.	n.s.	n.s.	-	χ^2(10, N=131) = 56.55 *** Cramér's V = .466 ***	n.s.	n.s.	n.s	n.s.
Setting	χ^2(14, N=131) = 141.87 *** Cramér's V = .739 ***	χ^2(10, N=131) = 34.44 *** Cramér's V = .364 ***	χ^2(6, N=131) = 47.7 *** Cramér's V = .428 ***	n.s.	n.s.	χ^2(10, N=131) = 56.55 *** Cramér's V = .466 ***	-	χ^2(14, N=131) = 89.07 *** Cramér's V = .585 ***	χ^2(6, N=131) = 31.89 *** Cramér's V = .350 ***	n.s.	n.s.
Activities	χ^2(49, N=131) = 287.09 *** Cramér's V = .562 ***	χ^2(35, N=131) = 65.91 ** Cramér's V = .318 **	χ^2(21, N=131) = 61.27 *** Cramér's V = .396 ***	χ^2(21, N=131) = 50.79 *** Cramér's V = .361 ***	n.s.	n.s.	χ^2(14, N=131) = 89.07 *** Cramér's V = .585 ***	-	χ^2(21, N=131) = 58.52 *** Cramér's V = .671 ***	χ^2(42, N=131) = 162.67 *** Cramér's V = .457 ***	n.s.
Stress	χ^2(21, N=131) = 71.90 *** Cramér's V = .429 ***	χ^2(15, N=131) = 41.06 *** Cramér's V = .324 ***	χ^2(9, N=131) = 18.41 * Cramér's V = .217 *	n.s.	n.s.	n.s.	χ^2(6, N=131) = 31.89 *** Cramér's V = .350 ***	χ^2(21, N=131) = 58.52 *** Cramér's V = .671 ***	-	χ^2(18, N=131) = 39.68 ** Cramér's V = .319 **	n.s.
Frequency	χ^2(42, N=131) = 149.75 *** Cramér's V = .438 ***	χ^2(30, N=131) = 80.25 *** Cramér's V = .351 **	n.s.	n.s.	χ^2(12, N=131) = 23.98 * Cramér's V = .304 *	n.s.	n.s.	χ^2(42, N=131) = 162.67 *** Cramér's V = .457 ***	χ^2(18, N=131) = 39.68 ** Cramér's V = .319 **	-	n.s.
Duration	n.s.	n.s.	n.s.	$H(3) = 13.66$ $p < .01$ **	$H(2) = 10.44$ $p < .01$ **	n.s.	n.s.	n.s.	n.s.	n.s.	-



(a) PT1 (b) PT2 (c) PT3 (d) PT4

Figure A.1: This figure presents four prototypes for learning tasks embedded into smartphone authentication when using "check word" tasks. **PT1** Single discrete input (tap); **PT2** Multiple discrete inputs; **PT3** Uni-directional continuous input (swipe); **PT4** Multi-directional continuous input.



(a) PT5 (b) PT6 (c) PT7 (d) PT8 (e) PT9

Figure A.2: This figure presents five prototypes for learning tasks embedded into smartphone authentication when using "multiple-choice" tasks. **PT5** Single discrete; **PT6** multiple discrete; **PT7** and **PT8** uni-directional continuous; **PT9** multi-directional continuous.

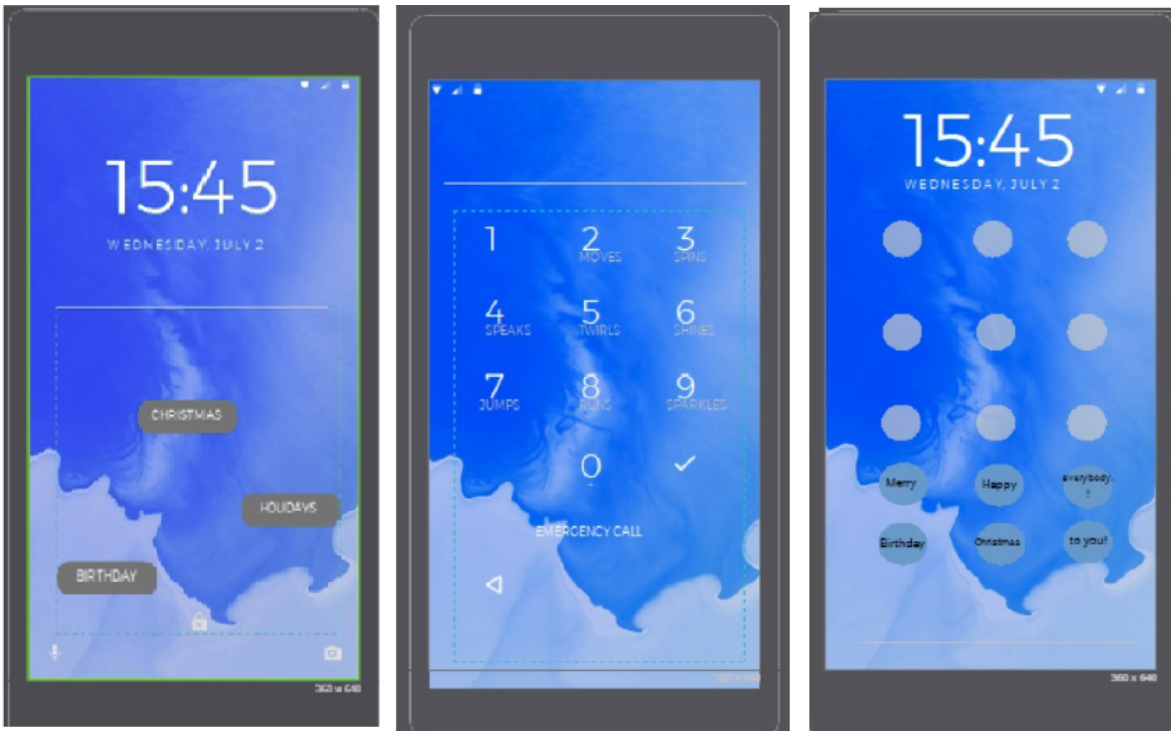


Figure A.3: This figure presents three prototypes for learning tasks embedded into smartphone authentication when using "sentence-building" tasks. **PT10** Multiple discrete, **PT11** and **PT12** both multi-directional continuous.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig und ohne unerlaubte Beihilfe angefertigt wurde.

München, den 3. Januar 2022

Christina Schneegass