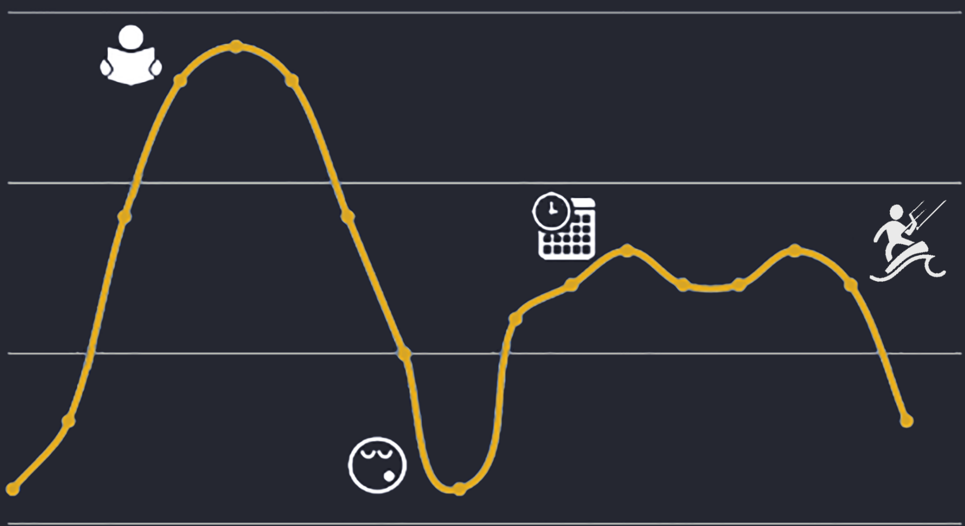
Cognition-Aware Systems

To Support Information Intake And Learning

Tilman Dingler



Cognition-aware Systems to Support Information Intake and Learning

Von der Fakultät für Informatik, Elektrotechnik und Informationstechnik und dem Stuttgart Research Centre for Simulation Technology (SRC SimTech) der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

vorgelegt von

Tilman Valentin Dingler

aus Würzburg

Erstgutachter:	Prof. Dr. Albrecht Schmidt
Zweitgutachter:	Prof. Dr. Marc Langheinrich
Externer Gutachter :	Prof. Steve Whittaker, PhD.
Tag der mündlichen Prüfung:	02.12.2016

Institut für Visualisierung und Interaktive Systeme
der Universität Stuttgart

2016

ABSTRACT

In today's information society, knowledge is created at an ever-increasing pace. As a result, most of us face a constant pressure to consume information and acquire new knowledge. But information gain and learning require time and mental resources. While the proliferation of ubiquitous computing devices, such as smartphones, enables us to consume information anytime and anywhere, technologies are often disruptive rather than sensitive to the current user context. For example, mobile applications trigger a plethora of notifications throughout the day, which often causes interruptions and breaks users' concentration. In addition, people exhibit different levels of concentration and cognitive capacity over the course of the day. During phases of low performance, the ability to concentrate is very limited, which negatively affects the effectiveness of information intake. Mobile applications do not take these variations in performance into account and often overburden the user with information or cause boredom due to a lack of stimulation.

In this thesis, we investigate how ubiquitous computing technologies can be used to help people deal with information intake and learning tasks through cognitive context-awareness. By harvesting sensor and usage data from mobile devices, we elicit people's levels of attentiveness, receptiveness, and cognitive performance. We subsequently use this cognition-awareness in applications to help users process information more effectively. Through a series of lab studies, online surveys, and field experiments we explore and quantify users' attention and cognitive performance during interaction with ubiquitous technologies. In our research we address three types of context awareness: (1) the user's momentary situational context by detecting what information content the user engages with, (2) the cognitive context exhibited by the user's momentary levels of attentiveness and engagement with technology, and (3) the user's patterns of general alertness and cognitive performance, which the user exhibits over the course of the day. We use these context factors to enhance systems with cognition-awareness and apply them to applications that support information intake and learning tasks.

In the course of this thesis, we develop three research probes with which we investigate how to build cognition- and content-aware systems. We provide empirical evidence for people being highly attentive to mobile phones, which is why we chose smartphones as the main platform for our probes. We further use context-awareness to detect moments of idleness and boredom, during which users engage with their mobile device to actively seek stimulation. More regular patterns are exhibited in users' circadian rhythms of alertness and cognitive performance, which describe systematic fluctuations because they depend on the user's internal body clock. We show how such chronobiological patterns can be elicited by conducting alertness tasks at different times of the day.

To validate our claim that such systems can support information intake and learning scenarios we develop and test four other research probes showing the feasibility of using information about the user's context to trigger content suggestions during opportune moments and adjust UIs in real-time to support reading tasks. Content suggestions in detected opportune moments (*e.g.*, bored states) lead to higher user acceptance and engagement than suggestions at random times. The success of these algorithms strongly depends on the content type and respective user state. Hence, we present and validate a conceptual framework, which we use to create algorithms that derive and classify cognitive user states based on phone usage patterns. Since people in high-alert states have more cognitive capacities, complex tasks can be handled more effectively. To explore direct applications of this hypothesis we build adaptive reading UIs that superimpose higher reading speeds and therefore cause higher cognitive load while allowing for quicker task completion time.

The tools and concepts described in this thesis allow researchers and application designers to build systems with cognition-awareness. Awareness of user's variations in levels of attention, receptiveness, and cognitive performance allows systems to trigger appropriate content suggestions, manage user interruptions, and adapt UIs in real-time to match tasks to the user's cognitive capacities. While we focus in our work on applications of these systems to support effective information intake and processing throughout the day, our tools can prospectively be applied to a broad range of applications ranging from schedule alignment according to the user's internal body clock, stress prevention through sleep/wake regulation, to recommending alertness-inducing activities.

ZUSAMMENFASSUNG

Die Geschwindigkeit, mit der sich Wissen vermehrt, nimmt immer mehr zu. In der heutigen Informationsgesellschaft nimmt der Druck stetig zu, dieser Entwicklung nachzukommen und sich ständig neue Informationen und Wissen anzueignen. Jedoch erfordert Informationsgewinn und Lernen Zeit sowie mentale Anstrengung. Während die Verbreitung von mobilen Endgeräten, wie etwa der Smartphones, es uns ermöglicht, Informationen jederzeit und überall zu konsumieren, sind es dieselben Geräte, die uns oftmals durch Benachrichtigungen aus unserer aktuellen Beschäftigung reißen und uns somit in unserer Konzentration stören. Darüber hinaus unterliegt unsere Konzentrations- und Aufnahmefähigkeit im Tagesverlauf stärkeren Schwankungen. Solche Leistungsschwankungen werden von mobilen Anwendungen derzeit nicht mit ins Kalkül gezogen. Vielmehr wird der Benutzer oftmals mit einer Fülle von Informationen überfordert oder im Gegenteil durch einen Mangel von Anregung schlichtweg gelangweilt.

Die vorliegende wissenschaftliche Arbeit befasst sich mit der Nutzung von ubiquitären Computertechnologien zur Unterstützung von Informationsaufnahme und Lerntätigkeiten mittels kognitivem Kontextbewusstsein. Aufgrund von Nutzungs- und Sensorendaten von mobilen Geräten untersuchen wir die Möglichkeit, das Aufmerksamkeits-, Aufnahmefähigkeits-, und kognitive Leistungsniveau des Benutzers zu erfassen. Anschließend beschäftigen wir uns mit Anwendungen, welche Nutzern helfen, Informationen effektiver zu verarbeiten. In einer Reihe von Laborstudien, Online-Umfragen und Feldversuchen erforschen wir das Aufmerksamkeitsniveau sowie die kognitive Leistungsfähigkeit von Benutzern während sie mit ubiquitären Computertechnologien interagieren. Dabei beschäftigen wir uns in unserer Forschung mit drei Arten von Kontextbewusstsein: (1) dem momentanen Situationskontext des Nutzers durch Erfassung des Inhalts, mit dem sich der Benutzer auseinandersetzt, (2) dem kognitiven Kontext, gemessen an der Aufmerksamkeit, welche der Nutzer den Geräten schenkt, und (3) den allgemeinen Wachsamkeitsschwankungen des Nutzers über den Tag. Mittels dieser Kontextfaktoren sind wir in der Lage, Technologien mit Kognitionsbewusstsein auszustatten, welche diese wiederum nutzen können, um Nutzer bei Informationsaufnahme und Lerntätigkeiten zu unterstützen.

Im Verlauf dieser Forschungsarbeit haben wir drei Prototypen entwickelt, anhand derer wir die Realisierbarkeit von kognitions- und inhaltsbewussten Geräten untersuchen. Wir liefern einen empirischen Beleg für die hohe Aufmerksamkeit, die Nutzer ihren mobilen Geräten entgegenbringen. Des Weitern zeigen wir, wie wir aus der Auswertung von Kontextdaten Einsichten in Situationen gewinnen, in denen Benutzer auf der Suche nach Anreizen zu ihren Geräten greifen. Weitaus systematischere Muster ergeben sich aus den circadianen Rhythmen der Wachsamkeit und kognitiven Leistungsfähigkeit, welche die regelmäßigen Fluktuationen beschreiben, die sich aus der inneren Uhr des Nutzers ergeben. In unserer Arbeit zeigen wir auf, wie solche chronobiologischen Regelmäßigkeiten durch Vigilanztests auf mobilen Endgeräten über den Tag verteilt gemessen werden können.

Um die Anwendbarkeit solcher Systeme im Hinblick auf effektive Informationsaufnahme und Lernszenarien zu untersuchen haben wir vier weitere Prototypen entwickelt und getestet, welche dem Nutzer in geeigneten Momenten Inhalte vorschlagen und Nutzerschnittstellen in Echtzeit so anzupassen, dass Lesetätigkeiten unterstützt werden. Inhalte, die zu geeigneten Momenten—z.B. wenn der Nutzer gelangweilt ist—vorgeschlagen wurden, führten dazu, dass Nutzer diese eher annahmen und sich länger mit ihnen beschäftigen, als wenn diese zu willkürlichen Zeiten vorgeschlagen wurden. Die Erfolgsrate solcher Erkennungsalgorithmen hängt stark von der Art des Inhalts und dem aktuellen Nutzerkontexts statt. Hierfür stellen wir ein konzeptionelles Rahmenwerk zur Entwicklung von Algorithmen vor, welche in der Lage sind, bestimmte kognitive Nutzerzustände anhand von der Gerätenutzung zu erkennen und vorherzusagen. Weil Menschen in Zuständen von hoher Wachsamkeit mehr kognitive Kapazitäten aufweisen, können komplexe Aufgaben effektiver abgearbeitet werden. Um direkte Anwendungen dieser Hypothese zu untersuchen, haben wir eine Reihe von adaptiven, elektronischen Leseoberflächen entwickelt, welche dem Nutzer höhere Lesegeschwindigkeiten vorgeben und ihn somit kognitiv mehr fordern. Im Gegenzug wird somit die Lesezeit verkürzt.

Die in dieser Forschungsarbeit vorgestellten Werkzeuge und Konzepte ermöglichen es Forschern und Systementwicklern, kognitionsbewusste Systeme zu entwickeln. Kenntnisse über die momentane kognitive Leistungsfähigkeit des Nutzers befähigen solche Systeme, geeignete Inhalte vorzuschlagen, Unterbrechungen durch Benachrichtigungen zu vermeiden, und die Komplexität der Nutzerschnittstelle an die aktuellen kognitiven Leistungsfähigkeiten des Nutzers anzupassen. Während sich die vorliegende Arbeit auf die Anwendung der entwickelten Systeme zur Verbesserung der Informationsaufnahme und -verarbeitung über den Tagesverlauf konzentriert, können diese Systeme zukünftig in einer großen Bandbreite von Anwendungsszenarien Einsatz finden—von Terminplanung im Einklang mit der inneren Uhr des Nutzers über Vermeidung von Stressaufkommen durch Schlafregulation bis hin zu Empfehlungen von aktivierenden Tätigkeiten.

ACKNOWLEDGMENTS

The work described in this thesis was thought up, conducted and written up in a great number of different places in collaboration with an even greater number of kindred minds. Numerous hypotheses, systems, and study designs originated or at least benefited from countless conversations with colleagues, friends, and acquaintances, often over coffee and/or fermented beverages.

Foremost, I would like to thank Albrecht Schmidt, supervisor and superhuman, for poaching me away from the Bay Area engineering madness and reintroducing me to the free-spirited world of academia; for always providing support and freedom to pursue this dissertation quest and having my back whenever bureaucracy struck.

I also owe gratitude to my external reviewers Marc Langheinrich—devil's advocate in the best sense—and Steve Whittaker who spontaneously cleared his afternoon schedule for putting sense into this thesis' story.

I would further like to thank Niels Henze, who taught me a great deal about academic work-life, essential sarcasm, and whose feedback is as merciless as his singing.

Alireza Sahami, the Persian prince who took me by the hand when my writing skill was still limited to source code and Tinder messaging.

Bastian Pfleging and Florian Alt, who made me feel welcome during my repatriation and both of which endured nightly tours through the streets of San Francisco and were the first to welcome me on planet Stuttgart.

Thomas Kubitza, with whom I mounted planes and cranes, and made a decent living off of numerous foosball opponents around the European bar scene.

Stefan Schneegass, BBQ instigator and favorite critic, who kindheartedly welcomed me into his family.

Markus Funk—fellow cheese lover—for keeping me in touch with everything hip.

Francisco Kiss—neighbor, colleague, dear friend—for our walks and talks, and whisky-induced guitar sessions.

Tonja Machulla, for her sleep sacrifices, openness in discussions, and management of her husband's leeriness. Lewis Chuang, for his patience with me and his wife.

Yomna Abdelrahman, for keeping my blood sugars high.

Anja Mebus for her endless patience with a foreign specimen like me. Murielle Naud-Barthelmeß, for keeping the troops in line in the nicest of French office regimes.

Patrick Bader, for enduring my office singing and eventually chiming in.

And all the rest of the hcilab troop who brightened my Stuttgart days: Alexandra Voit, Dominik Weber, Huy Viet Le, Jakob Karolus, Lars Lischke, Mariam Hassib, Mauro Avila, Miriam Greis, Pascal Knierim, Passant El Agroudy, Pawel Wozniak, Romina Kettner, Rufat Rzayev, Sven Mayer, Oliver Korn, Thomas Kosch, Valentin Schwind, Katrin Wolf, Norman Pohl, and Nora Broy.

I also want to thank my external collaborators, such as the entire RECALL team from the University of Lancaster, University of Essex, and the University of Lugano. Kai Kunze and Susana Sanchez, for building up the Japan connection. Martin Pielot, for inviting me to Telefónica R&D to work on anything but *boring* projects and giving me a first-hand glimpse into the intricate workings of industry research. Dagmar Kern, for not getting tired of repeating a 70-participant study.

I was further lucky to meet and work with countless talented students. Especially, the works of Katrin Angerbauer, Johannes Knittel, and Rufat Rzayev found its way into this thesis.

At last, but never least, I have drawn massive support from a colorful network of extraordinary people I am lucky to call friends spanning from San Francisco to Tarifa and Istanbul, where Tugce Tuğalan—the artist of this book's cover—resides.

And, to adequately conclude this list, I want to thank and honor the two people I owe everything to: Gisela and Wolfgang Dingler, for unlimited love and support, patience and prudence, and some megatons of noodle salad to keep me afloat.

Table of Contents

LIST OF FIGURES

List of Tables

List of Acronyms

ANOVA	Analysis of Variance
API	Application Programming Interface
BCI	Brain-Computer Interface
BPM	Beats per Minute
DVB-C	Digital Video Broadcasting - Cable
EEG	Electroencephalography
ECG	Electrocardiography
EDA	Electrodermal Activity
EMG	Electromyography
EOG	Electrooculography
EPG	Electronic Program Guide
EMA	Ecological Momentary Assessment
ESM	Experience Sampling Method
FA	False Alarm
fMRI	Functional Magnetic Resonance Imaging
fNIR	Functional Near-Infrared Spectroscopy
GNG	Go/No-Go-Task
GSR	Galvanic Skin Response
HCI	Human-Computer Interaction
HRV	Heart Rate Variability
LR	Logistic Regression
ML	Machine Learning
MOT	Multiple Object Tracking
MOOC	Massive Open Online Course

NASA-TLX	NASA Task Load Index
NSF	National Science Foundation
PVT	Psychomotoric Vigilance Task
RF	Random Forest
RQ	Research Question
RSVP	Rapid Serial Visual Presentation
RT	Reaction Time
SVM	Support Vector Machines
UI	User Interface
WPM	Words per Minute

I

Introduction and Background

Chapter 1

Introduction

In recent decades we have undergone remarkable changes towards today's information society. In 1981 Buckminster coined the theory of the *knowledge doubling curve*: since the beginning of mankind the speed at which information doubles has been getting increasingly faster [105]. From the invention of writing and then of printing all the way to the advent of the *World Wide Web*, this development has been accelerating.

The transition into a knowledge society requires us to effectively deal with this information growth and constantly advance our learning. But acquiring new knowledge, skills, and practices requires significant investments in time and mental resources. People develop their innate strategies for taking in and processing new information: while self-driven learners, for example, cope very well with learning tasks on their own, others prefer a more structured approach through regular courses and dedicated learning sessions. Information intake and knowledge acquisition is a highly personal process, there is no *one-size fits all* approach: people exhibit differences in aptitude, interest, and background knowledge, which requires learning processes to become more customized to the individual.

The *National Academy of Engineering* lists "*Advance Personalized Learning*" as one of 14 grand challenges for engineering in the 21st Century[1]. Current efforts to take into account individual learning styles make use of computer-supported

[1] http://www.engineeringchallenges.org/

instructions, often in the classroom or via the Web. An increasing supply of learning platforms is available online, such as self-directed learning through software, web-based resources, and Massive Open Online Courses (MOOCs). Learning platforms adjust curricula to the individual's needs and learning state: knowledge management systems, such as *kahnacademy.org* allow learners to keep track of their progress and facilitate the interaction between learners and instructors. Based on the personal learning history such platforms also employ recommender systems to suggest relevant learning content. Personalized learning has attracted research, philanthropic, and commercial interest. According to a report from *EdTechXGlobal* and *IBIS Capital*, the global education technology market will grow 17% per year to 252bn USD by 2020[2], which further sparks the number of emanating online offerings.

Meanwhile, information has become more accessible through the proliferation of ubiquitous computing devices, such as laptops, phones, tablets, watches, and smart eye-wear. Two Billion consumers have smartphones[3] and wearable devices are increasingly being commercialized, including wristbands, smart garments, watches and other fitness monitors: in 2016, 274.6 million wearable electronic devices will presumably be sold worldwide, which constitutes an 18.4% growth rate compared to 2015 [139]. Mobile devices possess rich multimedia output capabilities and make information available to their users anytime, anywhere. Hence, information can be consumed and learning sessions can take place throughout the day even while people are on-the-go. Mobile applications have become popular for reading (*e.g.*, Kindle app), following the news (*e.g.*, Feedly), and also as learning tools for language vocabulary (*e.g.*, Duolingo), digital flashcards (*e.g.*, Anki), and taking online courses (*e.g.*, Udemy).

However, while these tools are being constantly available, users find themselves highly receptive at some times during the day and unable to concentrate at other times. People exhibit different attentional phases throughout the day varying between focused states, in which cognitive activities can be performed with ease, and rather inattentive states, during which perception is constricted and cognitive processes run slower. While systems that assess users' cognitive states exist, for example, in automobiles where eye movements are tracked to monitor driver's fatigue levels, they are limited to specific application scenarios. Mobile applications on consumer devices do not take into account the user's different attentional phases and current cognitive context, thus either rely on explicit user

[2] `http://techcitynews.com/2016/05/26/report-edtech-spend-will-reach-252bn-2020/`

[3] `http://www.emarketer.com/Article/2-Billion-Consumers-Worldwide-Smartphones-by-2016/1011694`

action (*e.g.*, waiting for the user to launch the app), or trigger reminders regardless of the cognitive state the user is in, hence often causing interruptions. Similarly, interfaces on such devices do not automatically adapt to the cognitive capacities available and rather require users to explicitly customize the interface.

This thesis investigates how ubiquitous technologies can be used to help people deal with information processing and adapt learning tasks to the user through awareness of users' situational and cognitive context. Equipped with rich sensing capabilities, ubiquitous devices can make sense of user context and adjust their output in ways to provide cognitive support to users throughout the day. Therefore, we focus on how systems gain cognition-awareness in the first place and explore applications to facilitate information intake: how detecting cognitive context can be applied to making use of opportune moments for engaging users with information and learning tasks, and to adjusting information interfaces in real-time.

Awareness of cognitive capacities allows applications to match the user's current state with the complexity and presentation of a task at hand. By helping users deal with tasks in opportune moments where their performance levels match the task requirements, such systems enable quick and effective task handling. Therefore, cognition-awareness allows systems to suggest and schedule learning activities: for example, prompting users to repeat foreign language vocabulary or engage in reading activities when they are most receptive and likely to memorize consumed content. Further, User Interfaces (UIs) that adapt to the user's cognitive context can balance interface complexity and task efficiency: by increasing task complexity (*e.g.*, speeding up tasks) during phases of high cognitive performance interfaces can prevent users getting bored, while reducing complexity (*e.g.*, slowing down tasks) during phases of low cognitive capacity can prevent frustration.

To build cognition-aware systems to facilitate information intake and learning we first focus on the underlying cognitive phenomena of effective information processing. At its core lies the notion of human attention, which is crucial for perceiving, processing, and memorizing information. In the course of this thesis, we present a series of studies, in which we quantify people's attention throughout the day to identify opportune moments for information intake. Therefore, we explore characteristics of these moments and how to detect them based on smartphone usage patterns. Using machine-learning techniques we develop algorithms to assess the user's momentary attentiveness. But attention and receptiveness does not only depend on the user's current situation but is also impacted by diurnal rhythms. Following a sinusoidal pattern, our internal body clock determines hours of the day when we experience a particularly low or particularly strong sleep drive, which has a direct impact on our ability to focus. Hence, the second aspect of cognition-aware systems is the detection of general patterns across the day—the so-called circadian rhythm of alertness and cognitive performance.

In the course of this research, we develop tools that can be used and applied by researchers and system engineers to develop technologies capable of detecting users' cognitive context. Therefore, we create algorithms for detecting current levels of attentiveness and general patterns of alertness. We present cognition-awareness as an additional variable of context-aware computing. Further, we develop and test direct applications for such systems: the concepts and implementations presented in this thesis enable users to engage with learning and reading activities in opportune moments. We further explore dynamic reading interfaces whose mental demand on the user can be adjusted by balancing reading speed, cognitive load, and comprehension. Thus, cognition-aware systems can suggest users different reading modes according to different cognitive states. By adjusting when to engage users and how to present information processing tasks according to cognitive capacities, such systems customize information intake and learning on a deeply personal level.

1.1 Research Questions

To investigate the use of ubiquitous technologies to facilitate information intake and learning three main aspects need to be considered, namely the user, the user context, and the application. Table 1.1 lists the corresponding research questions, which have driven the research presented in this thesis.

Attention is a crucial factor for effective information intake and digestion. Therefore, we first focus on **human attention** in a technology context and how it can be quantified across the day (Research Question (RQ)1). With the goal of identifying opportune moments for information intake, we investigate the kind of states, in which users turn to their devices to seek stimulation. Here, we focus on the notion of boredom in a mobile context and whether such states can be detected based on phone usage patterns (RQ2).

Awareness of user states and intentions is a **context** dimension we further explore from two perspectives: first, we look at the current content the user engages with and how awareness of it can be used in multi-device environments to facilitate information intake and enrich the user experience (RQ3). The second context

RQ	Research Question	Chapter
	Human Attention	
RQ1	How can users' attentiveness be quantified across the day and reliably predicted from phone usage patterns?	3
RQ2	Does boredom measurably affect phone usage and which usage characteristics are most prevalent during such states?	3
	Context Awareness	
RQ3	How can awareness of the content which the user is currently exposed to be used to augment the user experience?	4
RQ4	How can technology be used to elicit the user's circadian rhythm of alertness and cognitive performance?	5
	Applications	
RQ5	How can opportune moments for content delivery be used to foster information intake and learning?	6
RQ6	How can reading UIs be adjusted in real-time to decrease or increase information bandwidth?	7

Table 1.1: Overview of research questions that build the base of this thesis.

dimension we look at is the user's variations of diurnal alertness patterns across the day. Depending on sleep routines and an innate bio-rhythm people exhibit changing phases of high and low phases of alertness. Awareness of these phases can inform technologies when to expect receptive user states and feed information accordingly. Hence, we investigate ways to elicit these diurnal changes of alertness, also called circadian rhythms (RQ4). The combination of predicted attention levels based on device usage patterns and the knowledge of user's circadian rhythms of cognitive performance adds cognition-awareness as an additional layer to context-aware systems.

In the final part of this thesis, we focus on the **application** layer. Considering cognition awareness as viable context dimension when users engage with technology and in information consumption, we focus on the type of adjustments technologies can undergo based on such awareness. First, we investigate how opportune moments for technology-triggered content delivery can help to suggest different types of content (RQ5). Here, we focus on a language learning scenario and information intake through reading. Since reading is the most predominant way of acquiring information on mobile devices, we further look at how UIs can dynamically adjust to the attentional and cognitive state of the user (RQ6). Besides studying different ways of adjusting information bandwidth while reading we focus on providing user controls beyond mere cognition awareness. Here, we investigate implicit mechanisms to control the reading flow of a reading technique called Rapid Serial Visual Presentation (RSVP).

1.2 Vision: Cognition-aware Systems

Education and continued learning are now more important than ever while we face challenges, such as moving towards a knowledge society, changing work environments, globalization, and the need for mutual tolerance and understanding. Meanwhile, technology is becoming more intertwined with our lives as devices, clothes, our environment and even our bodies become more and more equipped with it. In the work presented here, we investigate how ubiquitous technologies can be used to help people cope with the information growth and advance personalized learning. We believe that such technologies can help people meet the rising demands of our knowledge society, in which constant learning is crucial. In the following, we map out our vision of cognition-aware systems and how they can facilitate information intake and learning throughout the day.

Technology's ultimate goal is to support its users. The more a device knows about its user in terms of physical as well as psychological constraints and capabilities, the better it can provide assistance. A fully integrated device environment knows when to approach the user and how to do that in order to help users consume, process, and memorize information. It starts by detecting user states and identifying opportune moments (the *when*) for technologies to actively approach users and suggest different types of content (the *what*). The next step takes into consideration the current device environment, its capabilities and display parameters and adjusts the presentation of content so that it optimally fits the user's current physical and cognitive context (the *how*). An integrated system keeps track of information encountered and processed. This knowledge can help adjust the presentation of new information items to build associations between related information. Then, repetition sessions can be scheduled in accordance with learning theories to commit that information to long-term memory.

A holistically aware system, therefore, needs information on the user's knowledge background, preferences, pending tasks, and cognitive states (across time), but also about the environment (devices nearby) and the world (relevant information). It further provides mechanisms that allow users to customize the system and make manual adjustments. The idea behind context-aware computing is to support users in-situ according to their current situation. Cognition-aware systems focus on the user's mental state and current information processing capabilities to complement this trend towards a holistic context awareness. Systems that know about high and low attention phases throughout the day are able to support and enhance cognition and human memory in multiple ways:

Task Scheduling

Knowledge about the user's diurnal attention rhythms can inform systems to schedule tasks respectively across the day. By analyzing tasks and their cognitive requirements they can be matched with the user's attention phases. Matching is done with the goal of increasing the overall productivity of the user: complex tasks are met with phases of higher concentration and can, therefore, be mastered more effectively or in shorter amount of time; phases of lower concentration, on the other hand, can still be useful to perform daily chores, such as grocery shopping or answering routine emails, without wasting precious performance capacity. Keeping task requirements and available cognitive capacities in balance can help reduce boredom (in case of high capacity and low task demand) and frustration (low capacity and high demand). Further, smart scheduling of tasks can help users to experience so-called flow states in a systematic way. Such flow states have been described by Csikszentmihalyi [52] as situations in which the user is fully immersed in the task at hand.

UI Adaption

Knowing about the user's current cognitive capacity can further influence the presentation of a task. The higher the user's attention the more complexity could be displayed to allow a more efficient completion of the task. Reading activities, for example, can be sped up according to the user's ability to concentrate while making compromises on comprehension levels. A reading UI that adapts to the current capacity to absorb information could allow users to effectively take in, process, and retain more information in a shorter amount of time. On the other hand, an adaptive UI could also hide functionality during a design task when the user is detected to be tired. In such a state decreased UI complexity could help the user focus on the task at hand.

Disruption Management

Cognition-aware systems could be further used to manage external interruptions by keeping a tab on the user's current state and the importance of incoming notifications or alerts. In moments of high focus or task immersion notifications, incoming phone calls and other disruptions can be effectively delayed to more opportune moments to not disturb the user's flow. This can be beneficial for tasks that need an uninterrupted string of thoughts, but also for situations, in which interruptions through technology are merely inappropriate, for example when immersed in a conversation with a friend or loved one. Distractions could further be proactively prevented by blocking access to potentially distracting websites or applications to protect periods of high focus.

Self-Awareness

Self-awareness is the first step towards self-improvement. By eliciting diurnal attention patterns users can become aware of their own productive phases and hence make more informed decisions about their own task scheduling. For instance, when it comes to deciding when to schedule a reading session or when to deal with daily chores, such as doing laundry. Awareness of the potential lack of synchronicity between timetables and optimal cognitive performance can help users avoid frustration. Kreitzman and Foster [163] showed that working out of sync with our individual circadian rhythms of performance can even be harmful with negative long-term consequences for our health and well-being.

Personal Assistance

Cognition-aware systems become truly personal assistants, which learn about users' patterns and schedule their days in their best interests. Such assistants, therefore, become the connective tissue between user, devices and the world around. High profile characters, such as top managers or head of states often have access to an entire staff that focuses on managing their daily routines and structuring their day as effective as possible. The resulting daily agenda entails appointments and completion of tasks, but also sleep, nutrition, workout routine, information consumption, and other daily chores. By enabling technology to learn about the habits, activities, and cognitive states of an individual user, we can build systems that go beyond simple context-aware applications. Cognition-awareness allows us to build mobile personal assistants that accompany users throughout their day, detect their cognitive states and structure their task lists in a way so that each task is matched by the optimal user state. Such systems have the potential to help users be more effective at their tasks, increase their overall productivity, happiness (through reduced levels of frustration), and eventual well-being.

1.3 Challenges and Contribution

The life of today's information workers requires people to keep up with an abundance of information on a daily basis. Meanwhile, our knowledge society demands us to employ a habit of constant learning. While our time and resources are generally limited we need to look for ways to efficiently acquire new knowledge, skills, and practices. With the ubiquity of mobile devices, learning can take place on-the-go—anytime and anywhere. Hence, we tackle the question of how technology can help us in dealing with effective information intake and learning throughout the day. In the course of this thesis we focus on three major challenges:

1. Technologies—even though becoming increasingly context-aware—do rarely consider users' attention levels, receptiveness or cognitive capacities, all of which can change significantly across the day. The circadian rhythm of alertness and performance exhibits phases, in which people are more concentrated and can, therefore, be more productive than during other times of the day. However, technology does not take this into account and opportunities for effective knowledge transfer are missed. Instead, while being unaware of potentially productive states, people often spend idle moments looking for stimulation by aimlessly hopping between apps and services. The challenge is how technology can extract diurnal user patterns and therefore gain awareness of circadian rhythms of attention in order to predict productive phases.

2. Reminders and alerts often do not consider the user's current context. Despite an existing body of research in delaying notifications, little of these algorithms have made it into consumer-level products. Hence, technology often exhibits a distracting nature where users are at risk of getting interrupted in their task flow or where they neglect their social environment. There is a clear dichotomy between the near-constant availability of and through mobile devices and the imminent disruptions that come with this device ubiquity. Disruptions have been shown to have negative effects on productivity [180], which raises the challenge of detecting and utilizing opportune moments for content delivery.

3. Little research has been conducted on how to increase or decrease information bandwidth according to users' available cognitive capacities. For instance, in phases of high concentration, more information can be processed effectively, which in turn can lead to quicker task completion. In low concentration phases, on the other hand, decreased complexity can prevent frustration. Current UIs do rarely accommodate for adapting UI complexity to accommodate different cognitive states. Users have little means to adjust interfaces to their cognitive capacities in order to get tasks done more efficiently. Hence, the challenge lies in identifying possible UI adjustments as well as designing and developing effective user controls.

In this thesis, we tackle these challenges by bringing together theories from the field of cognitive psychology with technological capabilities. By applying memory and learning theories, utilizing ubiquitous sensing, near-constant device availability and adjustable output capabilities we set out with the goal of:

1. Quantifying human attention and identifying cognitive performance patterns in a technological context to enhancing context-aware systems by adding a cognitive dimension.
2. Identifying opportune moments for content delivery and providing content that is relevant to the user's context.
3. Creating adjustable information interfaces for users' varying cognitive capacities to absorb information more effectively and increase overall user productivity.

Our work focuses on the application layer and has its roots in the field of human-computer interaction (Human-Computer Interaction (HCI)), cognitive psychology and algorithms (Machine Learning (ML)). Therefore, we take a human-centered research approach by conducting formative studies, lab, and field experiments. In the course of this thesis, we present a series of user studies that allowed us to collect meaningful feedback from users, as well as develop and test algorithms and prototypes. Table 1.2 gives a summary of the prototypes created in the course of this thesis. Towards achieving the aforementioned goals our experimental approach yields the following contribution, an overview of which is depicted in Figure 1.1:

(a) Quantification and Assessment of Users' Attentiveness

In a field study, we collected ground truth for quantifying people's attentiveness and receptiveness to interruptions. By applying a machine learning model to predict users' attentiveness throughout the day we found that people are highly attentive to messaging for most of their waking hours [79]. While phases of inattentiveness only last for a few minutes, delay strategies for notifications or alerts are applicable without much risk of missing timely information. Therefore, alerts are often not required to regain user attention. In idle moments people tend to turn to their phones seeking stimulation, hence we developed a detection algorithm to identify and predict these moments [214]. We used experience sampling to identify times during the day, in which people are generally bored. We analyzed corresponding phone usage patterns and found features, such as the recency of communication, usage intensity, time of day, and demographic information to be prominent indicators for bored *vs.* non-bored user states.

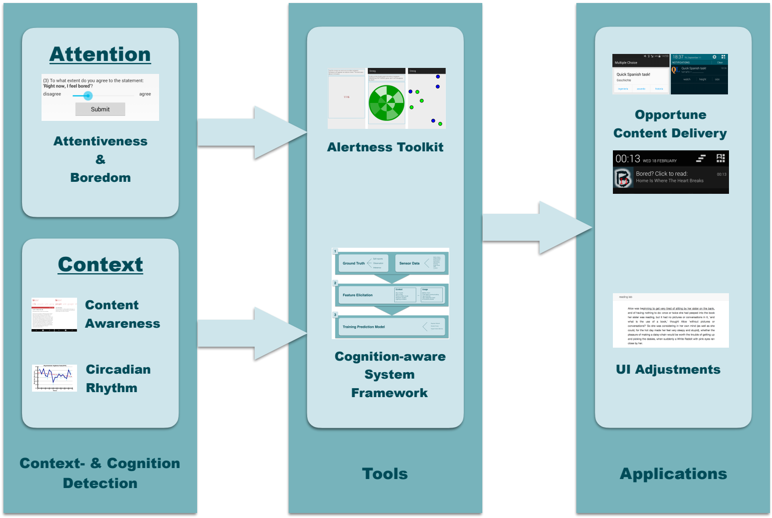


Figure 1.1: Contribution overview: research in human attention and context-aware systems build the bases for the creation of tools for cognition-aware systems and applications.

Awareness of attentional states can be used for benign interventions, such as triggering content recommendations in form of reading or learning materials.

(b) Tools for Researching and Building Cognition-aware Systems

In-situ attention detection is helpful for delaying interruptions. However, in order to effectively schedule tasks, technologies need to gain a more holistic awareness of attentional patterns across the day. Therefore, we look into circadian rhythms of alertness and performance and adopt a series of assessment tasks for being completed on a mobile platform. We validate these tasks in a user study and show how they can be used to extract a general model of the circadian rhythm of alertness [85]. We release a mobile toolkit for eliciting that rhythm as an open source library for HCI and psychology researchers for future work on building cognition-aware systems. We further present a framework for correlating technology usage patterns to cognitive states in three steps: ground truth collection, feature extraction, and model training [68]. In subsequent studies, we validate the approach to identify opportune moments for learning tasks and reading activities.

(c) Identification and Characterization of Opportune Moments for Content Delivery

Based on user context we set out to investigate opportune moments for information intake and so-called microlearning tasks. Therefore, we look at language learning throughout the day while people are on-the-go [87]. Being in transit and during idle moments, reviewing foreign language vocabulary is well received and shows the feasibility of scheduling learning tasks throughout the day without requiring users to dedicate big chunks of time. We further apply our boredom detection algorithm to suggest reading content in opportune moments leading to more articles being clicked and more time spent reading in states of boredom [214].

(d) Implications for Adaptive Reading

Phases of high concentration can be taken advantage of by either working on more complex tasks or by tackling a task in a more efficient way. The higher the user's cognitive capacity, the more demanding a UI can be designed with the goal of increasing task efficiency and prevent boredom. Since reading is a primary channel for information intake, we present a series of studies, in which we look at adaptive reading UIs. Therefore, we adapt speed reading techniques for electronic reading devices and assess the feasibility of actively guiding the user through text. Besides the established method of RSVP we introduce a kinetic stimulus that effectively increases user's reading speed while raising cognitive load and making compromises on text comprehension [86]. We show that increasing the reader's information bandwidth in this way helps users to get through text faster at moderate comprehension levels. Because a thorough understanding of a text is not always essential, especially when it comes to the different types of readings, such techniques can be used to sift through information effectively and deal with the daily reading load. To enable users to control reading flow in dynamic reading UIs we investigate implicit interaction through eye gaze. We show that such reading support helps readers increase text comprehension as compared to explicit touch controls [80].

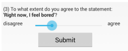
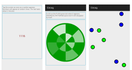
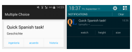
Prototype	Description	Chapter
	*Borapp* is an Android app we released on Google Play. The app records phone usage data and triggers ESM surveys throughout the day, through which we collect ground truth on users' receptiveness towards mobile messaging and self-assessed boredom states.	3
	The Windows Phone app *TVInsight* proactively displays relevant content in sync with the current TV program. By linking and retrieving additional content from Wikipedia and Google searches this second-screen app enhances the user experience through content-awareness.	4
	*Circog* is a mobile toolkit we developed for the Android platform and which triggers alertness tests throughout the day. We validated and released the toolkit as open source library to allow researchers and app developers to elicit users' circadian rhythm of alertness and cognitive performance.	5
	*QuickLearn* is an Android app for reviewing foreign language vocabulary throughout the day. It triggers interactive notifications in form of flashcards and multiple choice questions, which can be responded to directly in the notification drawer. We deployed this app on Google Play to research opportune moments for delivering learning content to users.	6
	We implemented a *kinetic stimulus* as a web reading interface that highlights the supposed reading position line by line on electronic reading devices and therefore guides the reader's eyes across text. By dictating reading speed of users cognitive load is induced, which serves as application scenario for UI adjustments based on user's cognitive capacities. We compare its effectiveness to reading with RSVP.	7
	For reading with RSVP on smartwatches we created an Android application that receives commands from a Pupil eye tracker to control the reading flow implicitly through the user's eye gaze.	7

Table 1.2: Overview of prototypes created in the course of this thesis.

1.4 Ethics

For most of the user studies described in this thesis, we followed the ethics process derived from a previous EU project, namely the pd-net project [170]. This process entails the submission of a detailed study plan document to a central platform where it was reviewed by our project partners prior to the experiment. An example document, which we used for the *QuickLearn* study described in Section 6.2 can be found in the Appendix. For so-called studies in-the-wild where we used application stores to distribute our technology probes, we included a consent form that would be shown after its installation. The respective probe would not start collecting data until users had given explicit consent. We provided further details with regard to the data types collected on accompanying websites (also see Appendix). The studies depicted in Chapter 3 and 6 were conducted in a corporate environment and therefore followed an internal examination process. However, all studies reported here were conducted in accordance with the declaration of Helsinki[4].

Privacy

In the course of our research, we deployed a number of research probes that collected a vast amount of data from the users' devices. This was necessary because many of the algorithms we developed required access to people's phone usage patterns. Getting people to agree to use apps and services that analyze their phone usage can be tricky and raises a number of privacy concerns. In general, algorithms that access such sensitive data to use them as a base for recommendations and therefore influence user behavior need to be critically viewed. The benefit of granting such insights needs to outweigh the risks. While people are generally concerned about sharing their location information, the gains of using navigation services, for example, make people compromise on their private data policy. However, for many of the algorithms proposed in this thesis data processing and ultimately state predictions can be done locally on the user's device. "*Privacy by Design*" has been proposed to ensure that attention to privacy and security is paid throughout the engineering process of ubiquitous systems [169]. One aspect of it is making sure that sensitive information does not leave the user's device, but is processed in-place. Such an approach could certainly be applied by cognition-aware systems. For more power-intensive

[4] http://www.wma.net/en/30publications/10policies/b3/index.html

data analyses that exceed local device resources homomorphic encryption could further be applied [109]. This scheme allows devices to encrypt data locally with a secret key and send it off to a server where arbitrary functions can be performed over it without description key. Hence, the user device is the only key holder and therefore has sole access to its own data.

User Awareness

When technology tracks and accompanies our every step as well as thought, individuals trigger non-deliberate events just by being in a certain location or feeling a certain way. The interaction with cognition-aware systems is not necessarily clear from a user perspective. When a system behaves differently because of a certain cognitive trigger, the user may be unaware of having triggered such a change. Therefore, the classical user-centered design approach, where we follow an iterative cycle comprising of studying, designing, building, and evaluating a technology with the user, needs to be updated. Sellen *et al.* [242] speak of "*folding human values into the research and design cycle*". The user remains at the center of the design model, but broader implications across time will need to be considered as well. Storing personal data may be secure at the time, but does not guarantee its inviolability for the time to come. Potential risks and implications should, therefore, be considered already during the design cycle of new technologies.

Developer Guidelines

Cognition-awareness and the resulting predictions can have an alienating effect on users. In the physical world, we are very much used to being influenced. In the digital world, this is still often controversial and considered manipulation. Examples of search results are one example. The big difference in our view is the visibility of the technologies and the understandability of how it influences our behavior for things in the physical world. As more and more things around us become computer controlled and more information we consume become digital, we have to face the issue that our experiences are strongly determined by software. For cognition-aware systems, it is important to be aware of this fact and to understand that content suggestions may influence users. In order for users to accept these systems, it is essential not to 'trick' the user or manipulate them into actions they would not want to do. The following basic rules give guidance on designing such systems:

- empower the user to explore WHY certain information is presented and why other information is not presented
- make it apparent to WHAT contextual factors the information is adapting and enable the user to personalize the adaptation

It is apparent that any information presented (or not presented) may impact the user. At the same time providing information that is contextualized will ease many cognitive tasks. There is no silver bullet here, but it is central for developers to consciously make these decisions when creating the system and to make them explicit in their system design.

1.5 Research Context

The work presented in this thesis was carried out over the course of about four years in the Human-Computer Interaction and Cognitive Systems group at the Institute for Visualization and Interactive Systems. The group is located at the University of Stuttgart under the supervision of Prof. Albrecht Schmidt. Further input came from the Graduate School of the Simulation Technology Program at the University of Stuttgart, which provided the author with the opportunity to gather interdisciplinary input from both technical, but also humanistic fields, such as philosophy. Several collaborations with experts from the field resulted in joint publications that contributed to this thesis.

RECALL

The major part of this work was conducted within the EU project *RECALL*[5] with funding through the Future and Emerging Technologies (FET) programme within the 7th Framework Programme for Research of the European Commission, under FET grant number: 612933. In *RECALL*, four partner universities (University of Lancaster, University of Essex, Università della Svizzera italiana, and University of Stuttgart) set out to re-define the notion of memory augmentation through ubiquitous technologies. By combining technological interventions with basic research questions in memory psychology, this 3-year research project (Nov. 2013 - Oct. 2016) focused on investigating and enhancing the way people use

[5] http://recall-fet.eu/

technology to remember and to externalize memory. The collaboration between project partners resulted in conjoint publications [69, 6], among others at the CHI 2016 conference [172]. Further, we jointly created and organized a workshop on "Mobile Cognition" at the International Conference on Human-Computer Interaction with Mobile Devices and Services [72] in 2015.

Telefónica R&D

In 2014, the author spent four months as an intern at Telefónica R&D in Barcelona, Spain. There, he worked together with Martin Pielot under the supervision of the scientific director at Telefónica, Nuria Oliver. The work conducted there was related to human attention research and is reflected in Section 3.3 and 6.3 of this thesis. The collaboration resulted in a number of publications [79, 214] as well as a patent filing.

Keio University, Japan

In the context of investigating knowledge acquisition points, the design and exploration of interactive reading UIs became prevalent. The author of this thesis worked closely together with experts in the field of reading and eye tracking including Kai Kunze and Susana Sanchez from Keio University in Tokyo, Japan. Mutual visits throughout the four years of conducting this research lead to a number of conjoint publications [86, 168, 231, 233, 145]. Together we successfully launched a continuing series of workshops on "*Augmenting the Human Mind*" at the ACM International Joint Conference on Pervasive and Ubiquitous Computing [84, 167, 77].

1.6 Distribution of Work

Parts of this thesis have been published in scientific conferences, and workshops: [68], [79], [80], [85], [86], [87], [160], and [214].

Other publications in scientific journals, conferences, and workshops by the author that go beyond the scope of this thesis include topics, such as lifelogging [6, 70, 172, 277], memory augmentation [69, 72, 76, 77, 81, 84, 167], reading on electronic devices [8, 145, 168, 233], peripheral displays [71, 74, 75, 83, 267], context awareness [106, 231, 232], auditory displays [73, 78, 264], and others [82, 164, 165, 166, 176, 240, 278].

In the following, the collaborative efforts and publications are listed that have lead to the respective research probes and user studies described in this thesis:

Chapter 3 - Human Attention. The study described in this chapter, which lead to two publications at MobileHCI'15 [79] and Ubicomp'15 [214] respectively, was conducted during the author's research internship in 2014 at Telefónica R&D in Barcelona, Spain. Idea, concept, implementation, and data collection stemmed from the two main paper authors Martin Pielot and Tilman Dingler under the general supervision of Nuria Olivier. For the data analysis and the training of the machine-learning model, we received significant support from Jose San Pedro.

Section 4.2 - Context-Aware Information Delivery. This chapter is based on the Master thesis project of Johannes Knittel [159] whom the author supervised at the time and which resulted in a publication at TVX'16 [160]. Design and implementation of the apparatus were lead by the student who ended up applying the resulting algorithm in his startup *FlickStuff* [6], which he co-founded right after completing his thesis.

Section 5.2 - Eliciting the Circadian Rhythm of Alertness. This project was mainly driven by the author in collaboration with Tonja Machulla whose input regarding concept and study design were invaluable and who played a major role in the study evaluation and data analysis. Valuable input for concept and framing by Albrecht Schmidt resulted in a paper, which was published in the IMWUT'17 journal [85].

Section 6.2 - Micro-Learning Sessions Throughout the Day. App and study design were supported by the visiting researchers Jennifer Sykes and Chun-Cheng Chang whose stay was funded by the National Science Foundation (NSF). The author mainly drove the concept, development, study, data analysis, and feature elicitation. Dominik Weber contributed his experience in Android programming and Martin Pielot trained the machine-learning model to elicit the feature ranking. With the help of Niels Henze, the project resulted in a paper presented at MobileHCI'17 [87].

[6] http://flickstuff.de/

Section 6.3 - Using Predicted Boredom to Suggest Reading Content. The study was designed and main parts of the apparatus were developed during the author's internship at Telefónica R&D. While the study was conducted by Martin Pielot after the author had completed his internship, the data collection happened in collaboration. Martin Pielot then finished the data analysis together with Jose San Pedro under the supervision of Nuria Olivier. The results were published in a conjoint conference paper at Ubicomp'15 [214].

Section 7.2 - Dynamic Reading Interfaces to Increase Reading Speed. The research around speed reading was mainly driven by the author under the supervision of Albrecht Schmidt and with concept refinements by Kai Kunze. Alireza Sahami contributed significant parts of the data analysis, while Thomas Kosch helped with the implementation of the study apparatus. Results were published at AH'15 [86].

Section 7.3 - Implicit Reading Support Through Eye Tracking. This project was conjointly conducted by all co-authors and resulted in a publication at ISWC'16 [80].

1.7 Thesis Outline

This thesis comprises eight chapters and is divided into six parts, the last two of which contain the bibliography and the appendix. The structure of the thesis closely follows the emergence of contributions as they are depicted in Figure 1.1. In the first part, we motivate the work, point out the greater vision of cognition-aware systems, their importance for information intake, state our contribution in this field, and describe the context, in which this research was carried out. The second part contains the research we conducted with regard to the quantification and detection of people's attentional states. Here, we describe the studies that lead to the presented tools and framework for building cognition-aware systems. In the third part, we apply these tools and concepts to applications to validate our approach and explore different application scenarios of cognition-aware systems with regard to information intake and learning. The fourth part contains an overall summary of the research contribution and discusses the overall approach and implications of this thesis.

Part I: Introduction and Background

Chapter 1 - Introduction The first chapter describes the motivation and vision for cognition-aware systems, states the context, in which this thesis was conducted, lists the research questions, and summarizes the key challenges and contribution we faced throughout our research.

Chapter 2 - Foundations In the second chapter, we introduce key concepts of cognition and memory that shaped this research. Also, we present the most relevant related work, which is mainly situated in the field of ubiquitous computing, learning applications, and context-aware systems. Finally, we briefly map out our approach and methods used.

Part II: Attention Research

Chapter 3 - Human Attention Attention is crucial for effective information intake and retention, which is why we first focus on how attention levels can be quantified, differentiated, and detected in a technology context. Therefore, we present a field study, in which we assessed states of engagement and boredom.

Chapter 4 - Context- and Content-awareness In this chapter, we focus on people's situational context and how to detect and support activities, during which people consume information. In a lab study, we investigate how content-awareness across devices can help enhance information intake and learning tasks.

Chapter 5 - Cognition-awareness People's ability to focus varies throughout the day. This chapter describes our approach to elicit users' circadian rhythms of alertness and cognitive performance. We develop and validate a mobile toolkit and propose a conceptual framework for building cognition-aware systems.

Part III: Applications

Chapter 6 - Opportune Content Delivery In this chapter, we apply the tools and concepts developed in our research to implement and test applications that identify opportune moments for content delivery throughout the day. Here, we specifically focus on delivering language learning and entertaining reading content.

Chapter 7 - Adaptive Reading Interfaces This chapter describes the concept of an application scenario for cognition-aware systems. By adjusting reading UIs in real-time we investigate the interplay between cognitive load and reading task efficiency. Therefore, we adapt speed reading techniques to reading on electronic devices and explore implicit control of reading flow through eye gaze tracking.

Part IV: Conclusion and Future Work

Chapter 8 - Conclusion and Future Work In this chapter, we summarize the findings and contribution of this thesis with regard to the research questions posed in the beginning. We conclude with a reflection on our approach, point out future work, and reflect on implications of cognition-aware systems.

Chapter 2

Foundations

The research presented in this thesis is rooted in the field of **ubiquitous computing** and **context-aware systems** in the discipline of computer science. To build cognition-aware systems and technologies that support information intake and learning, we apply concepts and theories form the field of **cognitive psychology** with a focus on perception, cognition, and memory.

2.1 Ubiquitous Computing

> *The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.*
>
> *Marc Weiser [268]*

Ubiquitous computing is commonly described as the third wave of computing. In the late 1950s, the era of mainframes began, where one computer was shared by several people. Only with the introduction of the *Altair 8800* by MITS in 1975, *personal computing* started becoming prevalent, in which one person was handling one computer. Nowadays, this has changed into a many-to-one relationship, where one user has a number of personal devices available, such as laptops, phones, tablets, watches, eye-wear, or other devices with processing capabilities. The ubiquity of these devices, that surrounds us marks the era of Ubiquitous Computing. Its very notion entails computing to appear anytime and anywhere.

As the inevitable consequence of the ubiquitous computing era, Weiser and Brown foresaw the coming age of *calm technology*: since we would soon be constantly surrounded by technology - in walls, chairs, clothing, light switches, and cars - it will be futile to ”get them out of the way” [269]. This calmness would then allow people to focus on being human. However, in an environment filled with ubiquitous computing devices, various appliances compete for our attention. This can have detrimental effects on our ability to focus. In the following, we will summarize previous work and challenges that arise from attention-seeking devices.

Interruptions and Disruptions

Interruptions generally occur when a person is detracted from a current primary task to another activity. Interruptions can occur in any setting, be it at work when focusing on a task, at home immersed in reading, or in transit when looking up the quickest route to a destination. Some interruptions only require temporary attention switches, while others—such as a colleague entering the office—can completely deter from the task at hand. This can go as far as forgetting about the resumption point of the primary task, which results in considerable time spent trying to get back into the previous task. In other cases, an interrupted task may even never be resumed at all.

Distractions or disruptions that lead to an interrupted task can be caused by external (*e.g.*, an incoming phone call) but also internal stimuli, *i.e.*, through self-interruptions. Self-interruptions occur in the absence of external triggers and are often a result of multitasking. While people who choose to self-interrupt were found to assess their productivity higher at the end of the day than those who get interrupted by notifications [181], frequent self-interruptions result in lower task accuracy [4]. A study by Iqbal *et al.* [143] found that it takes people up to 15 minutes to resume an activity after being interrupted by an incoming email or instant messaging notification. However, another study found that 64% of workplace interruptions are indeed beneficial [202], so they can not be dismissed easily.

Especially in the workplace, attentional states have been found to be related to mood and job performance. Mark *et al.* [180] investigated states, which make people more susceptible to distractions. Their results suggest that distractions depend on the user's current state of mind where certain attentional states precede their susceptibility to distracting stimuli. People immersed in rote work (*i.e.*, engaged but not challenged) are more likely to be distracted by Facebook or face-to-face interactions, whereas focused states are often followed by more time spent on email. They also found a connection between the number of application switches and prolonged communication behavior: the more people hopped between applications, the longer time they spent on Facebook and email. And the more task switches, the more opportunities presented themselves for this kind of distractions.

In mobile situations, additional environmental factors add to the list of possible distractions while performing a task. Oulasvirta *et al.* [207] found that on mobile devices users' task-directed attention can become fragmented into spans lasting only a few seconds. They proposed the *Resource Competition Framework* (RCF) to design for mobile user's limited cognitive resources. Building on Wickens' *Multiple Resource Theory* [273] their framework puts mobile task demands in relation to users' cognitive resources. Such a framework is useful to consult before designing applications that require user interactions in mobile settings.

Despite their distracting nature the proliferation of ubiquitous computing devices allows us to engage with information throughout the day and location-independently. Information consumption and processing are therefore deeply intertwined with our everyday life. However, we have limited time and resources to engage with the information available. There is simply too much information out, but furthermore, our cognitive resources are generally limited.

2.2 Cognitive Psychology

In this Section, we give a brief overview of the most important concepts with regard to mental processes as they are important for effective information intake and learning.

2.2.1 Perception

We perceive and process information from our environment through our five senses of sight, hearing, smell, taste, and touch. Information is passed in the form of signals in the nervous system resulting from physical or chemical stimulation from these senses. Perception does not only entail the passive receipt of these signals but is impacted by other cognitive functions, such as memory and attention.

Attention

> *Everyone knows what **attention** is. It is taking possession of the mind, in clear and vivid form, of one out of what seems several simultaneously possible objects or trains of thought. Focalization, the concentration of consciousness are of its essence. It implies a withdrawal from some things in order to deal effectively with others.*
>
> *William James [274]*

James' essential attribute of attention is the mind's ability to concentrate on one stimulus. Numerous sensations reach our sensors in every second we go about our day: we hear the voice of our conversation partner over the street noise coming in through the window. Meanwhile, we focus our eyes on our partner's facial expression and make out details about gestures and mimics that elude to the emotional aspect of what she is saying. All this is happening while our senses also take in the temperature in the room, the positioning of our arms on the table, our regular breathing, and incidental thoughts and associations coming to our mind. Each piece of this incoming information passes through a series of processing steps between its first perception to its processing and eventual consolidation in long-term memory. The stages of sensual perception to focused attention are connected to the three types of memory or memory functions: 1) sensory buffers, 2) short-term or working memory, and 3) long-term memory. Each function is discussed in detail by Baddeley [12]. In each of these stages, information is selected or discarded, so only a small fraction of the original scene makes it into our long-term memory [208]. Filtering is done through attention, where only the information is passed through that is of interest at a given time. Therefore, information that is not attended to, is lost. Schacter [234] called it absent-mindedness, where attention is focused on any other stimulus but one information. Hence, that information is not encoded correctly and later retrieval of it from memory is rendered impossible.

Attentional States

Mark *et al.* [182] presented a framework for classifying different attentional states in the context of work environments. They identified four distinct states, which can be aligned along the following two dimensions: *engagement* as the mental state of absorption in an activity [235] and *challenge* as the amount of mental effort being exerted to perform a task [131]. Figure 2.1 lists the four resulting areas where the upper right quadrant one describes '*focus*' as a state of high engagement and high challenge, the upper left quadrant two describes the attentional state of '*rote*', where high engagement but little challenge come together (*e.g.* mechanical type of work or thinking), the lower left quadrant three indicates the state of being '*bored*', *i.e.* low engagement and low challenge, and the lower right quadrant four shows situations of high challenge but low engagement lead to a state of '*frustration*'. In a study with 32 knowledge workers they found that their focus peaked during mid-afternoon (2-3 p.m.) and boredom was highest at the beginning of the day (9 a.m.) and early in the afternoon (1 p.m.). These attentional states were associated with certain emotions or valences. Here, participants were happiest during rote work and happy but also stressed when being focused.

Concepts related to attentional states include *cognitive absorption*, *mindfulness*, or *flow*. *Absorption* refers to when people feel deeply immersed in an activity, which comes with enjoyment, a feeling of control, curiosity, and with being oblivious to time passing [5]. *Mindfulness* is a psychological state with a strong focus on the present moment [57]. *Flow* is described as being totally immersed in an activity where "nothing else seems to matter" [52]. It refers to a pleasant but highly productive state, in which people feel challenged and stay within the boundaries of their skills.

Because perception and attention are vital for encoding and memorizing information, our memory is highly dependent on our attentional states.

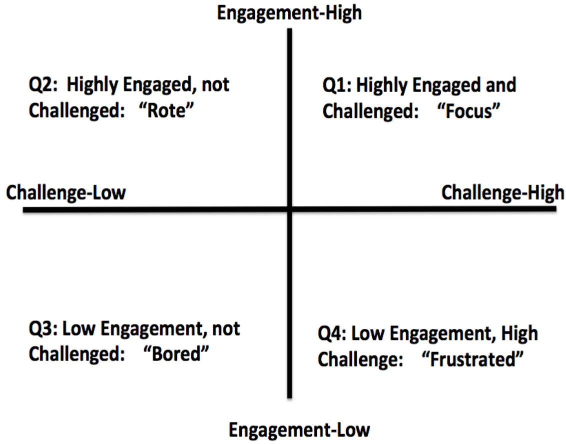


Figure 2.1: Theoretical framework of different attentional states by Mark *et al.* [182]

2.2.2 Memory

Neuroscience has made significant progress on revealing how the brain works, but there is still no complete understanding of how human memory functions. Information in the brain is stored in clusters of neurons, however, we do not know how, precisely, it is stored or encoded. In the following memory and its workings will be presented as it is most commonly and simply understood [12].

While being responsible for encoding, storing, and retrieving information, memory is also responsible for processing and acting on that data. Commonly, there are three main types of memory:

1. *Sensory Memory*: originating from sensory organs, such as eyes and ears, the brain receives this raw information and typically retains it for less than 500 milliseconds, but can store it for about 1 to 3 seconds. Sensory memory again is subdivided into iconic (visual), echoic (auditory), and haptic (touch) memory. Sensory memories decay quickly and are usually processed without conscious effort, which is also called preattentive processing (*i.e.*, prior to paying attention to the information). Here, only a subset of the information is made sense of, such as colors or shapes.

2. *Short-term Memory*: information is passed from sensory to short-term memory where it can be processed for up to several minutes, through rehearsal even up to several hours. Short-term memory is limited in its capacity; Miller found that between 5 and 9 items can generally be kept in short-term memory at a time [188]. This number can be increased, however, by a process called *chunking*, where items are grouped and remembered as chunks.

3. *Long-term Memory*: information in short-term memory is either thrown out or transferred to long-term memory where it is retained up to a lifetime. By rehearsing that information it becomes more likely that it will end up here. Emotions, meaningful, and multiple associations can further increase that chance.

A model of the brain and the distinction between implicit and explicit memory was by Larry Ryan Squire [249]. He categorized memory into two distinct parts: *implicit* and *explicit* or *declarative* memory, where implicit refers to motor skills which reflect in performance rather than through remembering (e.g., riding a bike or playing the guitar) and explicit refers to knowledge that we can describe and reflect on (e.g., knowing that eggs contain a high amount of protein). Endel Tulving further proposed the distinction in explicit memory between *episodic* and *semantic* memory [253] (see. Fig.2.2):

1. *Episodic Memory*: Episodic memory concerns personal experiences. Information stored here relate to a particular context, such as a time and place. "*I remember the day writing up my thesis while watching the kitesurfers go by,*" is a statement that resembles a personal memory, which can trigger associations to sensations, emotions, and other personal associations. Retrieval of information stored in the episodic memory system can also affect and modify the memory system.

2. *Semantic Memory*: Semantic memory refers to principles and, such as the shape of a guitar, and therefore holds abstract knowledge about the world independent of its context.

Information processing takes place in three stages: 1) receiving and encoding information through sensory systems, 2) storing that information in long-term memory and 3) retrieval of that information in the form of memories later on. Failure in one of these areas leads to what we call *forgetting*. With regard to knowledge acquisition and retention, forgetting is an undesirable trade, whose

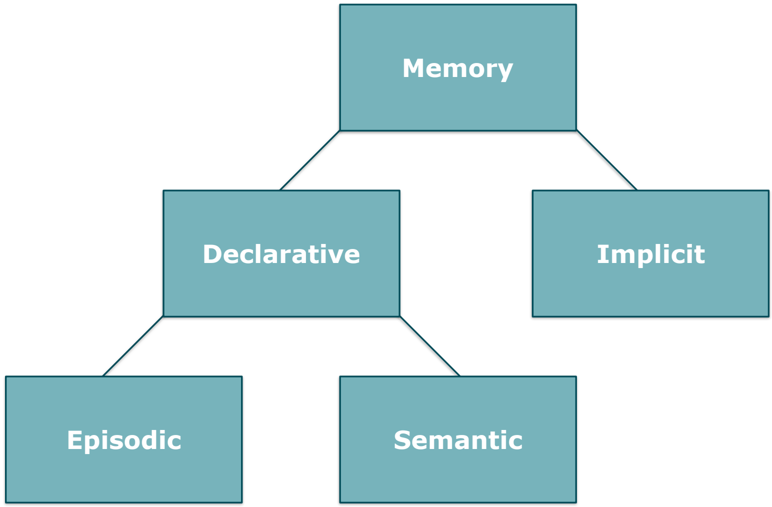


Figure 2.2: Types of memory and their relationship to each other.

root Schacter broke down into seven main causes [234], two of which are most relevant to our work:

1. *Transience*: information gets less accessible over time.
2. *Absent-mindedness*: due to a lack of attention, information when encountered is not encoded correctly and therefore later memory retrieval is being compromised.

Memory holds the information through which we understand the world. It, therefore, affects our perception and also our cognitive processes.

2.2.3 Cognition

Cognition has been defined as *those processes by which the sensory input is transformed, reduced, elaborated, stored, recovered, and used* [198]. It, therefore, entails the entirety of mental processing including incoming perceptual information, understanding, acquiring knowledge, memorizing, thought processing, retrieval, and thus the inner workings of our mind. Therefore, cognition comprises the more specific cognitive processes of perception, attention, and memory mentioned above, and further encompasses the use of existing knowledge and also the generation of new knowledge, hence learning.

Cognition and the workings of the mind have been studied as early as by the Greek philosophers, such as Aristotle. The term “cognition” goes back to the 15th century where it depicted “thinking and awareness” [228]. In the era of *behaviorism* (early 20th century) the focus of investigations lied on observable behavior. The mind was seen as a black box, where only input stimuli and output behavior could be observed and measured. Hence, in order to understand the mind’s workings psychology studies attempted to establish a relationship between inputs and outputs.

A paradigm shift occurred between 1950 and 1960 where *cognitivism* was introduced. In response to behaviorism and the introduction of the computer, cognitivism focused on mental processes. The mind was viewed as a computer where terms to describe internal processes were borrowed from the field of computer science (*e.g.*, memory, attentional bottleneck, perceptive capacity).

Another paradigm shift occurred from 1980 to 1990 with the rise of *neuroscience*. Computer usage had become commonplace in psychology and neuroimaging techniques had undergone significant improvements. Scientists were now able to focus on the brain’s activities. Since mental activity is a result of processes in the brain, neuroscientists focus on brain activities to gain insights into mental processes. Today, a variety of methods exist to measure electrical activity, some of which we present later in this chapter. These methods help us draw conclusions on underlying cognitive processes, including attention and learning.

2.3 Learning

Learning is generally defined as the acquisition of knowledge and skills through study, experience, or being taught. While there is a great number of learning theories about information intake, processing, and memorization, we will briefly focus on a few concepts most relevant to our work. Baddeley et al. [12] and Revlin [228] give an in-depth overview of the topic.

Learning activities are often grouped into two categories [196]:

- **intentional learning**: comprises activities specifically taken up for the purpose of learning.
- **incidental learning**: learning happens without explicit intent, such as while watching a movie with a foreign language character in from whom the viewer incidentally learns some phrases, for example.

Ubiquitous computing devices allow us to engage with information repetition anytime, anywhere. Cai *et al.* [38], for instance, created an Instant Message web application where users can learn words during wait times of an ongoing chat. Trusty & Troung [252] built a browser extension that replaces a selected set of English words with their foreign translation. Their examples show the feasibility of building learning applications with the specific goal of invoking incidental learning sessions throughout the day.

Ebbinghaus [90] famously investigated the nature of forgetting. Hence, the memory of information declines exponentially if no attempt is made to retain it. To prevent such rapid decline, repetitions need to be spaced out in a way that the information is encountered just when it is about to be forgotten [194]. Emotions and other associations can strengthen the memory of an information and therefore prolong the process of forgetting. Since people keep their mobile devices commonly at arm's reach for about 53% and within the same room for 88% of the time [65], these devices offer great potential for spacing out learning sessions throughout the day, week, and month.

Microlearning

Microlearning is the technique of breaking down a learning task into a series of quick learning interactions [20]. The, for example, daunting task of learning a foreign language could be broken down into numerous vocabulary learning units, grammar practices, pronunciation exercises, and so on. To yield optimal learning results, these units need not only be frequent but also spaced out in a favorable way [111]. Spaced repetition is crucial since people exhibit—as aforementioned—a negatively exponential forgetting curve [90], which is why repetitions need to occur at increasingly spaced intervals. By displaying learning items just as they are about to be forgotten we can make sure those items eventually end up and are accessible in long-term memory. Through microlearning these items become manageable and spaced repetition becomes feasible through the ubiquity of mobile devices.

Mobile devices provide a powerful platform for some forms of learning where individualization (or personalization) of learning content with ubiquitous access is critical [161]. The portability of the mobile device along with the low cost of retrieval makes them a great platform for microlearning. Ashbrook [9] demonstrated the importance of microinteractions, explaining that quick and easy access to a device is important for promoting frequent use. Prior to the ubiquity of smartphones, Cavus and Ibrahim sent SMS messages of English vocabulary words [42]. Edge *et al.* [92] created a mobile app that takes advantage of GPS location sensors to deliver contextual relevant vocabulary words. To ensure repeated exposure to a vocabulary word, Dearman & Truong [61] created a mobile app that displays the vocabulary word on the wallpaper of the phone. Especially vocabulary-based microlearning experiments have shown to improve vocabulary acquisition and recall [38, 92, 252]. Further, microlearning on a mobile device has accomplished high user acceptance [32, 42].

Concluding, mobile applications provide numerous opportunities to engage with learning content throughout the day. By spacing out learning content and repetitions, they can counteract the effects of forgetting and provide quick learning sessions in between activities through microlearning units. Incidental learning can further be fostered by learning in context: while mobile applications allow users to learn anytime and anywhere, consuming contextually relevant learning content can be triggered by the user's surrounding context [20].

2.4 Context-Aware Computing

A widely accepted definition of *context* was given by Anind Dey:

> ***Context** is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves.*
>
> *Anind Dey [64]*

By providing context information to computers Bowd *et al.* [3] state that we can effectively "*increase the richness of communication in human-computer interaction and make it possible to produce more useful computational services*".

In general, services support the user's task. Dey [64] describes such tasks in themselves as vital parts of context characterizing the user's situation: "*A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task.*" Dey further includes the user's emotional and mental state (*i.e.*, focus of attention) in his understanding of context [63]. Schmidt [238] regards the entity as part of a feedback loop: "*A change in the application will inevitably lead to a change in the context, perhaps as reaction to a changing situation.*" This reciprocal relationship implies that a change in the application can lead to a change in the context. Thus, such changes can be invoked in users as well.

A primary concern of context-awareness for mobile systems is awareness of the physical environment surrounding users and their devices. The key challenge of these systems is to record and make sense of contextual information. Advances in sensing, inferring, and using context information have been made by looking at different context dimensions, such as detecting people's locations [265], but also physical activities [243], and affective states [127]. Based on the users' location, for example, a navigation system can suggest routes to the nearest restaurant. Considering the time of day and lighting conditions, the device automatically adjusts the display's brightness to make content more readable. In order to enter the restaurant the user simply steps in front of the motion detector of the automatic door, which then slides open. These examples describe a network of context-aware systems that surround us in our daily life.

In their work Schmidt *et al.* [239] explore various aspects of usage context and propose a working model for context taking into account not only the physical environment, but also human factors, such as information about the user, her social environment, and current task at hand. For learning tasks, it has been proposed to consider various aspects of context to enable efficient acquisition of skills and knowledge. Merriam *et al.* [187] proclaimed that "*learning does not occur in a vacuum*", but is rather shaped by the context, culture, and tools. Wilson [275] proposed the notion of learning in context, also described as *situated cognition*. He emphasized the importance of interaction among people, tools, and context within a learning situation. Lave [171] conducted ethnographic studies on situated learning, in which she found that learning is a reoccurring process, in which learners act and interact within their social situations. She looked at how mathematical equations were applied and solved in the real world, especially during grocery shopping. There, grocery items, coupons, and social interactions with other shoppers and store workers built the learning context and helped the understanding of concepts. Thus, real-world context with social relationships and tools at hand have a positive impact on learning.

Such insights from contextual learning suggest that using information from the real world and taking advantage of devices in our environment as tools could foster learning. With the ubiquity of mobile devices and an increasing amount of content consumed on electronic devices, new learning scenarios arise.

2.5 Cognition-Awareness

Context-aware computing traditionally entails input dimensions, such as the user's location, physical activity, task, and device environment. In this Section, we focus on related work that considers the cognitive context as part of the user context. Such context comprises aspects related to mental information processing, such as attention, perception, memory, knowledge, or learning. Awareness of such processes can help systems to adjust to users in different ways:

1. Increase information bandwidth: phases of high attention and receptiveness can be utilized to push an optimized amount of content through a number of modality channels. Being highly alert allows people, for example, to monitor both visual and auditory channels and take advantage of simultaneous information streams.

2. Increase retention of information: by aligning information intake with phases of high concentration chances rise that this information makes it into long-term memory. Absent-mindedness is one of the main reasons why a piece of information is not retained [234]. By tracking attention phases and making information systems aware of these phases, they can adjust their presentation, information selection, and density accordingly. Detecting cognitive events, such as confusion, frustration, or realizations, such as "aha" moments, can help interface adjust in a way that they provide feedback or guidance on task focus, and suggest strategies in training scenarios. For learning, such supportive systems could increase understanding and retention.

3. Notification management: based on the user's cognitive states, interruptions can be deferred [201]. In phases of high user focus, the system could advise applications to refrain from interrupting the user. Email alerts and phone calls may be held back in such phases and the user's flow state can be preserved [52] to allow working in an effective manner.

4. Preserving mental health: stress occurs when there is a mismatch between task requirements and the user's cognitive capabilities [258]. Chronically high levels of mental load at the workplace can cause various health problems, such as stress, depression, or burnout. Circadian disruptions can further have devastating consequences for the emergence of schizophrenia [49] and diabetes [254].

In the following, we will give an overview of how cognitive states are commonly detected and what challenges arise from these techniques.

Detecting Cognitive Activities

To provide a "*window into our mind*" [261] technologies need to be able to sense and infer cognitive activities, which naturally take place inside of the user. Monitoring bio-signals with the help of sensors can give us indirect clues about different cognitive states. So-called *bio-sensors* can be highly specialized and are therefore only applicable under lab conditions, while others can be easily integrated into personal devices. Such sensors typically fall in one of two categories: invasive *vs.* non-invasive sensing, with invasive methods being mainly found in a medical context and hardly applicable in everyday consumer situations. Hence, we focus on non-invasive sensors with regard to their feasibility to inform context-aware systems.

In the following we briefly summarize the most commonly applied bio-sensors in HCI research:

Electroencephalography (EEG)

Brain cells communicate via electrical impulses which can be detected through electrodes placed on people's scalp. Since these impulses make up the brain's activity, they are commonly used in HCI especially for building Brain-Computer Interfaces (BCIs). From these measures, we can assess which brain regions are active during which kinds of tasks and also quantify the task workload. However, Electroencephalography (EEG) requires rather expensive equipment and is tedious to set up. Low-cost devices with basic functionality and a reasonable form factor, such as the *Emotiv*[7], are currently becoming more commercialized. BCIs are used in three major application areas [201]: (1) for interaction with

[7] `http://emotiv.com/`

computers with thought input, (2) for evaluating interfaces and systems (based on how hard someone has to work on a particular set of tasks), and (3) for adapting user interfaces according to brain activity measurements.

Functional Magnetic Resonance Imaging (fMRI)

Another way of detecting brain activity is using *functional neuroimaging*: during cerebral activity, neurons require nutrients to generate energy and produce action potentials. Glucose, oxygen and other substances are therefore transported to active neurons by means of blood perfusion. Hence, when a region of the brain is in use, blood flow to that area increases. Using a strong magnetic field Functional Magnetic Resonance Imaging (fMRIs) measure these changes in oxygenated blood flow and display activated regions accordingly. Activated regions allow us to infer the type of cognitive activity associated with them. However, fMRIs are bulky machines and therefore not applicable for being used in everyday mobile settings.

Functional Near-Infrared Spectroscopy (fNIR or fNIRS)

This technique is also based on the basic mechanism of functional neuroimaging. In contrast to magnetic resonance imaging, it uses light in the near-infrared spectrum to measure localized changes in oxygenated blood volume in the brain. Therefore, a ray of light is emitted at the scalp, half of which is absorbed by chromophores found in the nervous tissue. The compound of these chromophores and therefore their light absorption depends on the current saturation with oxygen and other nutrients (hemoglobin is a strong light absorber). A receiving photo detector captures the light wave resulting from the interaction with the chromophores. With respect to the original emitted, the characteristics of the light wave received differ due to the absorption in the nervous tissue. Methods like Functional Near-Infrared Spectroscopy (fNIR) present a non-invasive, low-risk method for studying cerebral processes. In contrast to fMRIs, fNIRs devices are less bulky and can be integrated into wearables.

Eye Movements

Sensing cognition is not necessarily limited to observing what happens in the brain. A number of physiological phenomena are caused by certain cognitive activities, which is why we can infer these activities by monitoring bio-signals that are not directly linked to the brain. Visual behavior, for example, is closely associated with cognitive processes, such as attention [177], relational memory [120], and learning [129]. There are two common ways of sensing eye movements, namely 1) via **video-based computer vision** techniques, and 2) via **Electrooculography (EOG)**. Video-based eye trackers generally use the eye's reflection of infrared illumination combined with high-resolution cameras and computer vision techniques for pupil detection and movement tracking. EOG, on the other hand, uses electrodes to measure changes in the electrical potential field surrounding the eyes, introduced by eye movements. Both techniques are capable of measuring the most important eye movement characteristics: fixations, saccades, and blinks. Early studies have shown the relationship between eye fixations and cognitive processes [149]. Others have focused on how saccadic eye movements and blinking are related to different levels of cognitive load [251]. Eye blinks can be further influenced by environmental factors, such as humidity, temperature or brightness, but also be an indicator of physical activity [35] and fatigue [237]. Rayner [223] investigated eye movements with regard to the underlying cognitive processes in reading and information processing. Other works have looked at inferring reader engagement, which is closely linked to attention [168]. Bulling *et al.* [33, 34] introduced eye movement analysis as a new modality for activity and recall recognition.

Physiological Sensors for Affective Computing

Although not necessarily directly related to cognitive activities, a range of physiological sensors can provide hints about a person's affect, fatigue, or stress levels. In the following, we will briefly discuss the ones that are most commonly used in HCI research and why they may make a valuable addition to cognition-aware systems.

Early studies found that systematic changes in **body temperature** correlate with diurnal variations in cognitive performance [157]. Spontaneous or induced changes in body temperature seem to have an inverse effect on reaction times, where an increase in body temperature leads to a decrease in reaction time. Kleitman [158] based his findings on metabolic activity of the cells of the cerebral cortex and suggested that by increasing body temperature, thought processes could be indirectly sped up as well: he assessed people's performance on card sorting, mirror drawing, code transcription, and multiplication speed. Wever [272] later suggested that those diurnal rhythms were related to the circadian system in humans. Influence of body temperature and internal biological time were found to affect performance and alertness [190]. Since changes in body temperature can be very subtle, measurements were required through a rectal sensor. Later in this Chapter, we will propose a less invasive technique to gather information on performance and the circadian rhythm associated.

One of the most indicative measures for a person's physiological state is the heart activity and performance. The heart rate is the number of times the heart contracts in one minute and is usually reported in Beats per Minute (BPM). It can be measured through a variety of methods: manually (*i.e.*, by feeling a person's pulse on an artery), with a photocell sensor (*e.g.*, photoreflectance or infrared sensor monitor), through video footage and video-processing algorithms [280], or by measuring the heart's activation pulse through electrodes placed on the body (*i.e.*, **Electrocardiography (ECG)**). If we take measures over a period of time we can calculate a person's **Heart Rate Variability (HRV)**, which describes changes in heart performance indicating how calm, excited, or exhausted that person is. The variability is often used to assess stress levels and mental load. Sensors can be placed on different body parts, such as on fingers, wrists, or on the chest, and is therefore feasible to be integrated into wearable devices. Similarly, electrodes placed along the body can also be used to measure electrical activity in muscles and assess muscle tension (**Electromyography (EMG)**), which elicits symptoms of excitement or stress.

Similar to heart rate, a person's **respiratory rate** can be a measure of the current physiological state. It counts the breaths a person takes per minute, which can differ across age groups and according to the general current health state (the respiratory rate for a healthy adult at rest ranges between 12 and 20 breaths per minute). Measured across a period of time the respiratory rate variability can be an indication of stress levels. Respiration can be estimated from wearable sensor data, such as from ECG, but also elicited from video footage [280] or accelerometer data.

Another strong indicator for stress or excitement levels is a person's **Electrodermal Activity (EDA)** or skin conductance, which describes the ability of the skin to conduct an electrical current. The sympathetic nervous system basically controls the amount of sweat emitted, which causes variations in skin conductance. Sweat glands can be measured and used as an indication for psychological or physiological arousal: if a person gets stressed, for example, the autonomic nervous system becomes aroused and causes an increased sweat gland activity, which in turn increases skin conductance. This technique is often used in lie-detectors. Skin conductance is also described as **Galvanic Skin Response (GSR)** and is fairly straight-forward to measure on different body parts. Hence, GSR sensors are feasible to be integrated into wearable devices and can, therefore, be used to assess emotional and sympathetic responses throughout the day.

Self-reports

Although not being a sensor in the literal sense, a lot of information about inner user state can be easily elicited by simply letting users assess their own situation. Daily diaries, for example, are used as a research methodology to make people describe their impressions, feelings, and general environment. For more in-situ feedback, **experience sampling** has been proposed and advanced with the ubiquity of mobile devices and the possibility to record short statements throughout the day [50]. Hence, participants are asked to stop at certain times during the day and report on an experience in real-time. A convenient tool to assess cognitive load, which is more often used in lab settings, is the **NASA Task Load Index (NASA-TLX)** [124]. In its original form, the NASA-TLX questionnaire contains a workload assessment on six subscales with regard to mental demand, physical demand, temporal demand, performance, effort, and frustration. These subscales are then individually weighted against each other by the participant based on their perceived importance. However, for most cases, the assessment on the subscales is sufficient though and also individual subscales can be dropped if they are less relevant to the task at hand. In these cases, the test is generally referred to as "*Raw TLX*" [123].

In-situ Elicitation of Cognitive States

As some of these cognitive assessment techniques show there are crucial challenges in trying to obtain the cognitive context in an unobtrusive manner. On top of that, the neural dynamics of the brain are generally complex processes, which are hardly accessible for non-invasive measurement techniques. Most of the sensors described above need to be attached either to the human body (*e.g.*, EEG) or be placed in the environment (*e.g.*, video cameras). In many cases such devices are bulky, which limits their portability and therefore their applicability in a mobile context. Another factor to consider is social acceptance. An fNIR sensor can be integrated into a headband placed on a person's forehead, but that does not necessarily make for an aesthetic appearance. Meanwhile, other sensors, such as rectal temperature sensors, are simply uncomfortable to wear throughout the day. An alternative approach is to associate observable activities with cognitive processes. Here, ubiquitous computing devices, such as the smartphone, can be used to track a wide range of activities. Such devices are also in close proximity to their users for most time of the day and therefore can deliver quite an in-situ picture of a person's daily doings. In our work, we investigate the use of ubiquitous computing devices to assess cognitive states on-the-fly by analyzing phone usage data. We show how such assessments can be made within a relatively short time window. More regular patterns of attention are exhibited by diurnal fluctuations of alertness—so-called circadian rhythms.

Circadian Computing

People's alertness, attention, and vigilance are highly variable and subject to systematic changes across the day [112, 156]. These fluctuations—in part caused by *circadian rhythms*—impact higher level cognitive capacities, including perception, memory, and executive functions. During phases of high alertness, we are able to perform tasks efficiently while during phases of low alertness we have trouble concentrating.

According to the prevalent theory on sleep/wake regulation, variations in alertness and sleep propensity are generated by two underlying processes: the sleep/wake homeostasis and a circadian process [28]. The homeostatic process manifests itself as a gradual decrease in alertness during wake periods. The longer we are awake, the stronger becomes the need for sleep. Alertness is further modulated by a circadian biological clock with a period length of about 24 hours. Following a sinusoidal pattern, it determines hours of the day when we experience a particularly low or particularly strong sleep drive. For many people, the alerting capability of the circadian process peaks in the late afternoon, thus partially counterbalancing the accumulated sleep pressure from the homeostatic process. This is commonly experienced as heightened alertness towards the evening after a post-lunch dip in alertness and concentration.

These rhythms can be different from person to person, however, they occur in individual, but distinct patterns. Figure 2.3 shows an exemplary curve with phases of high attention in the morning, a decline in the early and a performance recovery in the late afternoon, for example. The circadian rhythm of alertness and cognitive performance depends on a range of individual factors, such as sleep, nutrition, stress levels, or general health. Traditional methods to assess the circadian rhythm include extensive lab experiments, which can take weeks of being in controlled environments. Other methods can be equally cumbersome or even unpleasant, such as sleep-wake protocols or physiological markers (e.g., dim light melatonin onset, rectal temperature monitoring, cortisol level measurements [135, 158]).

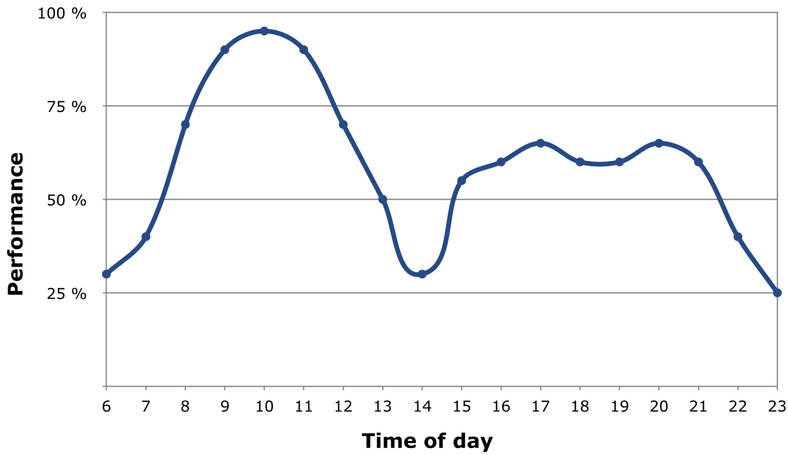


Figure 2.3: The circadian rhythm describes systematic performance changes across the day showing in our ability to concentrate and focus.

Current systems do not adapt to individuals' variations in sleep/wake cycles and the related diurnal patterns of alertness and cognitive performance. Instead, systems generally assume a constant level of cognitive performance and rarely accommodate for variations. In an attempt to reconstruct this rhythm in a more convenient and externally valid way, we will focus in this thesis on mobile devices and their capabilities to collect alertness data in-the-wild.

2.6 Methodology

In this Section, we describe our approach to build, apply, and evaluate cognition-aware systems. In each step, we follow the user-centered research approach [88] by conducting a combination of formative studies, lab, and field experiments with people using our research probes being at the center of this process.

2.6.1 Quantification and Analysis of Attention

At the core of information perception, processing, and memorization lie the ability to focus and pay attention. Because mobile devices, such as smartphones, are near-constantly available and have rich sensor capabilities, we start by measuring people's attentiveness towards these devices. To capture phone usage and measures of attentiveness throughout the day we conduct field studies because this data needs to be collected in-the-wild, where participants follow their everyday schedules and are not confined to an artificial experiment setting. Therefore, we release our data collectors as mobile apps on Google Play and invite participants to download and install them. Informed consent is given through the app and the collected data is securely transmitted to our servers.

Besides phone usage data we collect subjective user assessments in the form of Experience Sampling. Thus, the app prompts users throughout the day to fill in a quick survey about their current inclination towards boredom and other states. Again, these measures need to be collected in-the-wild, since we are interested in these assessments in the context of participants' everyday life.

2.6.2 Content-awareness between Devices

With the hypothesis that not only the attentive state but also knowledge about the content that is currently being consumed by the user, can help to facilitate information intake, we investigate ways to enhance information intake through content-awareness. Since TV consumption is a major form of information acquisition, we investigate the use of second-screen apps to enrich information intake. To elicit app features we first conduct an online survey, especially because it has the potential to reach a large number of people from different geographical regions to find out about TV watching habits and current use of second-screen apps. We further consulted the literature regarding this topic and combine our results to elicit a feature backlog for the eventual application. Therefore, we develop a system that harvests subtitles from TV programs in real-time to automatically create additional content. We needed to develop a system that records streams of subtitles, performs keyword extractions, and gathers additional relevant content through entity linking. Because the utility of this tool is highly dependent on the quality of the content generation, we evaluate the system with a pre-recorded, but real-world TV program. However, because we compare the use of the app to the traditional way of looking up information online, we conduct a controlled lab study with a repeated-measure design to limit the possibility of confounding variables, such as program content differences and user distractions. Measures are not limited to the relevancy of the generated content but include subjective user impressions and objective comprehension tests.

2.6.3 Tools for Building Cognition-Aware Systems

While we investigate the feasibility of analyzing the user's momentary attentiveness based on phone sensor data, there are also patterns in attention differences determined by people's internal body clocks. Knowledge of these diurnal patterns could inform cognition-aware systems about re-occurring productive phases across the day. Because these patterns occur in people's everyday life we need to deploy our research probe in the field. Hence, we install our apparatus on people's phone and collect quantitative alertness measures throughout the day. We further trigger surveys about the recent consumption of caffeinated drinks to be able to account for the influence of caffeine on the alertness measures. We conclude the study with a semi-structured interview where we ask participants to rate the different app tasks according to their subjective impression regarding likeability measures.

2.6.4 Proof-of-Concept Applications

To validate the tools developed in the first part of this thesis we integrate them into two applications which suggest different types of content (learning and reading) in moments classified as opportune (*i.e.*, bored). Because validation entails external validity we deploy our prototypes in the field. To collect more detailed feedback we recruit a number of participants to download and use these apps while collecting usage data and conducting interviews in between. Since we set out to train machine learning models on collected ground truth data we need to collect big amounts of phone usage data in order to get reliable results. Hence, we release the app on Google Play to the public, hence launch an in-the-wild experiment.

In another application scenario, we envision cognition-aware systems to also be able to align their interfaces with the users' current cognitive capacities. To explore potential interfaces for such adjustments we created dynamic reading UIs with the ability to superimpose variable reading speeds on the user. To measure cognitive load and effects on eye movements we use the NASA TLX questionnaire and a stationary eye tracker. Also, we measure text comprehension and subjective user assessments. To avoid distractions during the experiment we conduct a controlled lab study, where participants come in and read with the different reading UIs. To avoid confounds through differing reading skill or text difficulty, we employ a repeated-measure design and counterbalance not only the two conditions (*i.e.*, the reading stimuli) but also the allocation of text to the condition. Further, to make sure comprehension questions are validated, we use texts from an official corpus used by foreign language speakers to be tested on reading comprehension.

For the last study reported in this thesis, we investigate differences in implicit *vs.* explicit controls for adaptive reading UIs. Therefore we created an apparatus that uses eye gaze tracking to determine whether a user looks at a smartwatch to read or not. To test our two conditions we conduct a repeated-measure lab study where people wear the eye tracker and are asked to read different texts throughout the study. Again, we are interested in differences in text comprehension, which is why we need to control for confounding variables, such as distractions from the environment and text bias. However, because possible applications for implicit control of reading flow through eye tracking include a reading-while-walking scenario we introduce an artificial secondary task whose goal it is to introduce controlled reading interruptions.

Concluding, the first part of the thesis resembles our approach to build technologies for assessing and predicting cognitive states. Because these states occur mostly throughout the day and outside of the confinements of our lab, we conduct these studies in the field. The last part of the study comprises the application of these technologies for validation purposes and for investigating their feasibility and utility. Here, we apply a mixed method approach of controlled lab studies, controlled user groups as well as public experiments in-the-wild.

II

Attention Research

Chapter 3

Human Attention

Attention is generally described as the ability to concentrate on a specific stimulus while ignoring other perceivable information. It shows in our skill to focus on a conversation, a thought or a piece of text while blending out background noises, tactile impressions, and visual stimuli in our periphery. Hence, attention in itself is the process of selecting one over a range of competing stimuli [114].

Despite being focused on one specific channel of information we are still able to process competing channels in the background. Cherry [45] first defined the *Cocktail-Party Effect* as the ability to switch attention between multiple auditory streams, which allows people to 'tune into' one channel while 'tuning out' all others. The effect gained its name from the scenario of a cocktail party where people are focused on a single conversation while being able to make out the mentioning one's name in a nearby conversation.

We live in a world full of competing stimuli. In addition to our current physical environment, we are subject to a range of technologies that constantly try to grab our attention. The sheer amount of devices and services providing access to information as well as making requests for attention poses challenges to our cognitive processing capacities. In our current age of information technology information is produced and disseminated more quickly and more broadly than ever leading to what has become known as *'Information Overload'*. The term first introduced in 1964 by social scientist Bertram Gross [118] describes an effect where the amount of input exceeds the processing capacity of an actor. Due to the limited cognitive capability of a decision maker, for example, information overload can have detrimental effects on the quality of decisions [247].

However, we are drawn to our devices. Mobile phones, smartwatches, tablets, laptops, or connected TVs provide stimulation which is constantly available, even while on the go. Such devices are notorious for trying to grab our attention through various alerts in form of auditory beeps, visual notifications, and haptic vibrations. But even when we are not interrupted, according to Nielsen, 45% of our smartphone usage time is devoted to self-stimulation by engaging with news, entertainment, games, and social media[8]. This conditioning goes as far as that being alone with one's own thoughts for longer periods of time has been shown to cause aversive reactions. Being deterred from external stimulation has even lead participants in a study to rather self-administer electric shocks than to "*just think*" [276]. Technology companies profit from and actively encourage this conditioning since they monetize on users' attention. Hence, in today's so-called '*attention economy'* user attention has become a scarce resource [59].

With all these competing stimuli from services, technologies, and our environment it becomes increasingly challenging to focus on the right kind of stimulus. The ability to focus is crucial for memory tasks. A lack of attention while being exposed to incoming information is one of the main reasons for a bad memory, because information is perceived, but can not be committed to long-term memory [234]. Attention is influenced by a range of factors, such as the current environment, motivation, but also physical, cognitive, emotional, and social conditions [271].

In this Chapter, we investigate the nature of attention in a technology context as it is a crucial component for effective information intake. We first set out to understand how attention can be quantified in a mobile context. We further present an approach to identify idle, but attentive states. Therefore, we built a system that automatically detects moments, in which people actively seek out stimulation.

The two research questions we target with our investigation in attention are:

- **RQ1**: How can users' attentiveness be quantified across the day and reliably predicted from phone usage patterns?
- **RQ2**: Does boredom measurably affect phone usage and which usage characteristics are most prevalent during such states?

[8] `http://www.nielsen.com/us/en/insights/news/2014/how-smartphones-are-changing-consumers-daily-routines-around-the-globe.html`

Parts of this Chapter are based on the following publications:

- T. Dingler and M. Pielot. 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*, MobileHCI '15, pages 1–5, New York, NY, USA, 2015. ACM
- M. Pielot, T. Dingler, J. S. Pedro, and N. Oliver. 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, pages 825–836, New York, NY, USA, 2015. ACM

3.1 Related Work

The work described in this Chapter is based on principles of perception and attention as described in Section 2.2, and more specifically on interruption management and attentiveness towards devices.

3.1.1 Attention to Devices

In ubiquitous computing environments we are surrounded by a variety of devices: public displays, desktop computers, laptops, mobile phones, and increasingly wearable devices, such as smart watches and eyewear. Throughout the day we use these technologies to communicate, access, create and disseminate information, be it in the office, at home, or on the go. The consequence of constantly being surrounded by such devices is that they compete for our attention: we continually switch our attention between different devices and sources of information while doing different types of tasks. Especially the activity of information workers is characterized by frequent attention switches throughout the day.

A study by Dey *et al.* [65] found that people keep their smartphones within arm's reach for about 53% and within the same room for 88% of the time. On daily basis, users interact with their phones on average 58 times where 60% of these interactions constitute to quick glances since the phone remains locked [133]. A study by Böhmer *et al.* [27] found that users spend about an hour a day on their phone just using apps. Meanwhile, they use an app on average for less than 72 seconds at a time. Ferreira *et al.* [97] report that approximately 40% of application launches last less than 15 seconds showing that interaction happens in brief bursts of *micro-usages*. Such usage behavior shows the highly fragmented attention users pay to their devices.

Interruption Management

Interruptions negatively affect people's focus as they find it difficult to return to the activity prior to a distraction [56, 141]. Depending on the current task condition, people may either ignore an interruption, delay attending or immediately turn to it. Interruptions occur in our salient physical environment but increasingly emanate from our digital devices, especially in the form of alerts and notifications [55].

Hence, different notification-delivery strategies have been proposed to minimize the impact of interruptions: Horvitz *et al.* [137] proposed *bounded deferral*: if a user is predicted to be busy, alerts are being held back until a more suitable moment, but only for a maximum amount of time. In the context of mobile phones, Fischer *et al.* [99] found that *opportune moments* for delivering notifications occur right after the user has finished a task, such as writing a message. Previous work [209, 215, 216] has explored the use of mobile phone sensors and usage patterns, such as the user's location or recency of interactions, to automatically predict such opportune moments. Rosenthal *et al.* [230] created a classifier that learned from the user's behavior when to mute the phone for incoming calls or notifications in order to avoid embarrassing interruptions. After a learning period lasting two weeks, the application applied the learned preferences and changed the ringer volume proactively based on the trained classifier.

However, bounded-deferral strategies may not work if there are many long phases without opportune moments. It can be considered an ideal strategy only when users are typically attentive, and when phases of inattentiveness are brief. To see whether this is the case, we conducted the first study described in this chapter to determine the attentiveness of mobile phone users throughout the day. We then investigated the characteristics of opportune moments and how to identify them, which brought us to the notion of boredom.

Boredom and Stimuli-Seeking Moments

Fenichel [95] described boredom as a displeasure caused by a "*lack of stimulation or inability to be stimulated thereto*". It often comes with a "*pervasive lack of* *interest and difficulty concentrating on the current activity*" [100]. Eastwood [89] highlights that "*a bored person is not just someone who does not have anything to do; it's someone who is actively looking for stimulation but it is unable to do so*". When perceived as an undesirable state, boredom can instill an urge to escape the current situation [113]. This urge can cause people to actively seek out stimulation. Detecting these situations could lead us to such opportune moments, in which interruptions might be welcomed by users.

Boredom detection has been the focus of previous work. Bixler and D'Mello [24], for example, assessed boredom during writing tasks by logging keystrokes. They collected affect judgments from participants, in which boredom was named in 26.4% of the cases, second most often after engagement (35.4%). They were able to distinguish engagement-neutral and boredom-neutral states by looking at keystrokes together with participants' *stable traits*. Guo *et al.* [119] connected users' web activities to their general susceptibility to distractions. They found mouse movements, clicks, page scrolls, and other fine-grained interaction events to be indicators for boredom. Mark *et al.* [182] further showed how the type of digital activity can be a strong predictor for the attentional state of a person, such as being focused or bored. They found that boredom was related to frequent window switches, surfing the Web, and the time of day. People using Facebook generally did not report being in a focused state. However, online communication activities, such as Facebook visits, can function as a quick break when people are engaged in work [179]. These brief moments of often self-induced distraction can help people to maintain *emotional homeostasis* while multitasking [180]. Homeostasis defined as the ability of an organism to maintain an internal equilibrium is challenged as people get stressed, bored or frustrated. Hence, the attempt to move to another activity is aimed at getting back to a balanced state. So people sometimes actively seek out distractions even during work. Such distractions are considered low cost in terms of cognitive resources required, they promise to be fun, and leave the user in control over the duration of the interaction [179]. The day of the week also seems to play a role when it comes to bored and focused states. Here, Mark *et al.* [182] found that people were both most focused and bored on Mondays compared to the rest of the week. Most rote work was done on Thursdays.

Other work investigated affect detection based on device usage. Mobile social interactions and application usage seem to be closely linked to people's mood. LiKamWa *et al.* [175] inferred mood (valence and arousal) from SMS, email, and call interactions, as well as from applications routinely used. Bogomolov *et al.* derived daily happiness [26] and stress [25] based on mobile phone usage, personality traits, and weather data. As formulated by Marks *et al.* [180] certain attentional states are more susceptible to distractions than others. This openness to interruptions was further shown to be related to the following device usage patterns:

- time since recent device usage [10, 102]
- use of mobile phone messengers and the phone's notification center [213]
- activity levels, such as switching windows [142, 182]
- use of keyboard and mouse [102, 142]
- ambient noise level as a proxy for activity levels around the user [102]
- semantic location (*i.e.* home or work) [211, 230]
- ringer mode as an indicator for how people want to manage interruptions [211, 230]
- time-dependent variables, such as the hour of the day or the day of the week [10, 102, 138]
- proximity as an indicator for whether the phone's screen is covered (*i.e.*, stowed away) or not [211, 213].

Previous work shows how attentional states, openness to interruption, and boredom measurably affect the way people pay attention to their devices. In this Chapter, we present two studies looking at how attention can be quantified across the day and how we can identify such moments, in which users turn to their device in search for stimulation. By understanding attention better in a technology context we aim to inform the design of attention-aware systems that adjust information delivery to the user's current state. By knowing about opportune moments to deliver content, interruptions can be mitigated and information is presented at times when the recipient's attention allows for active perception and processing of that information.

3.2 Quantifying Attention

Mobile devices, applications, and services require and attract people's attention. But attention is generally limited and therefore valuable. Hence, in today's attention economy human attention is treated as a scarce commodity [59]. We set out to investigate how scarce it really is in a technology context and whether it can be assessed in a quantified manner. Therefore, we looked at people's attentiveness to their mobile phone across the day, and more specifically at their mobile messaging behavior.

3.2.1 Motivation

People predominantly use their mobile phones for communication purposes by sending out SMS and using various types of messengers. In a 2011 survey, teenagers were found to exchange a median number of 60 messages per day [174]. Due to these frequent interactions, people tend to expect responses to their messages within minutes [47]. To meet these expectations, people need to be *attentive* to their phones, which means checking and triaging new messages quickly upon arrival. Usually, people do so within a few minutes [19, 215, 232], but such behavior ends up often interrupting current activities. Such interruptions can exhibit negative effects, as it can be difficult to return to an activity prior to the interruption [56, 141]. Previous work has, therefore, proposed different notification-delivery strategies, such as *bounded deferral* [137], through which alerts are not delivered immediately, but rather delayed until a more opportune moment. These strategies only work, however, if sufficient opportune moments present themselves throughout the day. If messages are delayed for too long, the message content may become irrelevant and response times will increase, which in turn may violate expectations. If the maximum delay, on the other hand, is too short, messages may be delivered too frequently leading again to interruptions. Bounded deferral will, therefore, only be a viable strategy when users are generally attentive, and when phases of inattentiveness are kept brief.

To determine general attentiveness of smartphone users throughout the day, we conducted a study during which we collected phone-usage data from 42 mobile phone over the course of two weeks.

3.2.2 User Study

To be able to predict people's general attentiveness to mobile messages, we conducted a study in which we assessed how quickly users triaged incoming messages throughout the day. We, therefore, released a mobile app in 2014, through which we collected ground truth on the arrival time of messages and the time until users attended to incoming messages. Additionally, the app recorded contextual and phone usage data. Over the course of two consecutive weeks, we collected both ground truth and contextual data in order to train a machine-learning algorithm, with which we were able to determine users' attentiveness to mobile messaging throughout the day. From the context data collected, we extracted 16 features, with which we trained our model. For each minute of the day, we predicted how quickly a user would attend to a message by applying the model. Thus, the model helped us fill the gaps for all those moments where users did not actually receive any messages.

Participants

Through university mailing lists and from a study participant pool of Telefónica R&D in Spain, we recruited 42 participants (45.2% male, 23.8% female, 31% did not report their gender) who reported their age to be on average 28.7 years ($SD = 5.9$). The majority of participants was living in Europe based on the devices' language and timezone settings.

Data Collection

Our Android app was designed as a background service that ran on participants' phones and collected data from sensor and phone events, such as the status of the screen (on/off), data from the proximity sensor, access to the notification center, the ringer mode, the app in the foreground, and incoming notifications. Additionally, the app recorded Whenever an incoming message triggered a notification and when a notification was removed, including instances, in which messages were read on another device. We focused on notifications from messenger applications filtering out other types of notifications. Notifications that came in while the related app was already in the foreground, were ignored as well. The ground truth instances and the contextual data was recorded, stored locally, and sent to a server whenever a WIFI connection was available.

Procedure

After downloading and installing the app from *Google Play* participants were asked to sign an informed consent form when the app was first launched. In it, we provided information on the study's background and on the type of information being collected. The app then instructed participants how to explicitly grant access to notifications data. For two consecutive weeks of data collection, we compensated participants with a 20 EUR Amazon gift card.

Results

We collected a total of 55,824 messages from 42 participants. Figure 3.1 gives an overview of the total number of messages received during the different hours of the day.

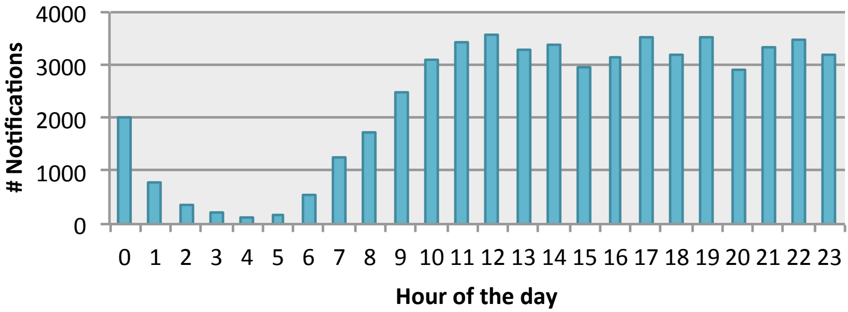


Figure 3.1: Total number of notifications per hour of the day ('0' indicates notifications arrived between 0:00 and 1:00 o'clock, for example).

Participants received on average 66.8 ($Mdn = 40$) messages per day. Figure 3.2 shows the distribution between different messenger apps. With 77.7% of the messages, WhatsApp dominated other messenger apps, while only 1.8% of received messages constituted text messages. In Europe, WhatsApp has mostly substituted SMS messaging.

Attentiveness

Responsiveness is merely an indicator of when users actually respond to an incoming message. Attentiveness, on the other hand, includes the user being aware of a message and having decided whether to act on or ignore it at the current time, a behavior also known as *triaging*. Because receivers' actual responses also depend on factors, such as the sender-receiver relationship, or the importance

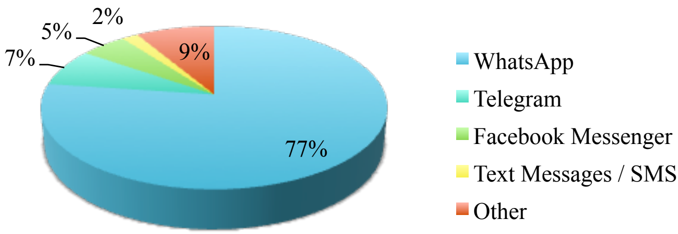


Figure 3.2: Messages by messenger service.

and urgency of the content, we focused on modeling attentiveness rather than responsiveness. Incoming messages could, therefore, be attended in three ways:

1. by checking the notification center, which shows sender and excerpt of the message
2. by opening the corresponding messenger application and
3. by reading the message on a different device (*e.g.*, WhatsApp messages can be read in a browser window rather than on the phone).

The dataset we collected contained 38,180 (68.4%) messages that were first attended to through the notification center and 14,134 (25.3%) that were attended to through another device. The remaining 3,510 (6.3%) messages were first attended to by opening the app they belonged to. Our participants attended to messages within a median time of 2.08 minutes, 25% of the messages within 12.0 seconds, 75% within 12.3 minutes, and 95% within 80.0 minutes.

We analyzed the time within which messages were attended to depending on whether this was done through the notification center, the respective app, or through another device using a Kruskal-Wallis test. The test revealed a statistically significant difference ($\mathscr{X}^2(2) = 2505.139, p < 0.001$). Pairwise Mann-Whitney tests (Bonferroni-corrected) showed that messages were attended to quicker through the app ($Mdn = 0.47$ min, $p < .001$) or another device ($Mdn = 0.75$ min, $p < .001$) than through the notification center ($Mdn = 3.2$ min).

3.2.3 Phone-Usage Patterns

To predict people's attentiveness to mobile messaging throughout the day, we created a model similar to the one presented by Pielot *et al.* [215]. The model uses 16 features based on the mobile phone sensor and event data: the screen status (on/off) and when this status last changed, the status of the proximity sensor (screen covered/not covered) and when it last changed, the time since the phone was last unlocked, the time since the last message arrived, the number of pending messages, the time since the user last opened the notification center, the hour of the day, the day of the week, and the ringer mode.

We used the median delay (2.08 min.) between arrival and attendance of a message for classifying attentiveness in a binary manner. We, therefore, labeled users *attentive* when they triaged messages within these 2.08 minutes, otherwise we labeled them as *non-attentive*. We then trained a *Random Forest* resulting in a 79.29% accuracy and $\kappa = .586$. Precision and recall for being attentive were .771 and .828 respectively.

Attentiveness Throughout the Day

We applied the trained model to computationally estimate the times that people were attentive throughout the day, minute by minute. We, therefore, went through all sensor data and computed the state of each of the features for the beginning of each minute of the day, which constituted 86,400 states per day. We then applied the classifier for each state to predict participants' attentiveness.

The model predicted participants to be *attentive* to messages on average for 50.5% ($SD = 14.6\%$) of the full 24-hours of the day. The quartiles were 40% ($1^{st}Q$), 49% (median), and 55% ($3^{rd}Q$). The majority of participants, therefore, attended to messages within 2 minutes for 12.1 hours per day or 84.8 hours per week, which corresponds to 75.8% of the hours typically spent awake (if we assume an average of 8 hours of sleep).

Figure 3.3 depicts the average attentiveness throughout all days of the week. We found a statistically significant difference between the days of the week ($F(6,20802) = 41.07, p < .001$). Bonferroni-corrected pair-wise t-tests showed that there were statistically significant differences between the weekdays and the days of the weekend ($p < .001$): participants were, therefore, significantly more attentive to incoming messages during the week (Mon - Fri) than during weekends (Sat, Sun). During the week, participants were predicted to be attentive 62%-67%

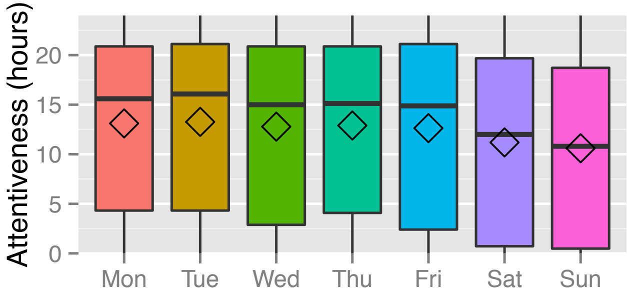


Figure 3.3: Average attentiveness by day. Dashes indicate the median, diamonds the mean levels of attentiveness. A value of, *e.g.*, 12 indicates that, on average, users are attentive to messages during 12 hours of the given day.

of the day, whereas on the weekend these numbers dropped to 45%-50%. The average attentiveness for each hour of the day is shown in Figure 3.4. The median predicted attentiveness ranges from 0% at 4:00 am to a maximum of 83% at 09:00 pm. Tests revealed a statistically significant differences ($F(23, 20875) = 189.6, p < .001$) as well: We found attentiveness to be highest during the evening, *i.e.*, between 6:00 and 9:00 pm. With a median attentiveness of at least 80%, this was significantly higher than during the rest of the day (pair-wise comparisons yielding at least $p < .01$, Bonferroni-corrected). Further, we found a statistically significant difference between night time (0:00 - 8:00 am) and day time (10:00 am - 11:00 pm) (all pair-wise comparisons yielding at least $p < .001$, Bonferroni-corrected). During nights and early mornings, median attentiveness was always below 50%, whereas during the day, median attentiveness was always above 67%.

We then looked at the duration of periods in which participants were predicted not to be attentive to messages, *i.e.* inattentive. Calculating the quartiles, participants were predicted to be back in an attentive state after 1, 2, and 5 minutes, in 25%, 50%, and 75% of the cases respectively. When entering a state of being inattentive to messages, participants, therefore, returned to a state of attentiveness most of the time after only a few minutes.

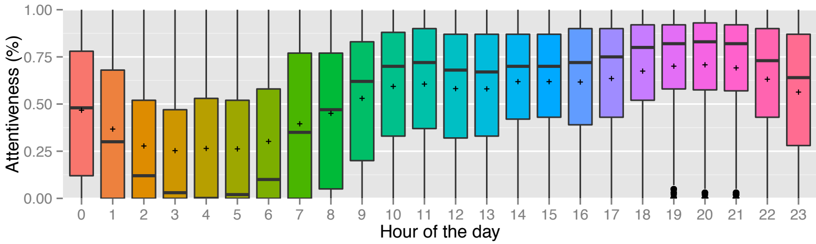


Figure 3.4: Average attentiveness by hour. The dash indicates the median, the cross the mean level of attentiveness. A value of .50 indicates that, on average, users would attend to messages within 2 minutes in 50% of the cases at a given hour.

3.2.4 Discussion

Based on the ground truth data we collected we trained a model that allowed us to predict people's attentiveness towards mobile messaging throughout the day. It showed that people are attentive to messages 12.1 hours of the day, while attentiveness is higher during week days than on weekends and people tend to be more attentive during evenings. Phases of inattentiveness seem to last for a relatively short amount of time since our participants returned to attentive states within 1-5 minutes in the majority (75% quantile) of cases. The model of attentiveness we constructed performed with around 80% accuracy significantly better than the 70.6% reported in previous work by Pielot *et al.* [215] due to a substantially larger data set (55,824 *vs.* 6,423 messages).

In 2006, a similar approach was taken by Avrahami and Hudson [11], who developed statistical models to predict users' responsiveness to incoming messages, which they used to compute the likelihood of receivers to respond to messages within a certain time period. They used a similar set of features, but the algorithm they presented was limited to Desktop usage, which did not take into account people's messaging behavior while being mobile and was, therefore, limited to only a subset of people's day. Other researchers attempted to determine the time interval between the arrival and acting on incoming messages/notifications with 6 minutes (average) for replying to SMS [19], 6 minutes (median) for attending to messages [215], and 30 seconds (median) until a notification is clicked (if it is clicked) [232]. With a median of two minutes delay until attending to messages, the work presented here is in line with these findings and stresses that people tend to attend to messages promptly. Most previous studies, however, only report measures of central tendency. The findings presented here, in contrast, give a more detailed estimation of people's attentiveness throughout the day. Hence, we provide insights into *when* people are more or less attentive and *how long* these phases of inattentiveness last. Our finding that people spend about 12 hours per day in highly attentive states further advances results reported by Dey *et al.* [65], who found in 2011 that phone users kept their within arm's reach for about 53% and in the same room for about 88% of the time. Keeping the phone close, therefore, also seems to lead to people attending new messages promptly for most parts of the day.

In 2013, Church and de Oliveira [47] reported that people exhibit high expectations towards the responsiveness of their conversation partners in mobile messaging. Strategies to deliver notifications in opportune moments [137, 141, 209, 216]—such as *bounded-deferral*, which may delay the delivery of messages—therefore, only work without violating expectations, if there is a sufficient number such moments. Our findings suggest that this is, indeed, the case: we found our participants to be attentive for large parts of their wake time and phases of estimated non-attentiveness during daytime typically lasted for only 1-5 minutes. In a majority of the cases, there will hence be a suitable amount of opportune moments sufficiently stacked together. Whether these moments are truly *opportune* or whether people simply give their phones priority over other activities, such as meetings or being out with friends, needs to be further investigated.

3.2.5 Study Conclusion

Our study on people's attentiveness provides quantitative evidence for people's tendency to exhibit high levels of attentiveness towards mobile messaging during large parts of the day: during 73.5% of their waking hours, people are highly attentive, while phases of inattentiveness tend to last only for a few minutes. Our findings, therefore, inform the design and development of such applications as any method for intelligent notification delivery can expect a generally high level of attentiveness from their users. The concept of *bounded deferral* would, therefore, work well in 75% of the cases with a bound of 5 minutes.

We do, however, need a better understanding of the underlying causes for such high levels of attentiveness, such as social pressure and positive reward loops. Therefore, we need ask questions, such as: how can we overcome such social mechanisms in order to create spaces where phone users can retreat to and feel free to think and reflect without external pressures? With being attentive to mobile messaging for 12.1 hours per day, people spend a great deal paying attention to their mobile devices. This made us wonder whether besides serving a need for communication the mobile phone may be picked up as a distraction. And if so, whether these moments, in which people turn to their phone to actively seek stimulation, could be predicted.

3.3 Detecting Stimulus-Seeking Moments

Despite the fact that the mobile phone is a major source of incoming distractions through alerts and notifications, people seem to also actively seek out stimulation in idle moments. Situations in which people are, for example, waiting (*e.g.*, for the bus, in a supermarket queue) and turn to their phone to kill time are unique in a sense that they actively seek out information to be consumed often without a specific goal in mind. Such behavior is especially common to battle temporary states of boredom. We were interested if such behavior might show certain patterns in phone usage characterized by, for example, frequent app switches or briefly checking social media sites. Hence, we set out to explore how such behavior could be detected automatically by the phone's sensors. Awareness of states, in which users actively seek out stimulation, could open up new ways of delivering and recommending content proactively.

In the following, we describe the design, implementation, and results of a two-week in-the-wild study, in which we collected over 40,000,000 phone usage logs and 4398 boredom self-reports of 54 mobile phone users. From this vast amount of data, we were able to train a machine-learning model to predict states of boredom with an accuracy of up to 82.9%

The research question we set out to answer was:

- **RQ2**: Does boredom measurably affect phone usage and which usage characteristics are most prevalent during such states?

3.3.1 User Study

To investigate whether states of boredom could be detected by analyzing mobile phone usage, we conducted an in-the-wild user study. A total of 54 participants installed an app called *Borapp* on their mobile phones, which we developed to collect data over the course of two weeks. Besides phone usage, the app recorded ground truth data about users' self-assessed states of boredom by using the *Experience Sampling Method* [50]. Our goal was to relate the phone's usage data to the boredom states reported by the users.

Method

Experience sampling entails prompting the user several times throughout the day to provide a subjective assessment of current activities or feelings. We used mobile phone notifications to trigger mini surveys throughout the day asking about the user's current state of boredom. On a 5-item Likert-type scale (0=*disagree*, 4=*agree*) participants answered the question: "To what extend do you agree to the following statement: *'Right now, I feel bored.'*?" (see Figure 3.7). Taking these self-assessments as ground truth we analyzed phone usage data recorded around the time of filling in these surveys. By extracting 35 features from phone sensor data we trained a machine-learning model which would allow us to predict bored *vs.* non-bored states.

Apparatus

To be able to prompt participants throughout the day we created *Borapp*, an app for Android phones with OS 4.0 or newer. The app consisted of a data collection service, a notification trigger service, and the main view. The data collector recorded various phone sensor data, the background service was responsible for scheduling and triggering the ESM surveys, and the main view allowed participants to track their study progress.

To reduce battery drain we divided the sensor data collected into two groups: 1) those that were permanently collected even when the screen was switched off, and 2) those that were only collected when the phone was in use, *i.e.*, the screen was on and unlocked. Table 3.5 shows the list of sensor data permanently collected, whereas sensors whose data was collected only when the phone was unlocked is displayed in Table 3.6.

To allow these sensors to collect the data listed, participants had to explicitly grant information access to the *Android Accessibility Services*, to notifications, and to location information. Due to privacy restrictions, these sensors are not exposed by the standard Android API. Through the accessibility service, foreground applications could be monitored, while notification access let us track which applications post notifications and when.

Sensor	Description
Battery Status	Battery level ranging from 0-100%
Notifications	Time and type (app) of notification
Screen Events	Screen turned on, off, and unlocked
Phone Events	Time of incoming and outgoing calls
Proximity	Screen covered or not
Ringer Mode	Silent, Vibration, Normal
SMS	Time of receiving, reading, and sending SMS

Figure 3.5: List of sensors permanently collected.

Sensor	Description
Airplane Mode	Whether phone in airplane mode
Ambient Noise	Noise in dB as sensed by the microphone
Audio Jack	Phone connected to headphones or speakers
Cell Tower	The cell tower the phone is connected to
Data Activity	Number of bytes up/downloaded
Foreground app	Package name of the app in foreground
Light	Ambient light level in SI lux units
Screen Orient	Portrait or Landscape mode
Wifi Infos	The WiFi network the phone is connected to

Figure 3.6: List of sensors only collected when the phone was unlocked.

Throughout the day a notification trigger service would fire notifications at semi-regular intervals. By clicking on these notifications users were taken to the mini-survey shown in Figure 3.7. Notifications were fired on average six times a day. Probes were only triggered during the day to avoid disturbance and more so when the phone was in use. Between probes answered we made sure at least 60 minutes had passed. Each survey response was recorded with the current timestamp in order to be later connected to the phone usage logs.

Borapp provided the main view for participants to be able to track their progress throughout the study. For successfully completing the study, the app needed to collect 14 days of data and 84 (14x6) questionnaires had to be filled in. Figure 3.8 depicts a screenshot of that view where the number of days and surveys completed were listed.

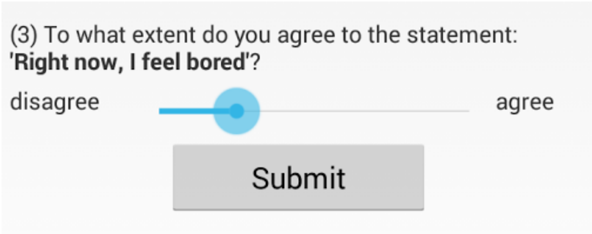


Figure 3.7: Screenshot of the ESM probe to assess bored states throughout the day.

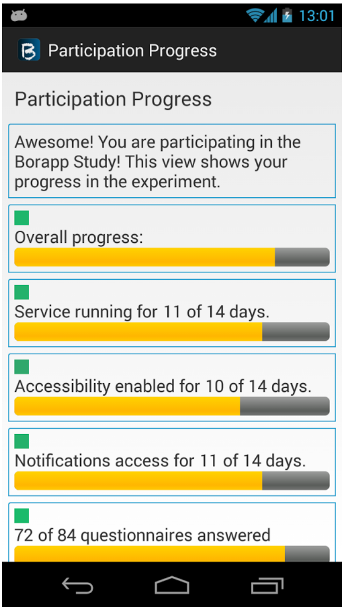


Figure 3.8: Screenshot of the progress view indicating participant's study completion status.

The app further provided detailed instructions about the study and device setup procedure. This included an explanation of the data collected, the purpose of the study, a consent form, and voluntary survey collecting demographic data.

Collected sensor data was written to a log file on the mobile phone's local storage. In regular intervals, the app checked whether the phone was connected to a Wifi network in order to upload the logs to our server and to avoid draining users' data plans.

Participants

After making the app available on *Google Play*, we distributed the link via a university mailing list as well as through a list containing volunteers from Spain who had indicated their general interest in research studies. The study was further disseminated through social network channels. We promoted the study by advertising a 20 EUR reward for each participant who completed the study, which entailed 14 days of data collection.

After about one month of running the study, we stopped collecting data and created a snapshot of the data of those participants who had completed the study until then. The raw dataset contained 43,342,860 data points from phone sensors and 4,826 survey responses from 61 unique mobile devices. In seven cases survey responses barely varied, which we interpreted as participants who had not taken the study seriously. After removing the data from these devices, we ended up with 54 remaining participants and 4,398 valid self-reports. Participants provided between 84 and 173 ($M = 110.3, SD = 25.8$) self-reports each. 39 participants reported their age in a range between 21 and 57 ($M = 31.0, SD = 7.9$) and to be female in 11 and male in 23 cases. The remaining 19 participants either chose the 'other' option or simply did not disclose their gender, since providing demographic information was optional. According to device locales (52% es_ES, 18% de_DE, 13% en_US) and timezones (79% UTC+1, 6% UTC+0 and 5% UTC+8), most participants were from Spain, Germany, and the United States.

Procedure

In June of 2014 we made *Borapp* publicly available for download on *Google Play*. The app was free to download and advertised through the email lists and social channels described. Since we were giving out money for successful study participation we limited the maximum number of sign-ups to a 100. To ensure that a participant was eligible for receiving the reward, the server conducted a participant count when the app was installed and first launched to check whether the new user was among the first 100 signing up for the study. The eligibility status was then directly communicated to the respective participant. However, people who were not eligible for the reward were informed by the app, but could nevertheless take part in the study.

Upon installation and eligibility check the app led users through the setup process. The first screen explained the purpose and procedure of the study, detailed the data collection process, provided contact information, and asked for explicit consent. The text explicitly stated what kind of personally identifiable information we would collect, namely device location. In the next step, participants were asked to grant access to the Android Accessibility Services as well as to notifications and provided detailed information on how to do that. The setup process was completed by an optional survey where participants could specify their age, gender, and leave their email address for reward collection. Once setup was complete the app started collecting sensor data and began triggering ESM probes via the notification scheduler.

Upon completion of the study participants eligible to collect their reward were sent a 20 EUR Amazon voucher. Successful completion required users to enable accessibility service and notification access, to keep the app running in the background for at least two weeks, and answer at least 84 ESM questionnaires.

Results

With 4398 filled in ESM questionnaires serving as ground truth and more than 40,000,000 phone usage logs, we conducted the data analysis in form of a machine-learning classification task. This approach enabled us to explore the relationship between different phone usage patterns and reported boredom and further allowed us to quantify the degree of which boredom can be inferred from mobile phone usage.

Feature Elicitation

We identified 35 features related to phone-usage patterns in the following seven categories:

1. **Context**: features relating to the phone's current context.
2. **Demographics**: explicitly collected features relating to the user.
3. **Time since last activity**: features relating to activities and their last occurrences.

4. **Usage intensity**: features based on direct phone usage.
5. **External triggers**: features related to phone events triggered from the outside.
6. **"Idling"** : features related to short, but frequent interactions with the phone hinting towards less goal-oriented activities.
7. **Usage type**: features related to direct user interactions with intent.

Detailed descriptions of each feature are listed in Table 3.1 and Table 3.2 under their respective categories. Features, such as time since last phone call, or ringer mode state, could be retrieved discretely at the time of the ground truth collection. Other features, such as battery drain or most used app, needed to be computed over a certain time window prior to the user submitting the subjective rating. To chose a time window we experimented with different window sizes of 1, 5, 10, 30, and 60 minutes in length. Here, we managed to achieve the best classification results with a 5-minute time window, *i.e.*, for all feature we took into account how the phone had been used five minutes before the respective ESM questionnaire was filled in.

For features related to the applications in the foreground or notifications triggered we used an app blacklist, that prevented, for example, system services from being taken into consideration (on some devices a notification event was triggered every time the keyboard was brought to the front, which we thereby effectively excluded from our analysis). Linear models are highly sensitive to outliers, which is why we checked numeric features for whether they required saturation, thereby removing outliers beyond a certain threshold. However, we ended up with the numeric features all having long-tail distributions, which means there were only positive outliers. Based on the skewness of each feature, we chose the appropriate percentile out of 90%, 95%, and 99%, and used that as an upper limit. A number of entries for features related to specific apps (*e.g.* last app in the foreground, most-used app, last notification) were sparse due to the fact that many of the recorded apps appeared only a few times or once in our overall dataset. That made it difficult to properly learn the meaning of sparse elements. Hence, we reduced the dimensionality of these features by assigning rarely used apps to the 'other' category. Again, this distribution was heavily skewed, which is why we kept the ten most frequent applications and mapped the rest into the 'other' category. For features describing application categories, we kept the three major categories (*i.e.*, communication, productivity, and society) which accounted for two-thirds of entries.

Context	
audio	Indicates whether the phone is connected to a headphone or a bluetooth speaker
charging	Whether the phone is connected to a charger or not
day_of_week	Day of the week (0-6)
hour_of_day	Hour of the day (0-23)
light	Light level in lux measured by the proximity sensor
proximity	Flag whether screen is covered or not
ringer_mode	Ringer mode (silent, vibrate, normal)
semantic_location	Home, work, other, or unknown
Demographics	
age	The participant's age in years
gender	The participant's gender
Last Communication Activity	
time_last_incoming_call	Time since last incoming phone call
time_last_notif	Time since last notification (excluding Borapp probe)
time_last_outgoing_call	Time since the user last made a phone call
time_last_SMS_read	Time since the last SMS was read
time_last_SMS_received	Time since the last SMS was received
time_last_SMS_sent	Time since the last SMS was sent

Table 3.1: List of elicited features related to context, demographics, and time since last communication activity.

Ground Truth

Predicting whether a person is in a bored or non-bored state is a binary classification. We collected the ground truth for these states through the ESM probes, in which participants were asked: “To what extend do you agree to the following statement: *’Right now, I feel bored.’*?” (see Figure 3.7). These self-assessments were answered on a 5-item Likert-type scale (0=*disagree*, 4=*agree*), which resulted in an average boredom rating of $M = 1.17$ and $Mdn = 1$. Hence, participants tended to generally disagree with the statement, as Figure 3.9 shows.

We classified participants as being bored whenever they tended to agree with the statement (scores 3 and 4), which we will refer to as *absolute* ground truth. However, we discovered that many participants had different anchor points: some rated themselves to be more bored on average than the rest. In previous work by Farmer and Sundberg [94] different predispositions of people towards boredom were found, which may explain our observation that some participants

Usage (related to usage intensity)	
battery_drain	Average battery drain in time window
battery_level	Battery change during the last session
bytes_received	Number of bytes received during time window
bytes_transmitted	Number of bytes transmitted during time window
time_in_comm_apps	Time spent in communication apps, categorized to none, micro session, and full session
Usage (related to whether it was triggered externally)	
num_notifs	Number of notifications received in time window
last_notif	Name of the app that created the last notification
last_notif_category	Category of the app that created the last notification
Usage (related to the user being idling)	
apps_per_min	Number of apps used in time-window divided by time the screen was on
num_apps	Number of apps launched in time window before probe
num_unlock	Number of phone unlocks in time window prior to probe
time_last_notif_access	Time since the user last opened the notification center
time_last_unlock	Time since the user last unlocked the phone
Usage (related to the type of usage)	
screen_orient_changes	Flag whether there have been screen orientation changes in the time window
app_category_in_focus	Category of the app in focus prior to the probe
app_in_focus	App that was in focus prior to the probe
comm_notifs_in_tw	received in the time window prior to the probe
most_used_app	Name of the app used most in the time window
most_used_app_category	Category of the app used most in the time window
prev_app_in_focus	App in focus prior to *app_in_focus*

Table 3.2: List of elicited features related to usage intensity, external triggers, idling and type.

had a different baseline level of boredom. Hence, we chose this *personalized* ground truth to reflect whether participants were feeling more bored than usual. Therefore, we transformed the absolute responses into personalized z-Scores, where 0 indicates that the participant considered herself not bored as a baseline boredom level during the study. Consequently, samples with values over $+.25$ were considered positive on this personalized scale, leaving us with a *normalized* ground truth. This distinction allowed us to derive how well we could predict whether people were more bored than usual (normalized) *vs.* how bored they tended to be in general (absolute).

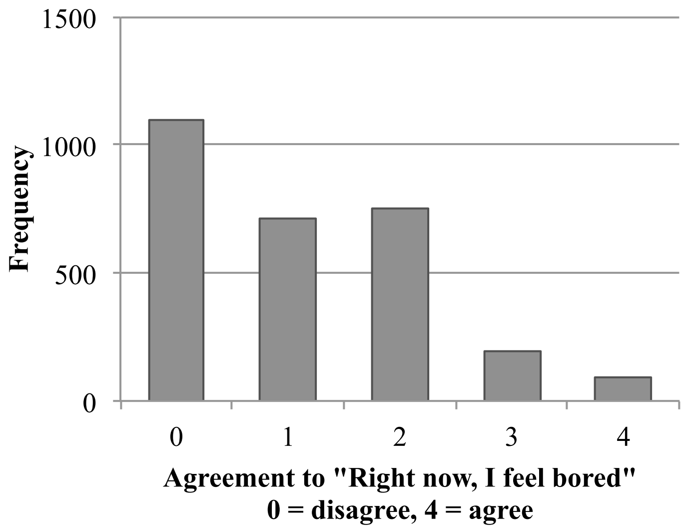


Figure 3.9: Histogram of ESM questionnaire responses.

Thus, we ended up with two datasets depending on whether we considered the ground truth to be *absolute* or *normalized*. Out of the total 4398 boredom assessments using the *normalized* ground truth produced 1518 (34.5%) instances classified as *bored*, and 2880 (65.5%) instances classified as *baseline*. This distribution is in line with boredom assessments reported by Goetz *et al.* [113], in which participants indicated to be bored in one-third of their responses. In contrast, the dataset produced by using the *absolute* ground truth with its 446 (10.1%) instances classified as *bored* was less balanced.

As mentioned earlier there seemed to be a general predisposition of people towards boredom, which affected their self-assessment scores. Bixler *et al.* [24] investigated the detection of affect and came to the conclusion that detection accuracy could be increased by taking users' psychological traits into account. Hence, we disseminated a post-hoc survey with a 28-item Boredom Proneness Scale (BPS) [94] to participants who had volunteered to leave their email address. Since this happened after the actual study had already been completed, we did not receive responses from all participants. Out of 54, 22 participants completed the survey. We added the resulting boredom *proneness* scores to participants' demographics and listed them as an additional feature to their self-reports. However, since our primary interest was in detecting moments, in which people used their phone to kill time, *i.e.*, their stimulation levels dropped, we chose the *normalized* ground truth for our primary data analysis without boredom proneness information. We were hoping this modeling choice would increase our model's applicability in a sense that it would not require obtaining boredom proneness scores before deployment and therefore was capable of detecting boredom deviations even for people who were not prone to be bored.

Boredom Classification

For classifying user states into bored *vs.* non-bored, we used three different classification methods and compared their respective performances:

1. A linear classifier: L2-regularized Logistic Regression (Logistic Regression (LR)) [125].
2. A non-linear classifier: support Vector Machines with Radial Basis Functions kernel (Support Vector Machines (SVM)) [260].
3. An Ensemble Learning technique: Random Forest (RF) [30].

For each classifier, we used the same model-building methodology: a nested cross-validation approach [43]. Thereby an inner loop performs a grid search over the space of model hyper-parameters and selects the best performing values. At the same time, an outer loop measures the performance of the model found in the inner loop. To prevent any positive bias when measuring the performance, the fold, which is being evaluated in each step of the outer loop, is not used for the training phase. In our analysis, we used 10-folds for the outer and 5-folds for the inner loop.

We achieved the best results with the Random Forest (RF) classifier. Similar to other ensemble learning methods, such as *Boosting* and *Bagging*, RFs use multiple weak-learners and aggregate their results, thereby optimizing the bias-variance trade-off. RF makes use of decision trees as base classifiers, which introduce randomization in several stages: first, for each tree of the forest different bootstrapped samples of the training data are used for its creation. Each node of that tree is then greedily split based on the best feature for only a random subset of all the variables. This process prevents a too strong influence of correlations between the different trees in the forest and helps to reduce the risk of over-fitting. The models created in this manner are inherently non-linear, tolerate outliers and implicitly support categorical variables.

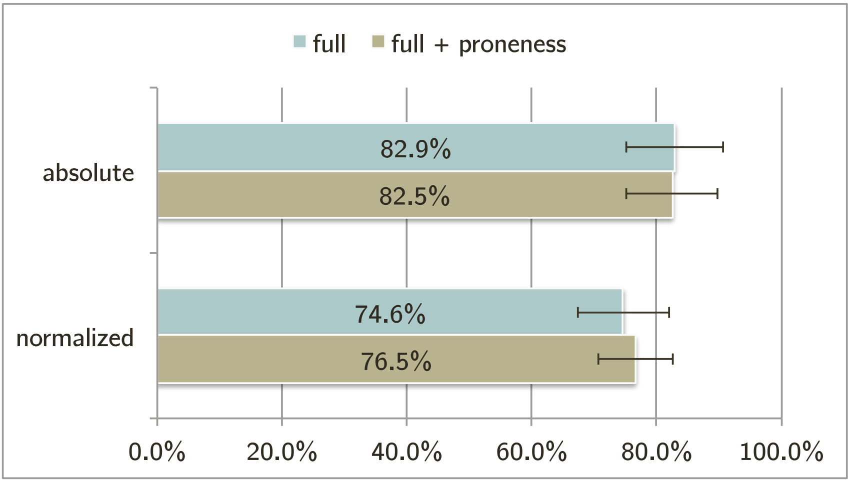


Figure 3.10: Area Under the ROC Curve (AUCROC) performance of the RF classifier for all four datasets.

With about two-thirds of participants' self-assessments being classified as not bored, we had to deal with an unbalanced dataset. The Area under a ROC curve (AUCROC) is an accuracy metric that can handle such asymmetrical data sets, whereby ROC stands for *Receiver Operating Characteristic*. Considering boredom proneness and the *absolute* as well as *normalized* ground truth we ended up with 4 datasets. Figure 3.10 shows the resulting classification performances of RF for all of these datasets in comparison. Using the *absolute* ground truth produced consistently better performance than using the *normalized* ground truth. Boredom proneness slightly reduced the variance.

Figure 3.11 depicts the precision-recall curve for our *primary data set*, *i.e.*, *normalized* ground truth without boredom proneness. The model shows a high flexibility for choosing different classification thresholds in order to trade precision for recall, which depends on the characteristics of the application setting. Scenarios, in which boredom detection is used to actively probe users, should prioritize precision to minimize the number of false positives, which may annoy users. Here, precision levels of 70.1% (for over 30% recall), or 62.4% (for 50% recall) in less restrictive scenarios can be reached.

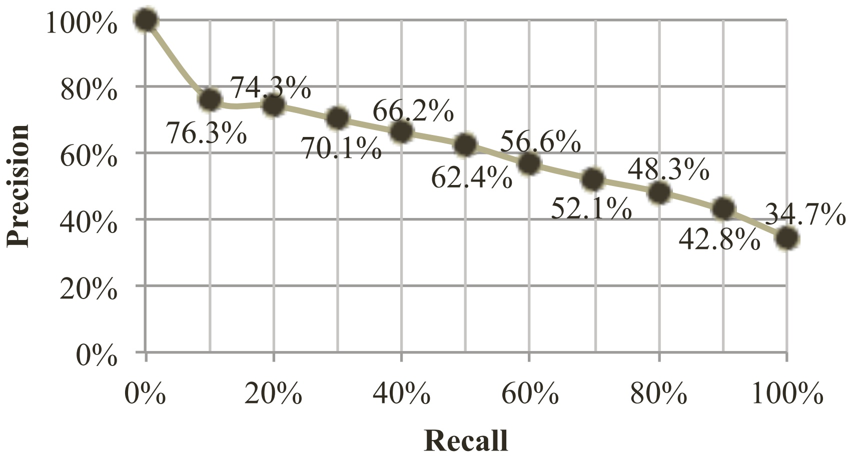


Figure 3.11: Precision-recall plot for the primary data set (*normalized* ground truth, no proneness).

Feature Analysis

Next, we set out to analyze which features were most meaningful to predict bored *vs.* non-bored states. Therefore, we used the RF to rank features according to their importance for the classification by using a measure called *Mean Impurity Decrease*. Since RFs consist of a number of decision trees, every tree node is a condition on a single feature, which effectively splits the dataset into two so that similar response values end up in the same set. *Impurity* is the measure based on which the (locally) optimal condition is chosen. When training a tree, we can, therefore, compute how much each feature decreases the *weighted impurity* in a tree. The *impurity decrease* from each feature can then be averaged across the forest and the features ranked according to this measure [30].

By applying this method, we computed the importance of each feature of our primary data set. The top 20 features and their importance measure are listed in Table 3.3 (Column: *Import*). By clustering these features into our groups depicted in Figure 3.1 and 3.2, we end up with the following most important feature categories:

- **Recency of communication activity**: last time a communication happened via phone or SMS and the last time a notification arrived (notifications were largely generated by applications from the communication category).

- **Intensity of recent usage**: *e.g.*, volume of internet traffic, number of phone unlocks, and level of interactions with applications in the last five minutes.
- **General usage intensity** reflected by, *e.g.*, battery drain, proximity sensor states (*i.e.*, phone screen covered or not), or time since last phone use.
- **Context / time of day** indicated by the hour of the day and light sensor data[9].
- **Demographics**: participants' age and gender.

Feature	*Import*	*Correlation*	*The more bored, the ..*
time_last_outgoing_call	0.0607	-0.143	less time passed
time_last_incoming_call	0.0580	0.088	more time passed
time_last_notif	0.0564	0.091	more time passed
time_last_SMS_received	0.0483	0.053	more time passed
time_last_SMS_sent	0.0405	-0.090	less time passed
time_last_SMS_read	0.0388	-0.013	less time passed
light	0.0537	-0.010	darker
hour_of_day	0.0411	0.038	later
proximity	0.0153	-0.186	less covered
gender (0=f, 1=m)	0.0128	0.099	more male (1)
age	0.0093	n.a.	+20s/40s, -30s
num_notifs	0.0123	0.061	more notifications
time_last_notif_cntr_acc	0.0486	-0.015	less time passed
time_last_unlock	0.0400	-0.007	less time passed
apps_per_min	0.0199	0.024	more apps per minute
num_apps	0.0124	0.049	more apps
bytes_received	0.0546	-0.012	less bytes received
bytes_transmitted	0.0500	0.039	more bytes sent
battery_level	0.0268	0.012	the higher
battery_drain	0.0249	-0.014	the lower

Table 3.3: Most important features in the primary data set sorted by their *Mean Impurity Decrease* score. The bigger (*i.e.* more blue) the value, the more predictive the feature is for boredom.

To analyze the relationship between phone usage patterns and boredom, we focused on the top 20 features. Therefore, we trained a Linear-Regression Model to yield the sign (positive or negative) for each feature, which is depicted in the

[9] The same physical sensor returns the ambient light levels and whether the phone screen is covered

Correlation column of Table 3.3. From these correlations we were able to draw the following conclusions:

- Communication: participants tended to be more bored the more time had passed since having received the last phone call, SMS, or notification, and the less time had passed since placing a phone call or sending an SMS.
- Notifications: the volume of notifications received in the last five minutes was likely to be higher when being bored.
- Fiddling with the phone: boredom was found to be correlated with more phone usage: the phone screen was less likely to be covered (which, for example, happens when the phone is stowed away), more apps were used, the last unlocking and checking for new notifications happened more recently, and the volume of data uploaded was higher when participants were bored.
- Data download and battery drain were lower when people were bored.
- Gender: men tended to be more bored than women.
- Age: for participants in their 20s and 40s boredom was higher than for participants in their 30s.
- Time of day: boredom was more likely the later it was in the day and the darker the ambient lighting conditions.

Looking at how app usage correlated with boredom we found that applications, such as Instagram, email, settings, the built-in browser, and apps in the 'other' category were correlated most strongly with being bored. In contrast, communication apps, such as Facebook, SMS, and Google Chrome, correlated with non-bored states.

3.3.2 Discussion

Our goal was to investigate different states of user attention and whether we could identify situations, in which users explicitly turned to their mobile phone to kill time and actively seek stimulation. Our field study described provides empirical evidence that 1) there are measurable phone usage patterns that allow us to infer boredom with acceptable accuracy and 2) states of boredom are especially related to communication activity, usage intensity, hour of the day, and demographics

Inferring Boredom

We trained a machine-learning model based on phone usage data from 54 participants that allowed us to distinguish bored *vs.* non-bored user states. Thereby, we investigated the impact of additional factors, namely whether we based our analysis on a *normalized* or *absolute* ground truth computed from participants' self-assessment scores. Our primary dataset produced a 74.6% AUCROC performance accuracy with *normalized* ground truth and without boredom proneness features. Using the *absolute* ground truth and the boredom proneness data our model even reached 82.9% accuracy with the *absolute* ground truth data performing consistently higher. This can be explained by the notion that higher boredom levels lead to a higher agreement with the provided statement of feeling bored (scores of 3 and 4). In contrast, *normalized* boredom lowered the model's detection performance but allowed detection when people deviated from their baseline state. To infer *normalized* boredom, user-dependent modeling was necessary since people had different anchor points regarding their perceived boredom levels.

Adding boredom proneness had a minor effect on the prediction results, namely that it decreased variance, *i.e.*, it helped to make the model more stable. We found the effect to be not as pronounced as in related work by Biller and D'Mellos [24], which could be caused by our relatively small sample size of 22 (40.7% of our participants). Matic *et al.* [185], however, recently presented a model for estimating boredom proneness based on phone usage, which could render a dedicated questionnaire obsolete and be a means to reasonably include boredom proneness data in prediction models.

The caveat of such in-the-wild studies—like the one we conducted here—is the lack of control over the participant's environment. While striving for ecological validity, possibly confounding factors are often overlooked or can simply not be identified. In our case, participants filled in the ESM questionnaires throughout the day, but in convenient moments. While we asked them to deliver 84 responses in total, participants were free to dismiss ESM probes of their own choosing. How they chose when to fill in a questionnaire might have introduced a bias we cannot detect for certain: probes might have been rejected whenever users were deeply focused on other tasks, for example. Especially using the *absolute* ground truth data may have amplified the effect of moments, in which the user was bored.

Our model's accuracy, however, proved significant. Related work focused on detecting happiness [26] and stress levels [25] throughout the day based on phone usage data. To achieve reasonable results they needed to include additional information about the user's personality traits and external sources, such as weather, which was not needed for our model.

Indicators of Boredom

With the hypothesis that boredom is reflected in the way people use their mobile phones, we looked into specific usage patterns. Therefore, we analyzed the top 20 indicative features, which were mostly related to five aspects, namely: recency of communication activity, recent and general usage intensity, usage context, and basic demographics. Essentially, the more time had passed since a last incoming communication and the more recently the user had reached out through communication channels, such as placed a phone call or sent a message, the more evidence accrued towards being bored. Hence, being reached out to seemed to mitigate, while conspicuously reaching out to people might have been a sign of boredom. We were also able to relate boredom to common "fiddling with the phone" activities in the form of phone usage intensity, which confirms previous work [31, 206] stating that people turn to their phones to kill time. With our work, we provide empirical evidence to solidify these claims and advance their findings showing that increased phone usage contributes to boredom detection.

We were further able to show that daytime had an effect on phases of boredom. Work by Mark *et al.* [182] showed that boredom levels vary throughout the day and that boredom is generally lower during late working hours. Based on features, such as time of day and lighting conditions, we were able to corroborate this finding and show that states of boredom tend to become more frequent as the day progresses. Also, demographics including age and gender seem to play a role. We found boredom to be higher in men participating in our study. Age groups in their 20s and 40s seem to show more signs of boredom than those in their 30s. Previous work has looked at the effect of age [270] and gender [16] as predictors for boredom in leisure time, which our findings are in line with.

The analysis described in our study is, however, limited to direct correlations of features and boredom. There may not be any direct correlation between some features and the respective levels of boredom, but features may rather become indicative when combined with others. Further, correlation does not imply causation: due to our observational approach, causal interpretations, such as incoming messages relieving us from boredom, are of rather speculative nature. Phone usage is not the sole predictor of boredom states, however. Contextual factors and demographics generally play into the mix.

3.3.3 Study Conclusion

We developed an approach to detect boredom based on how people use their phone throughout the day. Therefore, we conducted an in-the-wild study with 54 participants using the *Experience Sampling Method* to collect ground truth on boredom states and explore related phone usage context. The most indicative features for predicting states of boredom were recency of communication, usage intensity, time of day, and demographic information. By using machine-learning we created a model that was able to predict moments, in which users where bored *vs.* non-bored with an accuracy between 74.6% and 82.9%.

Such moments could be used for benign interventions, such as sending proactive recommendations. Boredom as a trigger could inform the design of mobile recommender systems with a better understanding of when and how to engage users:

- Providing content suggestions in moments of boredom: *e.g.*, articles to read or videos to watch.
- Making boredom-curing activity suggestions, such as contacting friends.
- Instead of suggesting *killing time* activities users could also be reminded to (re-)engage with planned activities: *e.g.*, clear the backlog of a todo or read-later list.
- Helping people make use of boredom beyond online recommendations by fostering introspection, reflection, and creativity. A possible suggestion, for example, could be to leave the phone alone for a while or even automatically switch it off.
- Preventing interruptions by, for example, blocking incoming communication requests in situations in which people are not bored, but focused.

Future investigations could focus on mapping out the numerous possibilities that boredom or engagement detection can provide. Technology could help make the best of these moments or attempt to intervene. In fact, we will come back to this idea in chapter 6, where we will discuss delivering content in opportune moments.

3.4 Chapter Summary

In this Chapter, we investigated people's attention in a technological context. Attention is a crucial factor for effective information intake and digestion. Therefore, we focused on how people attend to their mobile phones throughout the day and whether we can detect usage patterns that reveal underlying cognitive states, such as focus or boredom.

In a first step, we analyzed people's attentiveness towards mobile messaging as communication is one of the most prevalent mobile phone activities. Results show that people are highly attentive to messages for more than 12 hours of the day, which is most of the waking time. We further found that attentiveness is higher during the week than on weekends, as well as during the evening hours. Even when people are being found to be inattentive, they tend to return to an attentive state within 1-5 minutes in the majority (75% quantile) of cases. Our data shows that intelligent notification delivery, such as *bounded deferral*, is thus applicable. On the one hand, holding back messages in inattentive states prevents users from being interrupted from a current task at hand. And on the other hand, there seem to be sufficient opportune moments, in which users bring their attention back to the phone so that delayed delivery will not negatively impact the handling of urgent messages. We base this insight on a classification model we developed, which performed with almost 80% accuracy to reliably predict users' attentiveness from phone usage patterns, which confirms **RQ1**.

With incoming messages and notifications being external stimuli that try to grab the user's attention we were further interested in situations, in which people pick up their phones to actively seek stimulation. Such moments stem from internal triggers and may be characterized by a general susceptibility and desire for targeted distractions. We found these situations being generally associated with boredom. Hence, we investigated whether these specific states could be detected based on phone usage patterns. We combined the logging of phone usage data of 54 participants over the course of two weeks with the collection of self-reports on momentary boredom levels. By training a Random Forest classifier with data on 35 features related to phone usage we were able to predict bored *vs.* non-bored user states with up to 82.9% accuracy, which allowed us to endorse **RQ2**.

An analysis of the top 20 predictive features yielded that most phone usage patterns were mostly related to recency of communication activities, phone usage intensity, time of day, and demographic information. Hence, we found usage characteristics that could be linked to users being in a bored or stimulus-seeking state. Table 3.3 summarizes the most indicative 20 features and therefore answers the second part of **RQ2**.

There is a certain temptation to constantly refer to our mobile devices that connect us with the pulse of available information. Whether it is a good idea to be constantly connected, is the basis of broad discussions. Before electronic devices became ubiquitous it was possible for us to disconnect from online information spaces, but now we find ourselves being inevitably alerted by our devices. In a previous study [232] we investigated the nature of notifications on mobile phones and analyzed which types were highly valued or easily dismissed. We found notifications related to messaging, people, and upcoming events to be most appreciated. The resulting design guidelines advise developers on how to balance the informing and disruptive nature of notifications. But even when mobile devices remain silent, our studies show that people are drawn to check various information sources in all kinds of situations. Whether this is considered a bliss for productivity or the beginning of an unruly attention fragmentation makes for a controversial debate. Our studies have shown that attention is literally not scarce and people frequently check and re-check the information delivered by their devices. Whether we can make positive use of this behavior and nudge users towards more productive or mindful activities will be an interesting challenge for future applications. "The cure to boredom is curiosity" –a quote by the American writer Dorothy Parker- points toward an important function of boredom: boredom is an innate indicator for a state, which is not sufficiently satisfying and therefore nudges us to leap to action. Hence, boredom can urge us to initiate creative processes and self-reflection [263].

In our attention economy, however, there is a commercial value in knowing when people are receptive to stimulation. Pushing advertisements in situations where users are willing to absorb their content, creates interesting business cases. Ad networks could charge advertisers according to the recipients' attention levels. Users, on the other hand, might be more willing to skim through ads when downtime occurs and might welcome the idea of being left alone when they are focused on other things.

Engagement studies, such as our boredom study, show how attentional states are linked to people's current activities. Insights into when a person is bored or focused can provide us with a better understanding of when people are more productive and when downtime occurs. This can also be used to inform the design of tools and interfaces to promote a better experience when mobile, but also in the workplace. During highly focused states, for example, devices in the user's environment could be advised to prevent interruptions, unless they are of high priority or contribute to the current strand of work. An uninterrupted focus is crucial for learning tasks as absent-mindedness is fatal for successfully committing information to long-term memory. Proactive and situated recommendations could advise users to engage with information optimizing for receptiveness.

Such systems could push the user to work on yet another todo-list item even during idle times. They could also remind users to use these minutes to call up an old friend. Or maybe, the device could recommend being switched off entirely. Sometimes being with ourselves and with our innate thoughts might create the kind of calm we need to escape our busy daily routines and re-charge our mental capacities.

Chapter 4

Content-Awareness

In Chapter 3 we investigated to what extent users pay attention to their phones and analyzed the context in which they seek stimulation through their devices. In this Chapter we focus on the specific content people currently engage with and how awareness of such content can be used to support information intake.

While people often seek out information on their mobile phones in parallel to other activities, such behavior has become especially prevalent when watching TV. Most of the so-called second-screen apps provide additional information and services for a specific TV program and app content is mostly manually curated by the program or app publishers. On the one hand manual content curation is costly for publishers, and on the other hand dedicated apps need to be installed before usage and leave the user with a fragmented app landscape. In this Chapter we assess the feasibility of creating a single application that detects the current content being watched and delivers automatically generated relevant information to improve the TV experience in real-time. Thus, we eliminate the need for manual content curation and users are no longer required to install a dedicated app for each program or show. Therefore, we developed an extended entity linking algorithm to extract important keywords from subtitle streams, which we link to highly relevant Wikipedia articles and Google search results. These automatically generated contents are then delivered through our second-screen app called *TVInsight* (Fig. 4.1), which displays this additional information in a context-sensitive way, namely in line with the program and program position that is currently being watched. In a user study, we evaluated our system with regard to its effects on the user's TV experience, comprehension of the program's

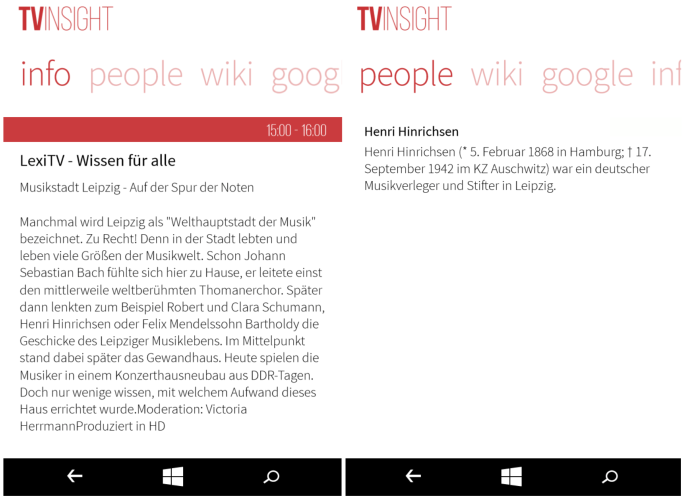


Figure 4.1: *TVInsight* is a second-screen app that shows additional program information by triggering web searches and Wikipedia look-ups in real-time and in line with the current program being watched. Left: general program information. Right: Wikipedia content according to people appearance.

content, and level of distraction as compared to conventional look-ups performed on a second device.

In this chapter we address the following research question:

- **RQ3**: How can awareness of the content which the user is currently exposed to be used to augment the user experience?

This chapter is based on the following publication:

- J. Knittel and T. Dingler. 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, pages 93–103, New York, NY, USA, 2016. ACM

4.1 Related Work

The work described in this chapter is generally rooted in the field of context-aware computing, and more specifically related to research regarding second-screen apps and entity linking.

Second-screen Apps

Some first guidelines for second-screen apps were formulated by Robertson *et al.* [229] who introduced a system connecting a Personal Digital Assistant (PDA) with a television device, where the TV was used more as an additional screen real estate. Utilizing the hardware of each connected device creates synergy effects that extend the experience of each device by itself. Cruickshank *et al.* [51] also looked at using a PDA as a second-screen device for browsing the TV program and changing channels.

In recent years a multitude of second-screen apps have commercially entered the market. They provide additional content, invite users to participate in real-time surveys, or comprise social features, such as social check-ins or sharing content. The market is fragmented with numerous shows and channels offering their proprietary second-screen app, mostly tailored to run on mobile devices. Geerts *et al.* [108] found that a general second-screen app is preferable with regard to every show providing its innate application while barriers to find additional program information should be kept at a minimum. Second-screen content should not solely mirror the program's content, but provide additional information. Further, the program progress should be taken into account.

By splitting functionality across devices the issue of diverting attention became more prevalent. Fleury *et al.* [101] found that users generally appreciate second-screen apps to draw their attention to additional content, but this should be done in unobtrusive ways. Van Cauwenberge *et al.* [257] looked into media multitasking and how using a search engine to answer questions while watching a documentary affects comprehension. They found that the increased cognitive load by using the search engine as a secondary task caused participants to not being able to comprehensively recite the facts of the documentary and performed worse on comprehension tests. A study by Google reported that 22% of second-screen usage was complementary to the current program [115].

According to a survey conducted by Nandakumar *et al.* [195] about 27% of TV show-related searches concern the characters and their relations, closely followed by searches about the plot with 23%. With a companion second screen prototype they showed synchronized and context-sensitive information mainly about the characters and found that this enhanced comprehension when used by first-time viewers of a late-season episode.

Content Linking

To reduce distractions, attempts have been made to automatically detect the TV program currently running in the background. Chuang *et al.* [46] developed a smartphone app using audio fingerprinting to recognize the current show in order to provide additional content by analyzing the video- and audio stream. While most additional content is manually curated, especially for canned user studies, Castillo *et al.* [41] presented an approach for automatically analyzing subtitles to find relevant news articles for a news program as well as music titles through mining song lyrics. Their algorithm was used in Yahoo!'s service *IntoNow* which was discontinued in 2014. Additional content can further be created by linking existing and relevant content together. Allan [7] formulated the need for automated procedures to link paragraphs to relevant documents in an automated way. Redondo *et al.* [226] performed named entity recognition on subtitles for news broadcasts and used structured data from DBpedia to generate a comprehensive set of relevant context items.

A comprehensive overview of common entity linking features was given by Shen *et al.* [244] who came to the conclusion that most algorithms were highly domain specific. Mihalecea *et al.* [53] presented a linking model integrating *Wikipedia* articles. To increase recall and precision values, other works have included PageRank values [178], the number of incoming links, or a combination of different features and classifiers [96]. Odijk *et al.* [203] made specific use of subtitles for linking Wikipedia articles by using a context graph based on the dynamic assembly of anchor phrases. However, the analysis of their approach was based on a curated set of well-defined topics but lacked an evaluation 'in the wild'.

Subtitles are traditionally used to make content accessible to people with hearing impairments. They have increasingly been subject to research focusing on word frequency analysis [147, 154, 199] showing that subtitles approximate spoken language. Hayati *et al.* [126] showed that displaying subtitles in foreign language movies improves auditory comprehension when shown in native as well as in foreign languages. Similarly, Mitterer *et al.* [189] found that subtitles in a foreign language support language perception and strengthen vocabulary. Kovacs *et* *al.* [162] developed a video player for foreign language learners which enhanced the traditional subtitles display by allowing users to jump back and forth through the video by clicking on the corresponding subtitle. Brasel *et al.* [29] investigated displaying advertisements in users' native language with subtitles and their effects on eye movements and recall showing that participants were able to recall more of the advertisement brands shown. However, they also showed a negative effect for recalling visual program elements.

Concluding, people seem to use a great number of apps and browser-based strategies to look up information based on the TV program they currently watch. But at this point, there is not a comprehensive solution proactively providing relevant content across TV programs. Manual curation would be unfeasible, which is why we focus on automated content generation techniques. By harvesting the availability of subtitles we present an approach in this Chapter for an entity linking algorithm for automatically generating additional program content. The solution we propose is independent of any particular channel or show. It prevents the user from having to download a variety of apps, but at the same time offers highly topic-relevant content. By proactively retrieving content, users have the information at their fingertips as they launch the app rather than having to actively look for content. This reduces potential diversion from the actual TV program. Limited distractions should benefit the TV experience—*e.g.*, through immersion—and improve content comprehension. Inspired by previous work in language learning we explore learning scenarios as application cases for context-sensitive second-screen apps.

4.2 Context-Aware Information Delivery

The ubiquity of mobile devices increasingly affects our TV experience. 90% of smartphone owners use their mobile device while watching TV, about 50% of this group browses the web while 30% use it for looking up additional information about the show, topic, and people involved [200]. People often use a second device, *i.e.*, phone or tablet, for looking up keywords online and gather additional information with relevance to the current TV program. In 2012 the PEW Research Center assessed the parallel usage of second-screen devices while watching TV and the relevance of that activity to the current program [245]: from almost 2000 participants, 22% used their cell phone to check facts mentioned in the program, 35% visited a related website, 20% looked up relevant social comments, and 19% posted their own comments about the program they were watching. However, the process of looking up relevant information can be quite slow, cumbersome and is error-prone; coming up with appropriate search keywords, getting the spelling right, and individually evaluating each search result can be time-consuming and requires mental effort, which diverts attention away from the current show.

The majority of commercial second-screen apps focus on one particular show or program. These apps are often hosted by the TV station itself or by an exclusive partner, so access to content is restricted to the particular app. Users who follow more than one show end up with a fragmented collection of apps on their devices. For content providers, on the other hand, curation can be expensive while content can be extracted in an automated way by intelligently linking existing resources together. Such an approach is cheaper and further allows live-shows to be augmented with additional information.

An increasing number of TV stations broadcast subtitles alongside their program. These subtitles bear great potential for content extraction algorithms since they contain crucial information about topics and people involved. Such contents can be used to augment the TV experience independent of the channel users are watching. Additional information can be purely informative, but can also be of educational use, especially while consuming foreign language shows, as they are often watched with the goal of language acquisition. Castillo *et al.* [41], for example, proposed a system that automatically analyzed subtitles to find relevant news articles for a given news program. Most systems and studies, however, are based on a very topic-specific domain, therefore limiting their general applicability in a diverse TV program environment.

In the following we report on an online survey we conducted to gather user requirements, the development of the application, and a final comprehensive user study assessing the effects on users' TV experience, comprehension of the program's content, and potential distractions compared to conventional look-ups performed on a second device.

4.2.1 Survey on Device Usage during TV Shows

To elicit features for a second-screen app we conducted an online survey in February of 2015 focusing on how people go about web searches while watching TV. We set out to analyze what type of information are looked up, on which devices and the timing of these searches. We were further interested in how people took advantage of subtitles, how frequently they used them, why and for which type of shows.

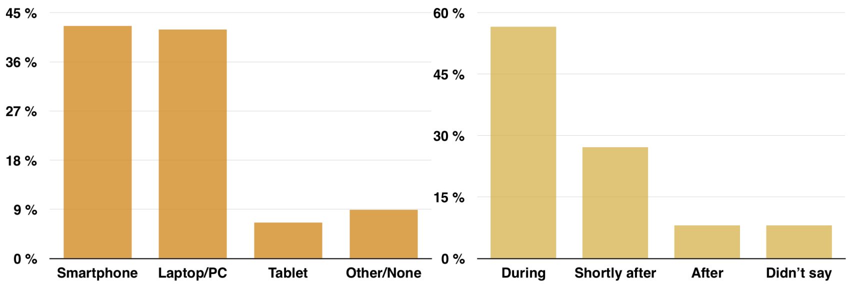


Figure 4.2: Types of devices used for web searches and the timing of the searches relative to the show watched.

We, therefore, set up an online survey using LimeSurvey[10] running on one of our university servers. After a brief explanation of the purpose of the survey and granting consent for the data collection, participants were redirected to three blocks of questions focusing on demographics, TV information search, and usage of subtitles. The survey was announced through university mailing lists and social networks. Over the course of two weeks, we recorded responses from 136 participants (60 females) with a mean age of 27 ($SD = 8.4$). 54% indicated to be students while 41% were working professionals. 40% stated they watched TV on daily basis, 88% at least once per week. Filling in the survey took less than 5 minutes.

Results

The data collected provides insights into how people search the web while watching TV as well as in what way they make use of subtitles.

Web Search

Of our survey participants, 19.9% stated they had looked up something online related to a TV show in the last 24 hours, 58% within the last week, and 80% within the last month. As to what they were searching for, 48.5% named people that were mentioned in the program, 28.7% searched for a terminology or had a specific question, and 10.3% looked up the current show's topic. Figure 4.2 depicts the device types used during triggering searches and at which point in

[10] http://limesurvey.org/

time: most people used their Smartphone (42.6%) or Laptop/PC (41.9%) to look up information, mostly while watching the show (56.6%). Figure 4.3 shows the program types in which participants usually use web searches according to what search category (terminology/question, people, topic) they look for. Participants indicated to mostly search the web during TV shows (33.5%), movies (19.1%) and documentaries (11.8%). Especially during shows and movies, people seem to look up people, during documentaries, there is a tendency to look up terminologies. As to where people end up when looking up such information, 61% indicated Wikipedia as number one while 10.3% named movie portals, such as *IMDb*[11], and only about 3% end up on the direct website of the TV channel or program. Most survey participants agreed that searches would not take too much time and usually satisfied their information need.

Subtitle Usage

Direct usage of subtitles was rather uncommon among participants. In Germany, where most survey participants resided, foreign media content is usually dubbed, which might explain this practice. More than 40% indicated to have never or more than a year ago made use of subtitles. 33% stated they had activated subtitles within the last month, only 8.8% within the last 24 hours. Subtitles seemed to be mostly used in movies (45.4%) and TV shows (38%), sometimes during documentaries (9.3%). In 84% of the usage cases, the program watched contained foreign language content, in 40% they were explicitly used for language learning purposes, the same amount accounted for a better understanding of acoustics.

Discussion and Feature Elicitation

Many consumers use secondary devices while watching TV for looking up additional program information mostly concerning people and the corresponding topic. Smartphones and Laptops seem to be the preferred devices, which confirms previous findings. The majority of web searches leads to Wikipedia where additional information about people and topics is to be found. It makes sense then to populate a second screen app with contents from or direct links to Wikipedia. Participants noted that they often did not launch web searches due to not having correctly heard a name or not knowing the wording of a topic. Thus, a second screen app should be aware of the program context including content specifics so that additional information can be provided as it becomes relevant to the current program position. Not surprisingly, subtitles seem to be mainly used to

[11] http://www.imdb.com/

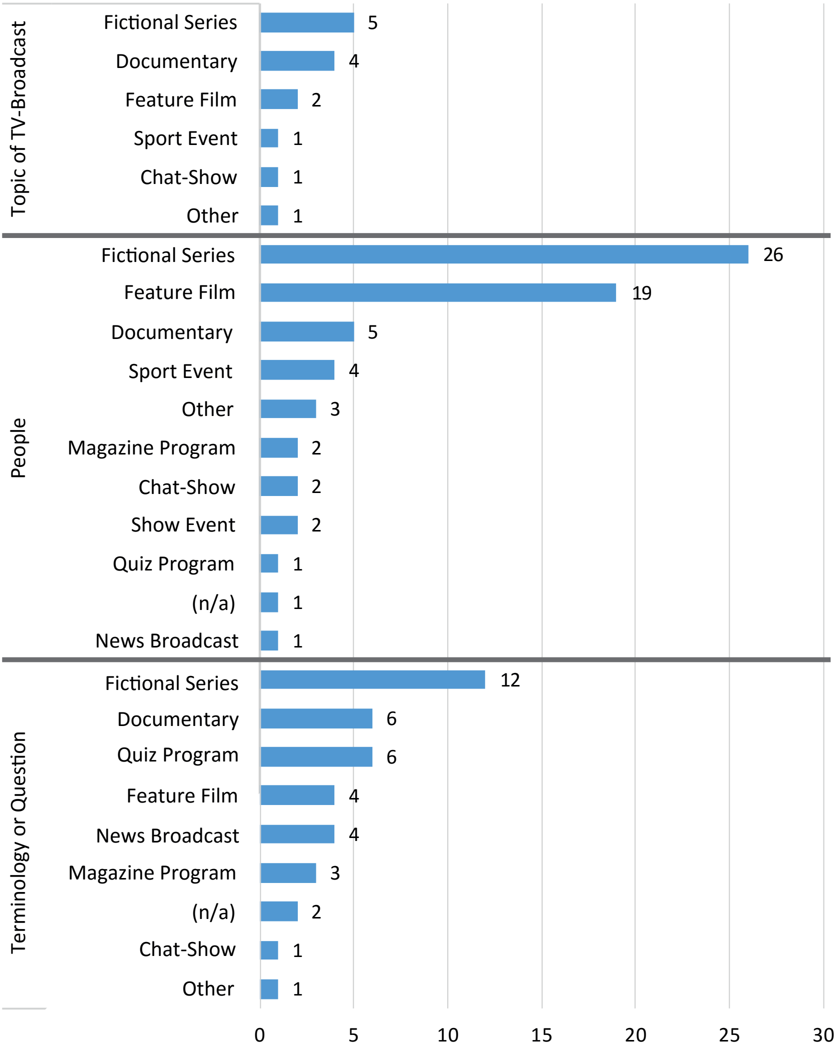


Figure 4.3: Types of shows according to search categories.

support comprehension of foreign language content. As related research states, subtitles often distract from the visuals. They do, however, provide important content keywords, which can be used as a starting point for web searches. A second-screen app can take advantage of the information contained in subtitles to automatically collect and provide additional information in real-time.

4.2.2 System Overview

Based on the insights from the online survey, we designed and implemented the second-screen app *TVInsight*. It provides additional information to the current TV program independent from the type of program or channel. Therefore, it needs to be aware of the current program context, *i.e.*, its content, the position, trigger automated web searches, and compile the resulting information into a comprehensive and easily accessible second screen experience.

Content Generation

An increasing amount of TV programs are broadcast with subtitle streams to make content accessible for people with hearing impairments. Our goal was to develop an entity linking algorithm specifically designed for real-time analysis of subtitles that works across a wide range of different TV-programs and genres. There are publicly available subtitle databases, such as *opensubtitles.org*[12], but they usually focus on movies and TV series and miss out on popular TV-program formats, such as live shows, chat shows, news broadcast, or documentaries. Hence, we developed a software which directly receives data streams from a TV tuner to decode and store subtitles broadcast via Teletext. Our system taps into cable TV streams using a Digital Video Broadcasting - Cable (DVB-C) extension card. For each transponder, the respective MPEG-2 transport stream multiplexes various data, video, and audio streams. We decode the Teletext streams and filter out packets containing subtitles. We further receive the Electronic Program Guide (EPG), which our software uses to associate subtitles with the corresponding TV program and store additional meta data.

Our software is written in C# and runs on a server which currently receives 30 TV channels. An admin interface (see Fig.4.4) shows current TV programs and most recently extracted subtitles. Here we can also tune to different frequencies or scan the transponder to list available channels.

[12] `https://www.opensubtitles.org`

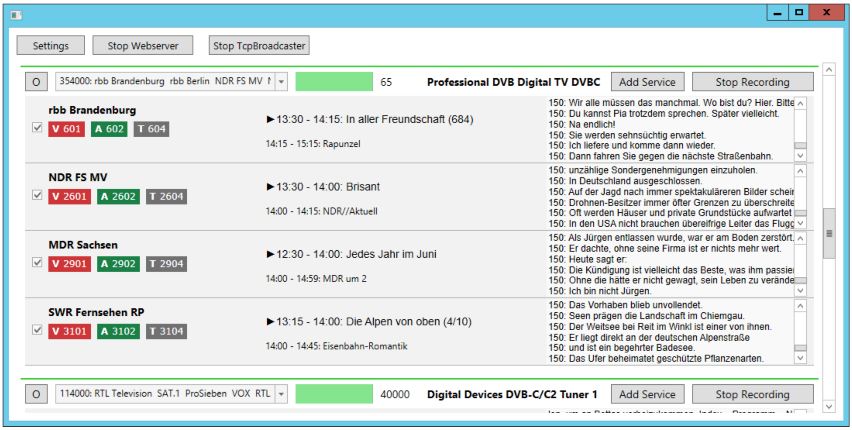


Figure 4.4: Admin view of the server software showing current TV programs and most recently extracted subtitles.

We intended to train our algorithm on a broad spectrum of TV programs. Hence, we let our server collect data over the course of four months, which resulted in 45.000 hours of programs with subtitles from 30 channels and a corpus of 136 million words in total. A word frequency analysis yielded about one million unique words (no stemming applied) and more than 200.000 words that appeared more than ten times. We focused on German channels, but similar corpora can be created using channels in other languages.

Entity Linking

Entity Linking describes the process of retrieving and linking phrases to respective counterparts in knowledge databases. We created an algorithm that continuously analyzes incoming subtitles in real-time and finds corresponding content on Wikipedia. To do this, we downloaded the publicly available german Wikipedia content[13] (~14GB), from which we extracted the links between articles, cleared out redundancies, extracted titles, definitions, images, and marked people entities. We wrote the algorithm from scratch in C#. Our approach is made up of three consecutive steps:

[13] http://dumps.wikimedia.org/dewiki/latest

1. Extracting Candidates

Similar to previous approaches like that of Odijk *et al.* [203], we used the collection of anchor texts (the displayed text linked to a specific page) to extract potential candidates. This step produced quite a number of candidates since commonly used words like 'this' and 'here' are often used as anchor texts within Wikipedia. On the other hand, anchor texts provide an extensive and high-quality list of synonyms and variant spellings of the same topic. We used stop words to immediately reduce the number of very unspecific candidates.

2. Selecting Target Page

For each candidate, there may be multiple possible destination pages. The page to which the corresponding anchor text was most often linked to within Wikipedia, was picked as the first-page candidate. In most cases, this was already the best target page. Garcia *et al.* [96] achieved a combined accuracy of 75% on several ambiguous data sets this way. To improve this, we took into account contextual information: to determine the second page candidate we intersected the context sets, comprised of the words of the first paragraph of the article on the one side, and the most recent words in the subtitle stream (sliding window approach) on the other side (ignoring stop words). The page whose intersection contained the least common words was considered the second candidate (if applicable). We used the word frequency of the generated subtitle or Wikipedia corpus, whichever was higher, to determine the least common word. In case there was a second-page candidate, a neural network decided which one was more relevant. To train the network, we manually annotated a set of about 2000 items which were randomly picked instances from our recorded set. Thus, we first extracted the candidates from our vast subtitle database, selected those with two-page candidates, filtered out very common anchor texts like '*this*', and manually annotated a subset of randomly picked candidates.

3. Determining Candidates To Be Linked

For each candidate and its chosen target page, we used a second neural network to decide whether an annotation between candidate and target site should take place. As stated before, the number of candidates extracted in the first step was rather large, thus, we had to narrow it down to items that were considered particularly relevant to the respective scene. Again, to train the network we randomly picked about 3500 candidates from our subtitle collection, performed the disambiguation and manually marked the items that should be linked.

Features

For the neural network classifiers, we elicited the following set of features based on previous work by Odijk *et al.* [203] and Fernandez *et al.* [96]. We further added some features to make the algorithm better suited for the analysis of subtitles.

- Word frequencies of anchor text and context intersection based on our subtitle corpus and Wikipedia.
- Probability that anchor text is linked to this page.
- Probability that anchor text is an anchor within Wikipedia.
- Indegree: how often the page is being linked within Wikipedia.
- Outdegree: number of anchor links on the page.
- Is the TV program mentioned on the page?
- Is the anchor text equal to the page name?
- Does the first paragraph contain the anchor text?
- Similarity metrics between anchor text and page title as well as context words and page title.

Performance

Our approach produced an in-memory directed word graph to efficiently extract the candidates and retrieve target page candidates. The run time of the classifier was negligible. Our implementation needed approximately 25GB of memory, but processed subtitles extremely fast since the time complexity lied in approximately $\mathscr{O}(1)$ per word. We optimized for performance rather than memory usage since available memory was less of an issue running the system in a cloud environment.

Results

To the best of our knowledge, our approach is the first one based on a comprehensive dataset taken from subtitles of a great variety of TV program genres. In contrast to previous work, we made heavy use of domain knowledge about subtitles, including relative word frequencies of the subtitle corpus and meta information about the corresponding TV program. Our approach for processing subtitles outperformed the one by Odijk *et al.* [203] with precision$= 0.79$ and recall= 0.77 vs. R-Precision= 0.71. The reported precision and recall values of our approach were the respective means of the 5-fold cross-validation runs. Further limitations of Odijk *et al.* are the presentation of multiple possible destination sites rather than definite decisions and the fact that their training data is solely based on manually annotated and topic-separated content from talkshows, which generally limits the external validity of their approach.

The Second-Screen App: TVInsight

The purpose of the entity linking algorithm described is to provide relevant Wikipedia articles based on an analysis of the current TV program's subtitles. The resulting articles are delivered through a second-screen app - which we called *TVInsight* - with the goal of supporting web-based searches during TV consumption. Therefore, we created a native Windows Phone 8.1 application which proactively displays relevant content (see Figure 4.1.

The front-end contains four menu items where consumers can choose between different categories of information (see Fig.4.1 and 4.5):

- *Info*: general information about the current TV program including channel, program times and a short description (as provided by the EPG).
- *People*: Wikipedia articles regarding people entities including an article preview and portrait picture. A tap opens the corresponding Wikipedia page in a browser view.
- *Wiki*: topic-relevant Wikipedia articles including an article preview.
- *Google*: keyword suggestions relevant to the current program. A tap opens up a browser view showing a list of corresponding Google results.

As subtitles are broadcast and received by our server, they are processed in real-time and the generated content is timestamped. The keyword search suggestions are based on the anchor texts of the detected annotations. Our goal is to offer a feed of articles that convey a much richer and more content-focused experience compared to traditional program information services (e.g. mobile EPG solutions), since we extract people and topic entities from the program and not just from official cast listings and shortened summaries. For synchronizing the app content with the currently watched TV channel, Chuang *et al.* [46] proposed audio fingerprinting. To ensure stable and equal conditions among all study sessions of our study, however, we recorded the documentaries in advance and synchronized

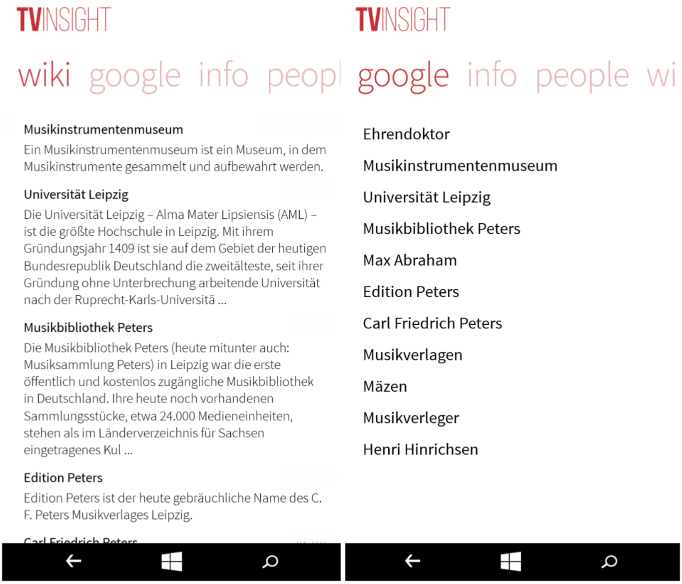


Figure 4.5: *TVInsight* shows Wikipedia articles relevant to the current TV topic (left) and extracted keywords to launch a corresponding *Google* search (right).

the video player with the app content according to the current viewing position through the network connection. As incoming subtitles were processed in real-time by the server, content was pushed out to the app synchronously.

Scenarios

Here we want to depict two scenarios showing how additional content delivered through our second-screen app can be used to augment TV programs and the user experience. By providing background information, the TV content shown can be better put into context by the user. Such context helps to build mental associations between new and existing information and increases the chance for long-term retention. The following scenarios explore this idea with an educational purpose in mind.

Language Learning

Learning language vocabulary is very much topic-driven. The vocabulary required to go grocery shopping is different from the vocabulary of a geopolitical discussion. An obvious extension of our second-screen app is using the keywords extracted from subtitles of foreign TV shows to compile a word list that reflects the topic of the show and can be used in-situ or post-hoc for studying.

General Knowledge Acquisition

The information need of regular TV consumers is pronounced [37, 200]. Results from our online study also show that users often act on cues provided by the TV program to trigger further searches. The conventional approach is to try to remember, for example, the name of a person in order to start a web search at some point bearing the risk of either forgetting to actually perform the search later on or misspelling the keyword. Previous work showed an increased cognitive load when search tasks were performed while watching, which made people miss important information [257]. Our second-screen app based on automated keyword extraction from subtitles allows us to retrieve and display or store important entities for later look-up. While the app proactively triggers web searches and retrieves, for example, relevant Wikipedia articles, contents displayed can be bookmarked for later review while minimizing interruptions in the current TV experience.

4.2.3 System Evaluation

To evaluate the utility of our approach and quantify the effects on consumers' TV experience, we conducted a user study in which we applied our entity linking approach to automatically display relevant contents from Wikipedia in synchronization with the current TV program. We were especially interested in objective as well as subjective effects on program comprehension and the utility of a proactive provision of additional content. In contrast to previous works we did not manually select the content stream but used the actual output of our content generation algorithm applied on an actual TV broadcast. We conducted this user study with the following four hypotheses in mind:

- *H1*: Using *TVInsight* leads to better comprehension of the current program's contents based on *objective* assessments.

- *H2*: Using *TVInsight* leads to better comprehension of the current program's contents based on *subjective* assessments.
- *H3*: Using *TVInsight* is less distracting than manually searching the web using the Smartphone browser app.
- *H4*: Using *TVInsight* leads to a better user experience than using the Smartphone browser app for look-ups.

To compare our approach we tested *TVInsight* against a baseline (no tools available) and the smartphone's browser search capabilities. Participants were asked to watch different documentaries while being able to use the available tools for additional content retrieval.

Method

For this study we employed a repeated-measure design with the tool available for web searches as the independent variable, which resulted in the following three conditions:

- A: No tools available (baseline): participants watching a documentary without a secondary device.
- B: Smartphone browser: participants were given a Smartphone with Internet access and asked to use web searches as they saw fit to retrieve additional information.
- C: *TVInsight* app: participants were given a Smartphone equipped with our prototype and asked to use it as they saw fit.

To avoid learning effects between conditions we counterbalanced the sequence of conditions between participant. Furthermore, the order of the documentaries, ad breaks and questions were randomized to mitigate the influence of participants' preferences and reduce learning effects.

In each condition, participants were asked to watch a documentary. As dependent variables, we measured comprehension in the shape of multiple-choice questionnaires about the documentary content and applied a memory test up to a week after the study. Comprehension questions were designed with regard to testing literal (recalling what has been explicitly stated in the text) and inferential comprehension (requires readers to understand relationships that are not explicitly stated in the text) [17]. Each test consisted of 15 questions with three levels of difficulty: easy, intermediate, and advanced questions, which we asserted during pilot studies. The memory test consisted of statements which required simple true/false responses. Further, we collected subjective feedback through a questionnaire after each study condition as well as through a semi-structured interview at the end of the study.

Participants

We recruited 30 participants (9 female) with a mean age of 22.5 ($SD = 4.0$) years through university mailing lists and social networks. From our participants, 93% indicated German to be their first language, 76% reported to at least occasionally watch TV. The total study took about an hour, for which we compensated participants with 10 EUR.

Apparatus

During the study, all participants were seated in the same room in front of a 22-inch monitor connected to a notebook on which they watched the documentaries, answered the questionnaires and provided subjective feedback. Therefore, we created a proprietary .NET program through which we made sure that the allocation between documentaries, condition, and the viewing sequence was counterbalanced. Questionnaires were directly applied to that same software. For the app and browser condition, we handed out a *Microsoft Lumia 640* on which our *TVInsight* prototype was running as described above. The software further took care of the time synchronization between the played documentary video and delivering the content to the mobile app in real-time to avoid glitches and ensure equally stable conditions for all participants.

Procedure

After participants signed the consent form, we explained the nature of the study and sat them down in front of the screen where we collected basic demographic information through an opening questionnaire. To give an idea of how the comprehension test will look like, three sample questions not related to the following topics were shown. Before each condition we made sure participants could familiarize themselves with the tool at hand and were instructed to use the tool as they saw fit during the documentary in order to improve their comprehension score later on. In each condition participants were asked to watch a 10 min documentary in which we embedded two commercial breaks, 1 min each, in order to simulate

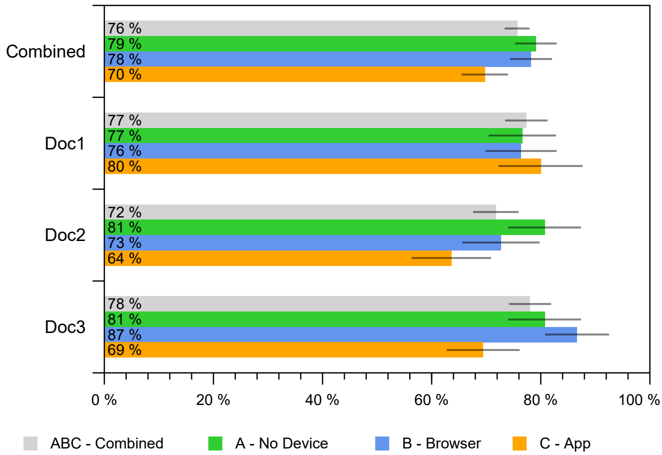


Figure 4.6: Percentage of correctly answered comprehension questions, segmented by condition and documentary.

a somewhat realistic TV experience. Ad blocks were equally spaced throughout the documentary. The content of the ad breaks had nothing to do with the respective documentary and solely served as a time break for participants to use the second-screen device. The documentaries were in German and included topics about birds of prey, a recent naval accident, and about the history of a publishing house. They all aired on the same TV channel several weeks ago. We made sure that no participant had watched any of them before. At the end of a documentary participants had a 30-second window to finish their current web or app search before the comprehension test started. Comprehension questions were designed solely based on the content of the respective documentaries and before the entity linking algorithm was applied to generate the actual second-screen content. The output of the algorithm was deliberately left unmodified to simulate a real-time content extraction setting, thus, some of the articles and search suggestions were evidently irrelevant. The content of the app was synchronized with the viewing experience based on the display time of the respective subtitle lines. We did not allow the use of the app or browser while filling in the comprehension test. After each comprehension test participants were asked to provide a subjective assessment of their comprehension and experience with the tool at hand in form of a 5-point Likert-type scale as depicted in Figure 4.8. After having completed all three condition blocks, we asked participants to provide feedback on the overall

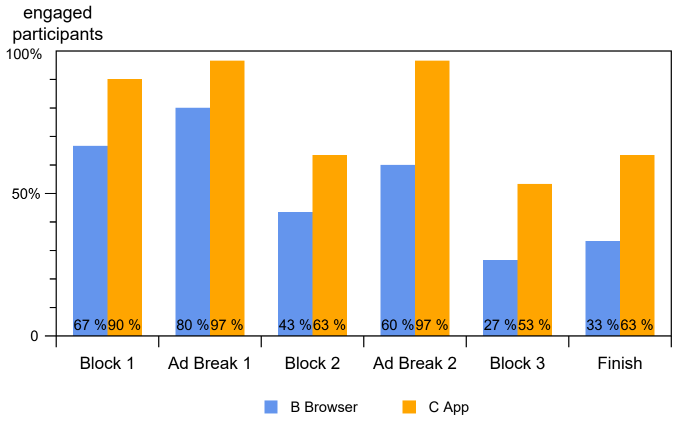


Figure 4.7: App and browser usage across time including advertisement breaks.

experience, what they liked and didn't like about the TVinsight App and for which types of programs they could imagine using it. After seven to eight days time we called participants via phone to answer 18 quick questions taken from the pool of comprehension test questions, but this time the questions were transformed into true/false statements.

Results

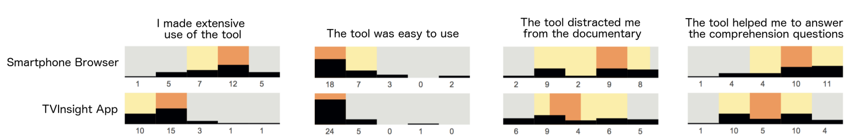


Figure 4.8: Subjective assessment after each condition in form of 5-scale Likert-type scales with *I totally agree* on the left and *I don't agree* on the right. The orange color depicts the median, yellow the interquatile area.

The selected documentaries contained a mean number of 839 words ($SD = 54$), from which our entity linking algorithm extracted a total of 67 ($M = 22.3, SD = 2.5$) relevant and 9 ($M = 3, SD = 1$) irrelevant Wikipedia articles, resulting in a

88.2% hit rate. The Benjamini-Hochberg procedure [134] was applied with a false discovery rate level of 0.05 to account for multiple testing.

Device Usage

Figure 4.7 shows the usage of both smartphone browser and *TVInsight* app split in content blocks, ad breaks, and post-documentary searches across all documentaries. TVinsight was used continuously more often than the browser ($\chi^2 = 25.7, p < 0.001$, Pearson's chi-squared test) which is statistically significant. During ad breaks the usage of the TVinsight app is significantly more pronounced than during the documentary ($\chi^2 = 7.1, p < 0.01$, Pearson's chi-squared test), the same holds for using the browser. Participants used *TVInsight* in 97% of ad breaks whereas only 80% used the browser during the first ad and 60% during the second.

Comprehension Scores

Figure 4.6 gives a detailed overview of the comprehension scores for each condition and documentary. Using the *TVInsight* app results in a statistically significant decrease in performance on the comprehension tests compared to no device usage ($t = 3.05, p < 0.003$, paired Student's t test) and to using the browser ($t = 2.57, p < 0.008$). Using the browser had no impact on the test scores compared to the baseline. Looking at the results by documentary, participants did not differ significantly between the three conditions, whereas the scores of the app users dropped significantly on the other two conditions.

Subjective Feedback

Figure 4.8 lists the results of the Likert-scale statements with participants' subjective feedback with regard to the tool available (Browser or *TVInsight*). When using *TVInsight* participants reported a better subjective comprehension of the program content than when using the smartphone's browser for searches. This difference was statistically significant (Mann-Whitney $U = 679, p < 0.001$). As with regard to ease of use, the app was rated significantly better than using the browser (Mann-Whitney $U = 782, p < 0.001$). Perceived distraction was significantly higher when using *TVInsight* than when the browser (Mann-Whitney $U = 829, p < 0.001$). When asked at the end of the study which tool (or no tool at all) they preferred for being able to answer the comprehension questions, 40% chose *TVInsight* and just 7% the browser.

Memory Test

Most participants left us their phone number with a preferred timeslot for being called for the final memory test a week after the study. We reached 20 out of 30 participants. The percentage of correctly answered questions on the baseline condition dropped from 78% to 73% compared to a slight increase from 64% to 66% on condition C (TVinsight app). However, neither change in the memory scores was statistically significant.

Qualitative Assessment

Overall, participants agreed that using real-time look-ups contributes to the user experience while watching TV. Especially the mechanism of delivering content right in time when the respective topic or person was mentioned on the program was positively mentioned in the interviews by 9 participants, while 6 explicitly stated they liked that the app *"allows for quick results/access"*. The majority rated the app content as informative and well fitting, only 7% disagreed. The user interface was described by 12 participants as *"clearly structured"* and "*easy to use"*. "*Inappropriate topics*" and "*too few people entities*" were the most common negative remarks made about the content. As things to improve participants mentioned a stronger integration of contents with regard to the Google search keywords. Instead of routing the user to the external browser application, search results should be directly integrated into the app, similar to the Wikipedia articles. Also, there is room for improvement with regard to personalization, *e.g.*, through dynamic bookmarks or a customizable entity menu. As to which types of programs participants could imagine *TVInsight* to be especially useful, documentaries were mentioned 21 times, quiz shows or news 8 times respectively, sports 4 times, and movies and entertainment 3 times. Rather inappropriate TV formats would be movies (mentioned 15 times), entertainment shows (3) or TV series (2).

4.2.4 Discussion

Due to the size of our data corpus and our extended entity linking approach, articles extracted from Wikipedia were relevant, which our hit rate of 88% confirmed. Also, participants overwhelmingly stated that articles were a good fit for the documentaries. Further, the real-time aspect of proactively showing additional content with the current program was well received, which created an overall positive TV experience when using *TVInsight*. Also, the content of the documentary turned out to be more relevant in order to answer the comprehension questions. Compared to traditional web searches with the smartphone browser, our second-screen app scored higher both in objective usage measure, but also in subjective ratings.

The extent of how to use the available tools, however, was up to participants in order to self-select an optimal strategy for multitasking. This lead to an imbalance between the browser and app usage with our app ending up being much more frequently used. Some participants barely used the browser for web searches which is why there are few measured differences between the use of no tools and the browser. One reason for rejecting the browser may be the fact that search keywords need to be formulated by the user whereas *TVInsight* continuously presented new keywords along with the current program. Hence, extended usage of *TVInsight* also took effect in the comprehension scores, which is why we need to reject our first hypothesis (*H1*). The objective comprehension scores, therefore, confirm the distracting nature of second-screen applications.

TVInsight was perceived as more distracting than the browser. However, this does not allow us to fully reject *H3*, especially due to the relative imbalance between app and browser usage. When there is no need to actively type and search for keywords, more time is spent in browsing available contents. Furthermore, users were not alerted to new second-screen content, it just appeared. Neate *et al.* [197] found that users would like to be actively notified of new content, preferably with an auditory icon or visual indicator on the TV itself. However, both methods are not feasible for independent second-screen apps with automatically generated content. The extensive usage may also be due to some novelty effect, which causes the app to be more appealing than the actual documentary. Long-term studies in real-world scenarios will be able to give more insights into the nature of distraction.

Since there was a tendency of increasing subjective comprehension of the TV content through our *TVInsight* app, we were able to find evidence in favor of *H2*. This is certainly a reason why many participants found the app was helpful for completing the comprehension test. The well-received app usability and the positive feedback regarding the real-time feature of proactively pushing additional content to the second-screen benefited the overall user experience, which allows us to confirm *H4*; the content relevance and the notion of real-time delivery of additional content lead to a better user experience than using the smartphone browser for look-ups.

We performed the study on documentaries since they are a natural fit to observe learning effects from an educational standpoint. It should be noted, however, that our content extraction algorithm was not specifically optimized for documentaries. Rather, the training set was sourced from a variety of different program genres. Hence, it is reasonable to assume that the participants' ratings and opinions towards the displayed information generalize to other types of programs.

The perceived distraction varied greatly which suggests that the capacity to multitask strongly depended on the individual. Hence, users should be in control of adjusting information density and focus.

While many participants criticized irrelevant content, there was hardly any comment about missing topics. This indicates that precision is more important than recall when evaluating content retrieval algorithms for second screen use cases.

We were not able to detect any statistically significant differences between conditions with regard to the memory scores, but users of *TVInsight* tended to perform equally well in the direct comprehension tests and the memory test a week later. Participants using the browser or no tool at all, on the other hand, tended to perform worse in the memory than in the comprehension test. Further studies will need to be conducted to assess the long-term utility of *TVInsight*, especially when equipped with bookmarking features, where users could save or even share content as they watch it. We are further in the process of developing *TVInsight* with another focus on integrating audio fingerprinting for automatic channel recognition. This feature will be vital before further studies in the large can be conducted by releasing *TVInsight* on mobile app stores.

Concluding, based on insights from our study we derived the following five design guidelines for context-aware second-screen applications:

1. *Synchronization*: second-screen content should be sensitive to the current program context and be delivered in real-time.
2. *Bookmarks*: in order to strengthen memory and support learning scenarios, users should be given the chance to save content for later retrieval. The real-time aspect allows users to bookmark content while watching rather than pushing it to later where it is potentially already forgotten.
3. *Less is more*: too much information distracts users, which is why a second-screen app user interface should be clearly structured with a focus on essential information, but with the option to have users look for deeper content if they so wish.
4. *Precision > Recall*: irrelevant content distracts users and harms the overall user experience more than possibly missing information.

5. *Personalization*: the need for information and the capacity to multitask strongly depend on the individual. Hence, users should be in control of adjusting information density and focus.

4.2.5 Study Conclusion

Second-screen apps have become increasingly popular in recent years. Instead of requiring users to install numerous different apps tailored to each of their favorite TV shows, we proposed a context-sensitive second-screen app that creates additional content automatically by linking existing resources and pushing that content to the user in real-time when it is most relevant. We, therefore, described an entity linking algorithm that extracts keywords from live subtitles and uses Wikipedia to provide additional program information. Having built a large database with data from 30 TV stations over the course of four months, we were able to extract particularly relevant content. The utility of the prototype was confirmed in our user study where we investigated its effect on user experience and content comprehension. The resulting insights can be used by app developers to create second-screen apps that take advantage of existing content resources, and bear the potential to keep user distraction at a minimum by proactively providing content in a context-sensitive way. By expanding the number of TV stations recorded by the server the subtitle corpus can be refined in order to improve the entity linking and therefore the relevancy of the provided contents. Additional resources, such as social network chatter, historical archives, current news articles, or product databases can further be linked to provide users with a holistic second-screen experience.

Our prototype shows the feasibility of a technology that picks up contextual clues from the environment and augments the user's current experience with relevant information. We believe such techniques could be used to augment not only TV experiences, but also lectures, meetings, and conversations, wherever technology can access real-time audio or another type of content stream in order to trigger contextual queries. The study described in this chapter focused on external context, *i.e.*, information from the environment, in order to support information intake. Another way of thinking about context is taking into account the user's internal processes. In Chapter 3 we already discussed how attention can be tracked and predicted in a technology context. In the next section, we describe our approach to go one step further and investigate how technology can detect and make use of users' cognitive states. By considering both external (*e.g.*, content options) and internal (*e.g.*, receptiveness) context, technology can help bring the two together and optimize for efficient information intake.

4.3 Chapter Summary

In this chapter we investigated the use of knowledge about the content, which the user is currently engaged with, to support people absorbing information. Awareness of the content at the time of exposure allows systems to provide relevant services. Our motivation was to investigate whether providing additional background information helps people put content better into context. Such context linking allows users to create mental associations between existing and new information, which increases the probability for long-term retention.

We set out to test our assumption by developing a system for augmenting the TV experience. While watching TV people often turn towards so-called second-screen applications to look up additional information about the current program. By making use of subtitles our system extracts keywords through an advanced entity linking algorithm. These keywords are then linked to additional online resources, such as relevant articles on Wikipedia, which are sent to and displayed by a corresponding mobile app. We assessed the utility of using such an app in a lab study. With 88% of the automatically generated content being highly relevant, our entity linking algorithm outperformed previous approaches. Besides using neural networks for linking entities such performance was due to the high volume corpus of 45,000 hours of TV shows which we collected over the course of four months. The lab study yielded effects on user experience and content comprehension, which demonstrates the feasibility of using content awareness as a context dimension to enhance information intake (**RQ3**). We derived a set of guidelines that can be used to build context-sensitive second-screen apps, whose applicability is not necessarily limited to the special use case of TV consumption. Real-time synchronization between content displayed, bookmarking capabilities, and taking into account user preferences can help users create content associations beyond what is displayed on a primary device. By providing additional services in line with the content that users momentarily engage with, the research probe we developed serves as an example of technologies that exhibit context- and content-awareness.

Chapter 5

Cognition-Awareness

So far we have investigated users' context with regard to attentiveness towards their mobile phones and to the content, they currently engage with. While such factors allow us to infer information about people's cognitive states at the time, there are other aspects that influence cognitive performance. People's alertness, attention, and vigilance are highly variable and subject to systematic changes across the day. These fluctuations—in part caused by *circadian rhythms*—impact higher level cognitive capacities, such as perception, memory, and executive functions.

In this Chapter, we investigate ways to measure these rhythms and introduce the notion of considering the user's cognitive states as part of the context. Cognition-aware systems detect different aspects of mental information processing, such as engagement, cognitive load, memory, knowledge, and learning [36]. For effective information intake, the cognitive state of the user is crucial. Whether information is effectively processed and retained long-term or whether it is barely brought to the user's attention highly depends on the user's current capability to focus [234]. By identifying productive phases during the day, cognition-aware systems can suggest times of the day to engage with information consumption. Awareness of cognitive states can further be used for task scheduling to support work rhythms [21], preventing interruptions [138, 180], and inducing flow states [52].

Typical methods that extract patterns from diurnal fluctuations in alertness levels are time-intensive and take place in artificial lab settings. For less intrusive and time-consuming elicitation of these patterns, we created a toolkit that can be deployed on mobile devices. In this Chapter, we present a study, in which we elicit users' circadian rhythm of alertness. To allow researchers and system builders to create cognition-aware systems we released the toolkit as open source library, that can be integrated into existing applications and games.

Patterns of alertness levels across the day provide systems with a general awareness of diurnal fluctuations of users' cognitive capacities. These patterns can be combined with information about the momentary cognitive state based on the current user context. In the second part of this Chapter, we present a conceptual framework for building algorithms that detect users' momentary states. The framework is derived from our approach to detect bored states (Section 3.3), but allows for more general applicability.

In this Chapter, we address the following research question:

- **RQ4**: How can technology be used to elicit the user's circadian rhythm of attention and cognitive performance?

Parts of this chapter are based on the following (pending) publications:

- T. Dingler, A. Schmidt, and T. Machulla. Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance. volume 1, pages 47:1–47:15, New York, NY, USA, Sept. 2017. ACM
- T. Dingler. Cognition-aware systems as mobile personal assistants. In ACM, editor, *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '15 Adjunct, New York, NY, USA, 2016. ACM

5.1 Related Work

While an in-depth introduction to cognition-aware systems can be found in Section 2.5, this Chapter more specifically refers to related work in sleep/wake regulation and circadian rhythm elicitation.

Human cognitive performance is affected by sleep-wake homeostatic and by an internal circadian rhythm. Kleitman [156] was among the first to establish a link between cognitive performance, chronobiology, and sleep. He found speed and accuracy of cognitive performance to follow diurnal variations with highest performances in the afternoon and poorest in the early morning. In later studies, he noticed that performance is dependent on body temperature [158].

Traditional methods to assess the circadian rhythm include extensive lab experiments, which can take weeks of being in controlled environments. Other methods can be equally cumbersome or even unpleasant, such as sleep-wake protocols or physiological markers (*e.g.*, dim light melatonin onset, rectal temperature monitoring, cortisol level measurements [135, 158]). To measure people's alertness level in a less invasive way, the psychomotor vigilance task (PVT) has been proposed [67]. It measures the simple reaction time to a visual stimulus. In its original version, the task lasts ten minutes and is thus a test of vigilance (the ability to sustain attention over time) as much as a test of psychomotor speed. During the task a visual stimulus is presented randomly every two to six seconds. While the original experiment setup uses a physical button [67] to provide a response, an implementation for touchscreens using the *touch down* event has been proposed by Kay *et al.* [152].

Abdullah *et al.* [2] recently demonstrated the general feasibility of using PVT in-the-wild to measure diurnal alertness fluctuations. They showed alertness fluctuations by harvesting mobile phone data with influencing factors being time, body clock type, sleep, and caffeine intake. Murnane *et al.* [192] correlated these same fluctuations with mobile app usage patterns.

To the best of our knowledge, our current work is the first to provide a robust statistical model that specifies the contribution of time-of-day to different performance measures and relates the measured fluctuations in alertness to prevalent theories of sleep/wake regulation. We investigate an in-the-wild approach to elicit performance variations based on three quick tasks performed on users' smartphones, with which we aim to eliminate the need for extensive time and resources spent on sleep labs studies or artificial tasks. By having created and released a toolkit we allow measurements of cognitive performance to be included in a wide range of applications and games, which allows researchers and application builders to add cognition-awareness to their systems.

5.2 Eliciting the Circadian Rhythm of Alertness

To extract patterns of diurnal alertness fluctuations, we created and validated a test battery that can be deployed on mobile devices. It consists of three sustained attention tasks as described below (see Figure 5.1). To validate this test battery, we built an Android app (Android version 4.1 or higher) that administers the tasks as well as a number of short questionnaires concerning the users' demographics, sleep, and alertness self-assessment. The app prompts the user at random times during the day through notifications to complete the test battery. The time between task reminders is between 60 and 90 minutes. To respect sleep times notifications are only scheduled between 8 am and 9 pm. We included a logging mechanism that saves measurements locally on the device and transmits the logs to a remote server when a WIFI connection is available.

We selected three tasks for inclusion into the toolkit - a psychomotor vigilance task (PVT), a go/no-go task (GNG), and a multiple object tracking task (MOT). We wanted to assess the tasks for their utility as quick measurement tools that cause as little interruption to users' daily routine as possible. We, therefore, limited the first two tasks to about one minute each and the MOT to two minutes. The **PVT** is the gold standard for assessing alertness levels. It measures the simple reaction time to a visual stimulus. In its original version, the task lasts ten minutes and is thus a test of vigilance (the ability to sustain attention over time) as much as a test of psychomotor speed [67]. During the task a visual stimulus is presented randomly every 2 to 6 seconds (see Figure 5.1, *left*). While the original experiment setup uses a physical button [67] to provide a response, our touchscreen implementation uses the *touch down* event as proposed by Kay *et al.* [152]. The **GNG** task falls into the class of choice reaction time paradigms. It uses two or more distinguishable stimuli, each associated with a unique answer option - in our case, a plain green circle, for which the participant needs to perform a speeded *touch down* gesture (”go” trial) and a patterned circle, for which this behavior needs to be inhibited (”no-go” trial, shown in Figure 5.1, *middle*). Hence, this task measures reaction time, as well as executive functioning. In our implementation, we use between 8 and 12 stimuli, approximately half of which are no-go stimuli, appearing at random intervals of 1 to 8 seconds. If ignored, stimuli are shown for a maximum of three seconds. The **MOT** is a strenuous sustained attention task that requires participants to divide their attention across multiple moving objects [219]. In our implementation, eight blue circles are shown. A subset of four target circles briefly flashes to indicate the objects to be

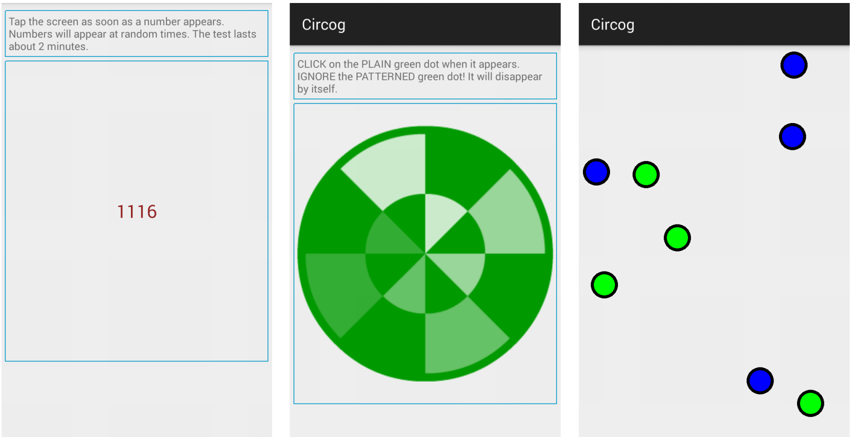


Figure 5.1: Our toolkit comprises three tasks to measure alertness and cognitive performance variations across the day: a Psychomotoric Vigilance Task (left), a Go/No-Go task (middle), and a Multiple Object Tracking task (right).

tracked (see Figure 5.1, *right*). Then, all circles start moving in random, but linear directions. After ten seconds the circles stop and the test person has to identify the target circles. The task is repeated five times, the performance measure is the number of correctly identified targets.

We released the app and the contained toolkit library under an open source license on *Github*[14]. By including it in their source code, application builders can *register* their application to collect performance measurements. These can either be provided by their own application (*i.e.*, reaction times in games) or, alternatively, one of the toolkit tasks can be triggered to collect these measurements over time. Their application can further *request* the user's *current performance state*. If the application has already collected enough measurements (usually across a week), an individual assessment of the user's performance is returned. As long as there are not sufficient data points available to derive a robust individual model of the user's performance fluctuations, the library returns a generic version model (as described in this paper) adapted to the user's current timezone.

[14] `https://github.com/til-d/circog`

5.2.1 User Study

We conducted a user study to validate the effectiveness of the three attention tasks regarding their ability to measure systematic fluctuations in alertness within a short duration of time. Since our goal was to measure fluctuations across the day we opted for an in-the-wild study, where participants were asked to perform the aforementioned tasks in their daily context. As dependent variables, we collected task performances together with the time of day when the tasks were completed, subjective sleep and alertness assessments, and task preferences.

Procedure

We recruited 12 participants (4 female, mean age: 24 ($SD = 2.67$)) through university mailing lists. All participants were briefed about the purpose of the study and provided informed consent. The task order was randomized each time the app was opened. A service kept running in the background that managed the posting of notifications to remind users to perform the task sequences from time to time, up to six times a day. The prompts were shown in the notification drawer until clicked or dismissed. A click on the notification launched the task sequence. Before the first task sequence of the day, a survey was shown with questions about the user's wake-up time, the number of hours slept and rated the quality of sleep (1=*poor*, 5=*very good*). Each task sequence was preceded by a short self-assessment regarding "*How alert are you feeling right now?*" (1=*super sleepy*, 5=*super alert*) and a checkbox labeled "*I had a caffeinated drink within the last hour*". The study ran for a total of 14 days; participants were free to start at any time by installing the app and completing the tasks for at least seven days. We awarded 50 cents for each task sequence completed at a maximum of six sequences per day, resulting in up to 42 EUR. At the end of the study, we sent out a questionnaire to assess participants' subjective impressions of the different task types. For each task, we collected Likert-style feedback on participants' evaluation of task difficulty, exhaustion, and fun.

Results

In the following, we analyze each task according to its effectiveness to measure systematic changes in cognitive performance across the day. We further provide an assessment of the influence of caffeine, sleep and the accuracy of participants' self-assessments, where applicable. On average, participants performed the tasks on nine days (SD = 3.9) with a minimum of 2 and a maximum of 13 days, resulting in a total of 367 PVT, 364 GNG, and 367 MOT tasks. We removed incomplete tasks and data points, for which the subjective alertness rating was missing.

We examined all performance measures for the influence of the sleep/wake homeostatic and the circadian process. The homeostatic process should manifest itself in a performance deterioration with time spent awake. For instance, simple reaction times in the PVT should increase throughout the day. Therefore, we first examined the data for a linear trend over time. Trend analysis was performed by fitting a linear mixed model to the raw data with the measure of interest as the dependent variable and the fixed factors *time of measurement*, *self-rated alertness*, *consumption of a caffeinated drink in the previous hour*, *sleep duration*, and *self-rated sleep quality* as well as the random factor *subject*. p-values were obtained by using a likelihood ratio test of the full model against a null model without the fixed effect of interest. If this comparison was non-significant, the fixed effect was excluded from further analysis. We did not find any significant interaction effects between the studied factors.

If a linear trend of time was found, it was removed from the data before further analysis. In a next step, we looked for variations in performance due to the circadian process. The circadian process should result in non-linear variations of cognitive performance across the day. For instance, performance should decline in the early afternoon and improve towards the early evening. For this analysis, we fitted a second linear mixed model with the ordered categorical predictor variable *hour of the day* as a fixed factor and *subject* as a random factor. For each fit, we report the results of an omnibus test (Analysis of Variance (ANOVA); indicates whether performance differs between any two-hour slots of the day) as well as contrasts between successive hours (indicate whether performance changes between hour x and hour x + 1). The level of significance was adjusted for multiple comparisons using the Holm-Bonferroni procedure. In the following, we first report the influence of time/hour of day on the different measures of performance, followed by an analysis of the influence of the control variables.

Variability of Performance Measures across the Day

Psychomotoric Vigilance Task (PVT). Figure 5.2 shows average reactions times (Reaction Time (RT)) as a function of time. We find that participants' RT increase by 1.9 ms (± 0.7) per hour of day (*time*: $\chi^2(1) = 6.7, p = 0.009$), a performance deterioration reflecting the homeostatic process. After this linear relationship

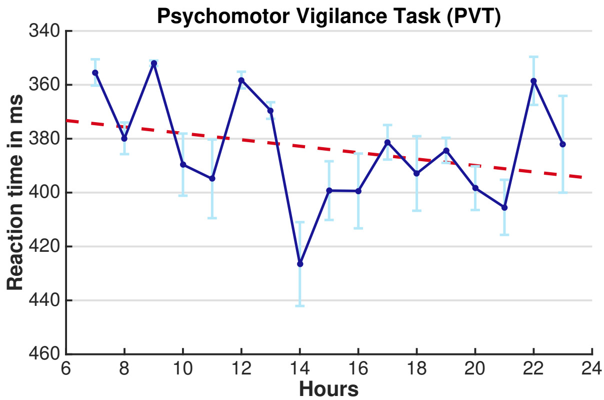


Figure 5.2: Performance variations across the day in blue: mean reaction times from the Psychomotor Vigilance Task (PVT). Error bars indicate the standard error of the mean. The red line depicts a linear fit to the data.

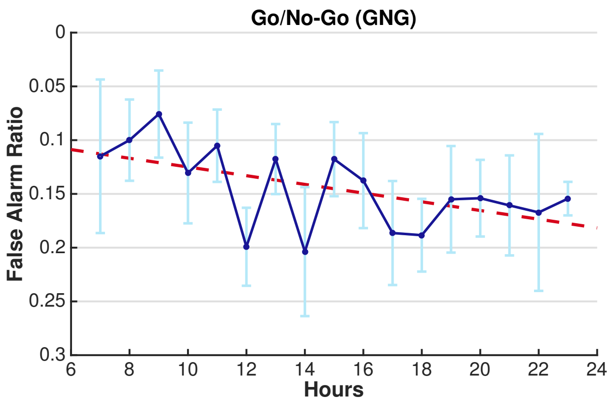


Figure 5.3: Mean false alarm ratio from the the Go/No-Go (GNG)task.

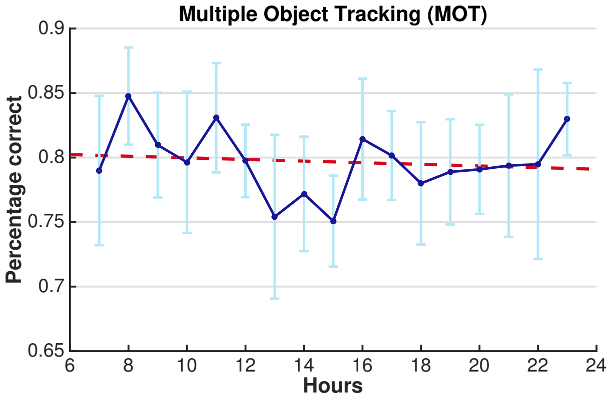


Figure 5.4: Percentage of correctly identified targets from the Multiple Object Tracking (MOT) task.

is removed from the data, we find a significant effect of hour ($F(16, 3175.3) = 1.8, p = 0.016$), indicating that there are further differences in RT performance across the day. In particular, there is a significant increase of 54 ms from 1 to 2 pm ($t(3173.6) = 3.3, p < 0.001$). Mean RT are highest just after 2 pm and lowest after 10 pm (399 ms vs. 316 ms). This agrees with previous reports regarding a post-lunch dip in alertness and performance [259] as well as reports of improved performance towards the evening in younger adults (which our sample predominately consists of).

We further find that consuming a caffeinated drink decreased RT on average by 17.6 ms ($\pm$ 13.7 ms; $\chi^2(1) = 10.8, p = 0.001$) and that subjective alertness ratings are a significant predictor of PVT RT: RT decreases by 10.4 ms ($\pm$ 4) per level of alertness rating ($\chi^2(1) = 5.8, p = 0.015$).

Go/No-Go-Task (GNG). For the GNG task we analyzed RT as well as false alarm rates. False alarms are ”go” responses to ”no-go” stimuli. Both indices show a pattern similar to the PVT RT. Figure 5.3 shows the average false alarm (False Alarm (FA)) rate across the day. For the RT, there is a linear increase of 0.8 ms/hour throughout the day, resulting in an overall difference of about 13 ms across a typical period awake. Additionally, there is a pronounced deterioration of performance in the early afternoon. However, in contrast to the PVT, neither the linear increase in RT is significant ($\chi^2(1) = 0.6, p = 0.44$) nor are there any significant differences in performance between consecutive hours.

A possible reason for this finding is that the GNG task is subject to speed-accuracy trade-offs. When participants are compromised in their alertness, they may either decrease their reaction speed or compensate by keeping the same level of reactivity but concede to making more false positive decisions – that is, reacting to non-target stimuli. To test this hypothesis, we performed a linear mixed effect model analysis with *RT* as the dependent variable and *false positive rate* as a fixed and *subject* as a random factor. We find strong indication of a speed-accuracy trade-off: as the false positive rate increases from 0 to 1, RT decreases by 129.6 ms ($\chi^2(1) = 9.6, p = 0.002$). Further, we find that participants become increasingly likely to make a FA as the day progresses ($\chi^2(1) = 4.3, p = 0.038$, FA increase by 3% ($\pm$ 2%) per hour of day). This reflects increasing impulsivity, that is, a failure to inhibit wrong responses.

Fluctuations in the FA rate can result from changes in perceptual sensitivity or from adjustments in participants' answer patterns. To distinguish between these two possibilities, we analyzed the data using signal detection theory [117], which allows for the joined analysis of RT and the false positive rate in terms of two measures: d-prime and criterion. d-prime measures sensitivity. It indicates how well an observer discriminates between signal and noise (in our case, between target and non-target stimuli). The criterion is a measure of response bias, or in other words, the tendency to react over the tendency to not react. It is important to note that a change in criterion is a purely behavioral adjustment with no concomitant perceptual change.

Our analysis reveals that participants' response tendency does not change significantly over the day - neither is there a linear shift in criterion ($\chi^2(1) = 2.3, p = 0.129$) nor does the criterion change within successive hours. However, we find that sensitivity varies: d-prime decreases by 0.015 per hour of day ($\chi^2(1) = 4.2, p = 0.04$). This means that participants loose in their ability to discriminate between stimuli as homeostatic sleep pressure accumulates.

Multiple Object Tracking (MOT). Figure 5.4 shows the average proportion of missed targets. We find no evidence for performance to worsen linearly throughout the day ($\chi^2(1) = 0.01, p = 0.91$). However, we find significant differences between successive hour bins. In particular, there is a significant improvement in performance between 3 and 4 pm ($z(365) = -4.1, p < 0.001$) just after performance had reached its daily low, when 25.6% ($\pm$16.6%) of targets are misidentified. Performance is best in the morning at 8 am, when only 12.9% of the targets are misidentified.

Statement	PVT	GNG	MOT	p
This task **took me a lot of time** to complete.	1 (1.08)	2 (1.03)	**3 (1.27)**	< 0.001
This task was **difficult** to complete.	1 (0.79)	2 (0.97)	**3.5 (1.71)**	< 0.01
This task was **exhausting** to complete.	1.5 (0.7)	2 (1.0)	**3 (1.8)**	< 0.05
This task was **fun** to complete	2 (2.0)	2 (1.77)	5 (2.49)	0.073

Table 5.1: Subjective user ratings in 7 point likert-style: 0 = totally disagree, 6 = completely agree (median values with SD)

Subjective Feedback

For evaluating participants' subjective assessments of the tasks regarding difficulty, exhaustion, and fun (0 = *totally disagree*, 6 = *completely agree*) we applied Friedman tests with post-hoc analyses using Wilcoxon signed-rank tests with a Bonferroni-corrected significance level set at $p < 0.017$.

Table 5.1 gives an overview of participants' subjective feedback. There was a statistically significant difference in perceived **task difficulty** ($\chi^2 = 10.7, p = 0.005$) with MOT ($Mdn = 3.5, SD = 1.71$) being rated more difficult than PVT ($Mdn = 1, SD = 0.79, Z = -2.68, p = 0.007$). With regard to perceived **exhaustion**, we found a statistically significant difference ($\chi^2 = 6.9, p = 0.032$) with MOT ($Mdn = 3, SD = 1.8$) being rated as more exhausting than PVT ($Mdn = 1.5, SD = 0.7, Z = -2.413, p = 0.016$). As to which task was the most **fun** to complete we did not find any statistically significant difference ($\chi^2 = 5.243, p = 0.073$), but MOT reached the highest rating ($Mdn = 5, SD = 2.49$). 90% of participants explicitly stated to prefer the MOT task (GNG: 10%, PVT: 0%).

5.2.2 Discussion

Our validation study shows the tasks' feasibility to extract circadian fluctuations and a homeostatic decrease in performance across the day. To be as unobtrusive as possible we limited all tasks to 1-2 minutes in duration, which at the same time limited the amount of data collected. Nevertheless, using the PVT we found both effects, while the GNG showed a linear decrease in performance and the MOT resulted in circadian variations. As alertness decreases we find slower processing speed in the PVT, the GNG shows poor target discrimination and decreased response inhibition as evident in impulsive responding, and the MOT shows a decrease in the ability to divide attention across space. We confirm that PVT is most economical as it provides the most data in the shortest time. It is, however, subjectively rated worst, likely due to its monotonous and simple nature. Adding GNG and MOT to our toolkit provides a more holistic assessment of cognitive performance. They extend simple alertness assessments by adding higher cognitive functions (such as executive control and divided attention). Subjectively, participants preferred MOT and GNG as they are perceived to be more challenging. In contrast to the PVT, their nature allows them to be adapted in terms of their difficulty and the challenge they pose to the user. This makes them attractive to be integrated into applications and even games. Their metrics (hit/miss and reaction time) can be obtained through a variety of mechanisms, such as game performance or typing behavior. Integration of such metric collection in other applications will allow us to build implicit, cognition-aware systems, where explicit measurement that requires users to interrupt ongoing activities may become obsolete.

5.2.3 Study Conclusion

We developed a mobile toolkit for assessing alertness and cognitive performance of users by using a combination of three tasks. The toolkit allows those tasks to be performed in-the-wild and effectively extracts circadian fluctuations and a homeostatic decrease in performance across the day. We show that with a small dataset and with the sensitive statistical methods we applied we can reconstruct users' circadian rhythm of alertness and cognitive performance. We released the toolkit as an open source library to allow researchers in psychology and medicine, as well as application builders, to create cognition-aware systems. Systems that are aware of users' performance rhythms can adapt interface complexity and information bandwidth to match the user's current state. Due to the task variety, similar metrics can be collected in games and applications and therefore make data collection even less obtrusive.

In future work, we are planning on collecting phone usage data along with cognitive alertness metrics. Using the performance assessments as ground truth we can train models that predict the user's alertness based on the individual's circadian rhythms. In the last part of this Chapter, we will present and discuss a framework for obtaining cognitive states and building machine-learning models to predict those states in-situ.

5.3 A Conceptual Framework to Derive Cognitive States

When measuring cognitive states and activities we need to rely on indirect observation, inference, and self-reports. Therefore, we adapted a classic machine-learning approach to a framework for building algorithms capable of detecting and predicting cognitive states. This is basically done by correlating sensor data with an observed user state through supervised machine-learning techniques. The framework depicted in Figure 5.5 entails three steps:

1. Collecting ground truth together with context sensor data.
2. Extracting features from that sensor data.
3. Training and applying classification and prediction models.

In the following each of these three steps will be described in more detail along with a few examples.

5.3.1 Ground Truth Collection

To be able to correlate context data gathered from ubiquitous sensors to cognitive states, we need a labeled dataset. Such a dataset consists of feature entries and a corresponding label containing the ground truth. The ground truth entails the existent cognitive state of interest (*e.g.*, current level alertness) at a certain time with the simultaneous occurrence of feature characteristics. Such ground truth in form of the user's current cognitive state can be assessed in three ways: 1) through **self-reports**, 2) direct **observation**, or 3) **inference**.

Self-reports

Self-assessments are a widely used way to collect information about users' feelings, thoughts, activities, and experiences without having to invest much time and resources building a complex recognition system. Assessments can take the form of daily diaries or users can be asked to stop at certain times during the day to give an assessment of their momentary experience. This method is also called the Experience Sampling Method (ESM) or Ecological Momentary

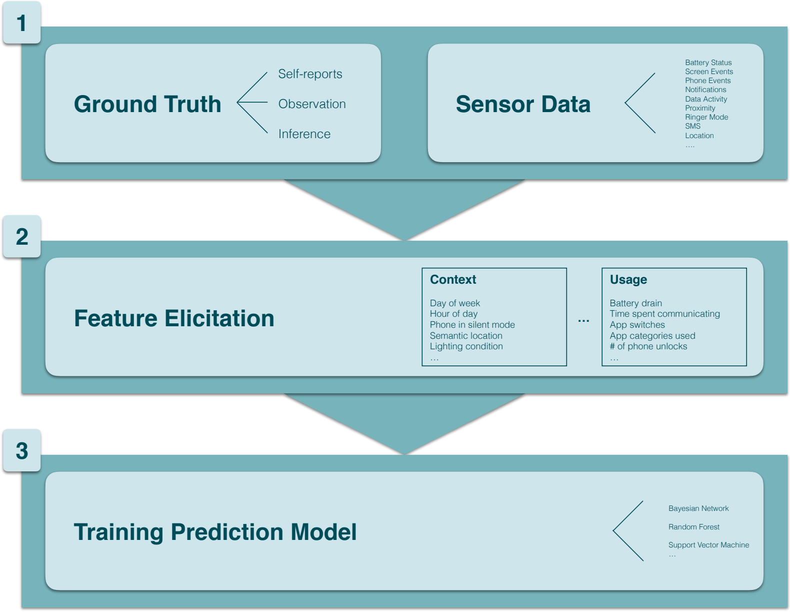


Figure 5.5: Conceptual framework to derive cognitive states in three steps: 1) collecting ground truth together with sensor data, 2) extracting meaningful features from the data collected to create a labeled dataset, and 3) training prediction models that can be integrated in applications.

Assessment (EMA). Experience sampling is used to gather in-situ feedback in the form of short statements. With the availability of mobile devices, such statements can easily be collected throughout the day [50]. The immediacy of these short self-assessments reduces the cognitive biases associated with other recall-based self-report techniques, such as interviews, surveys, or diaries. Experience sampling has been shown to have both internal and external validity, however, the interruptions caused by the sampling process have been recognized as an issue [130].

In a ubiquitous computing setting, there are two general approaches to trigger ESM probes: interval- and event-triggered. Self-assessment surveys can be triggered through notifications, for example, in regular intervals, *e.g.*, every waking hour, or based on device events, such as phone unlocks or incoming

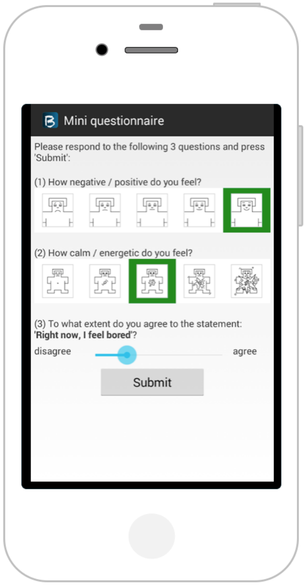


Figure 5.6: Ground truth collection in form of experience sampling prompting users to provide self assessments.

phone calls. Intille *et al.* [140] recently presented an experience sampling method for smartwatches, in which they found that interactions on the watch throughout the day have a high compliance rate, but are also perceived as highly distractive. Distractions and interruptions caused by triggered self-assessments should be kept to a minimum, which is why on mobile devices an ESM probe should contain only a limited number of questions. Likert-style ratings (*e.g.*, "*right now I feel bored*" (Figure 5.6)) and multiple choice questions are preferable to free text answers, especially on mobile devices where text input is rather cumbersome. Also, self-reports can be subject to individuals' biases and (un-)intended falsifications.

Observation

A more reliant way of collecting ground truth without the risk of interrupting the user is through direct observation. Since this technique is rarely feasible when conducting field studies in-the-wild, we often need to resort to implicit observation. Through logging user actions on mobile devices, for example, we can

Figure 5.7: Ground truth collection in form of observing or logging what users actually do (e.g. app interactions).

document user behavior in everyday context without interruptions. For example, we can directly record app usage behavior on users' mobile phones (Figure 5.7)) in order to gather information about what times during the day users consume news, engage with games, or communicate with peers. Boehmer *et al.* [27] describe an approach to analyze mobile app usage across the day by logging user data and also taking into account contextual information, such as location. However, many behaviors or intents are not directly observable. A lot of user actions happen in the real world and are not directly captured by mobile sensors, which is why we at times need to rely on inferences from sensor data rather than from observable behavior.

Inference

As opposed to inferences from observable behavior, inferences from sensor data are made on a lower level abstraction. Physical activity, for example, can be derived from accelerometer data (Figure 5.8), where movement patterns can

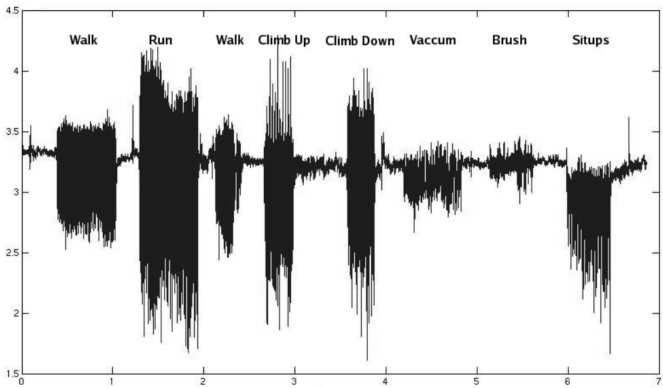


Figure 5.8: Ground truth collection in form of inferring user states or activities from raw data (*e.g.*, accelerometer data showing sleep, walk or running activities (src.: [221]).

be mapped to activities, such as standing, walking, running, climbing stairs, and brushing teeth [221]. Therefore, classifiers are applied in order to extract the ground truth in form of discrete states from raw sensor data. The AWARE framework by Ferreira *et al.* [98] gives researchers a tool to obtain and pre-process mobile phone sensor data to derive user activities. Similarly, Mathur *et al.* [184] used EEG data to collect ground truth on user engagement to correlate it in a second step to mobile phone usage. Thus, making inferences from sensor data requires an intermediate step of pre-processing that data in order to extract ground truth states.

When collecting ground truth data either through self-reports, observation, or inference, we record additional context data through mobile sensors as they are present in phones and wearables; these typically possess rich sensing capabilities, are near-constantly available, and provide means through Application Programming Interfaces (APIs), for example, to access sensor states, observe user interactions, and use the device's output capabilities to provide users with feedback and prompt for explicit user input. Examples of sensor data can be acceleration, lighting conditions, phone usage intensity, apps being used, or timestamps. So whenever the ground truth is collected by one of these three means, we take a snapshot of the available sensor data at the point of collection. Sometimes it makes sense to take into consideration the sensor data in a certain time interval (*e.g.*, 5 minutes) before or after the ground truth is being collected.

5.3.2 Feature Extraction from Sensor Data

Sensor data provides us with contextual information, based on which we can now define features that may be relevant to analyze in co-occurrence with the ground truth collected, *i.e.*, the cognitive state. A feature is described as an individual measurable property of a phenomenon observed [23]. A phenomenon as measured by sensors can be the source for multiple features: from a timestamp, for example, we can extract multiple features, such as the month of the year, the day of the week, or the hour of the day. Similarly, app usage data gives us insights not only into the kind of app, used, but also contains information about the app category (*e.g.*, news, games, productivity apps), usage duration, sequence, and frequency. Defining informative, discriminating and independent features is a vital step for training effective models. It often requires experimenting with multiple possibilities and combining automated techniques with intuition and domain knowledge. The goal of defining and extracting features is to produce a labeled dataset, in which each feature is listed in combination with the characteristics of other features along with the co-occurring ground truth as a label.

5.3.3 Training and Applying Prediction Models

With this labeled data set we can now set out to build detection and prediction models. By using machine-learning techniques we can reconstruct distinct usage patterns that correlate with the user's cognitive state as described by the ground truth. For the work described in this thesis, we have achieved good results with Random Forest classifiers or Decision Trees depending on the types of data collected [87, 214]. *Weka* [136] has proven to be a powerful software offering a variety of tools for analyzing data, training algorithms and exporting prediction models. The more data available for training these models, the better the accuracy of the prediction. Once the cognitive state in question can be detected and distinguished with sufficient accuracy, we can export the prediction model and integrate it into live systems (*e.g.*, mobile phone apps) where its applicability can be tested in the wild.

This overall procedure focuses on creating general prediction models, but once we have a proof-of-concept of mapping ground truth to sensor data, the training of the algorithm can also be conducted on the fly with users' personal datasets. Therefore, the system needs to continue collecting ground truth along with sensor data in order to train these models in-situ. Hence, individual prediction models are feasible to be created and applied directly on the user's device, which has the additional advantage of *privacy by design*, since no data necessarily need to leave the user's device.

5.3.4 Applying the Framework: The Augmented Narrative

We applied the aforementioned framework to an experiment in measuring reader engagement while reading text [168]. Based on physiological sensors, such as eye movements, skin temperature, heart rate, and GSR or EDA we investigated whether it is possible to detect how immersed people were in the texts they were reading. Not only could this give us valuable information about the current text and the reader's levels of interest, but also allow us to adapt the reading UI in real-time by staging interventions to bring the reader's diverted attention back to the text. Reliable detection would allow electronic reading interfaces to take into account readers' engagement levels by providing, for example, additional textual information in engaged states or try to bring back readers' attention by adding sound effects to trigger mental imagery.

Ground Truth and Sensor Data Collection

In the experiment we had five readers (3 female) with a mean age of 29 ($SD = 3$) read six different texts of different presumed engagement levels. For ground truth collection we assessed participants' level of immersion in the text by applying an immersion questionnaire after reading each text as proposed by Jennett *et al.* [148]: in this questionnaire participants self-reported their subjective levels of empathy, frustration, boredom, and enjoyment on a Likert-style rating scale from 1 to 5. Each text was further rated by participants according to their subjective engagement. Along with the self-reports on immersion as ground truth, we collected contextual data by equipping study participants with a temperature sensor mounted on the nose and behind the ear, attaching a sensor to a finger and a heart rate monitor to the wrist (see Figure 5.9). Additionally, we attached a Tobii EyeX eye tracker to the bottom of the screen, on which the texts were shown, in order to record eye movements.

Feature Extraction

Hence, for each of the six texts, we collected ground truth on readers' immersion and engagement along with contextual data in form of physiological sensor data.

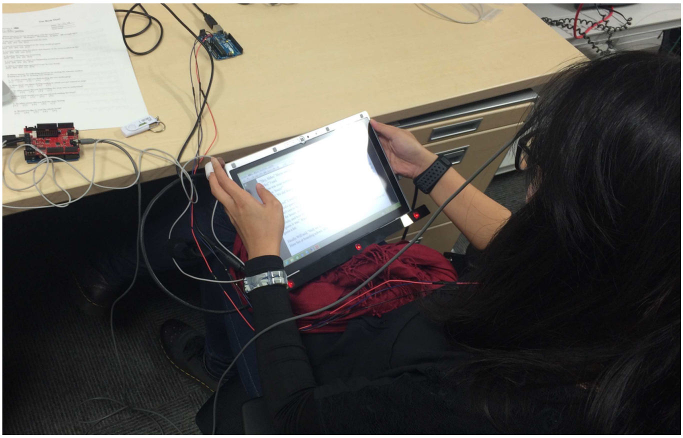


Figure 5.9: Study participant reading on tablet while collecting temperature, GSR, heart rate, and eye movement data.

In the next step we extracted a range of features from the sensor data: for one, we extracted nose temperature changes over time and calculated the slope of the temperature curve using linear regression, which showed to be a significant indicator for engaging *vs.* non-engaging texts over all users ($p = 0.03, F = 0.84$). From the eye movement data we extracted blink frequencies and calculated the frequency change over time, which also turned out to be a significant indicator for engaging *vs.* non-engaging texts ($p = 0.05, F = 0.96$). Further features we extracted and that looked promising were: number of fixations, the median of fixation duration, variance of fixation duration, and the number of saccades that were not in the main reading axis. Since we did not find strong correlations between engagement levels and GSR or heart rate data, we did not use them as features for the classification.

Classification

Based on the features extracted we trained a support vector machine (SVM) with a radial basis kernel. Therefore we applied a leave-one-out user independent strategy by training the model on the data provided by four users and testing it on the remaining one. We approached the classification as a two-class problem: engaging *vs.* non-engaging. Further, we trained a second SVM to assign an engagement value from 1 to 6 (1=low, 6=high engagement) to each text by training on five texts and testing on the remaining. The goal was to automatically sort the texts according to user preferences based on engagement scores.

In a leave-one-out cross-validation the SVM classifier was able to predict text engagement correctly. The relative rating of texts according to predicted engagement scores worked for 3 out of the 5 users (60%). The two users, for which sorting did not work, had the most bored and most engaged ratings correctly classified, but the remaining text ratings were falsely classified. Due to the simplification of reading engagement and the small sample size we refrain from claiming to have produced a reliable reading engagement predictor. However, the experiment served as a proof-of-concept for the framework we applied.

5.4 Chapter Summary

In this Chapter, we considered the user's cognitive state as an additional dimension of context. Being in a fatigued or in a highly focused mental state has an impact on how information can be received, processed, and retained. By being aware of users' current states technology can adapt information selection and presentation in ways to match users' processing capacities. Inferring cognitive states by using bio-signals is often invasive and cumbersome. In Chapter 3 we discussed utilizing people's near-constantly available mobile phones in order to detect moments of low and high attention. In this Chapter we focused on more long-term patterns of cognitive performance across the day, thereby laying the groundwork for **RQ4**: through alertness task tracking we were able to reconstruct people's circadian rhythm of alertness and cognitive performance, which generally describes the diurnal changes in the ability to concentrate. We argue that by providing technologies with an awareness of users' cognitive states applications can adapt to users' current cognitive capacities. Information interfaces, for example, could be adjusted according to the user's current cognitive capacity: in phases of high concentration, complex information can be efficiently displayed, whereas in phases of low concentration complexity can be removed to prevent information overload and frustration. Systems that are aware of the user's circadian rhythms can provide support in scheduling tasks across the day, manage interruptions, and help users become more self-aware of their diurnal rhythms. Since circadian disruptions have shown to be related to the emergence of schizophrenia and diabetes [49, 254], self-awareness through such systems could help people take preventive measures.

For supporting researchers and application builders to create cognition-aware systems we presented two tools: 1) a mobile toolkit comprising a sequence of tasks designed to elicit users' circadian rhythm of alertness and cognitive performance and 2) a conceptual framework comprising three steps for building algorithms that detect cognitive states from contextual sensor data. In the next part of this thesis, we present concepts and our applications of the presented tools and show how they can be used to facilitate effective information intake.

III

Applications

Chapter 6

Opportune Content Delivery

In Chapter 3 and 4 we laid the groundwork for detecting attention levels as an awareness dimension for context-aware systems. Technologies with the ability to assess different user states, such as being alert, receptive, tired, or bored, can use this information to facilitate information intake. While the user's capacity to focus influences the ability to process information effectively, current attention levels can be used to determine the timing of when to deliver particular types of information. In this Chapter, we present two research probes, in which we investigate content delivery in opportune moments as a direct application for systems with cognition-awareness.

We specifically focus on systems that proactively push content to the user. Based on the user's receptiveness to such content suggestions we gain insights into what makes moments opportune for delivering different types of content. We apply the prediction algorithm developed in Section 3.3 to establish links between different cognitive states and the user's openness to content suggestions.

While previous work has looked at making sense of users' mobile phone activities to predict human interruptibility [11, 102, 215], we focus on predicting people's general openness to different content suggestions, namely learning content in form of foreign language vocabulary and articles to engage in reading activities. Hence, we built a language learning app to trigger vocabulary reviews throughout the day. This investigation indentifies opportune moments for delivering learning content. We further report on a study conducted together with Pielot *et al.* [214], in which we applied our boredom prediction algorithm to suggest reading articles of general interest in situations of detected boredom.

In this Chapter, we address the following research question:

- **RQ5**: How can opportune moments for content delivery used to foster information intake and learning?

Parts of this chapter are based on the following (pending) publications:

- T. Dingler, D. Weber, M. Pielot, J. Cooper, C.-C. Chang, and N. Henze. 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, pages 28:1–28:12, New York, NY, USA, 2017. ACM
- M. Pielot, T. Dingler, J. S. Pedro, and N. Oliver. 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, pages 825–836, New York, NY, USA, 2015. ACM

6.1 Related Work

The work in this chapter is mainly inspired by and based on previous research in learning theories (see Section 2.2), technology-mediated learning, and work regarding user attention and interruptions.

The Effects of Repetition on Vocabulary Learning

Research has shown that the extent to which vocabulary gains are made is positively related to the number of times language learners encountered each vocabulary word [266]. Incidental vocabulary learning is a gradual process in which gains are made in small increments with repeated encounters needed to achieve full knowledge of a word [193]. In addition to repetition, word review can be enhanced with proper spacing. Spacing effects refer to the retention advantage for information that is repeated in a distributed fashion relative to information that is repeated in a massed fashion [39]. Dempster [62] demonstrated that language learners retained a greater number of vocabulary words when words presented were distributed and spaced apart instead of being presented in succession. Spacing effects not only apply within a learning session but also between sessions. Bahrick & Phelps [13] demonstrated that words were better retained when studied and re-learned at 30-day intervals when compared to 1-day or 0-day intervals.

The use of computers greatly enhances spaced repetition vocabulary learning. It is a lot easier for a computer to keep a record of individual learner's performance and control the sequencing of vocabulary words appropriately. Language learners who study with physical flashcards need to monitor their own learning progress, in turn running the risk of inefficient learning [194]. A study conducted by Nakata [194] has shown that students were able to recall more vocabulary words using the computer when compared to word lists.

Flashcard and Multiple Choice Learning

Unlike lesson-oriented learning, flashcards operate at the granularity of facts (*e.g.*, word translation) and measure the learner's ability to provide the correct response in the presence of a stimulus [92]. The use of flashcard is a common way to learn vocabulary as it allows the user to group difficult words together and allow them to be reviewed more frequently than easy words [194]. Furthermore, flashcards enhance vocabulary learning by reducing the list effect where learners tend to recall words more easily due to the order they appear on the word list.

Numerous researchers have explored microlearning applications based on flashcard presentation. Basoglu and Akdemir [18] had students use a mobile phone with a flashcard vocabulary application. Results indicated that using mobile phones as a vocabulary learning tool is more effective than one of the traditional vocabulary learning tools. Edge *et al.* [91] created a mobile flashcard application with an adaptive spaced repetition algorithm.

Multiple choice learning has not been extensively applied to mobile devices. Vocabulary skills, however, are commonly assessed through multiple choice tests [225]. Primarily, multiple choice quizzes highlight deficiencies in language acquisition, thereby providing a similar baseline as flashcard learning.

Notifications & Interruptions

Microlearning mobile applications have shown a lot of promise and potential to improve vocabulary learning, however, microlearning research studies typically are not optimized for sustained or continuous use. One way to promote sustained use of the app is through push notifications, however, the language learning app must also compete with numerous notifications pushed by other apps. There is research suggesting that frequent push notifications cause interruptions and can induce stress [281]. Increased number of notifications was also associated with increased negative emotions [212].

The stress and frustration induced by frequent and interruptive push notifications can be reduced. Studies have shown that pushing a notification during a breakpoint reduces frustration [141] and are dealt with significantly more quickly [99]. Strategies, such as *bounded deferral*, have been proposed to hold beck alerts when the user is predicted to be busy until more suitable moments [137]. Poppinga *et al.* [216] collected data from smartphone sensors and applied machine learning models to automatically predict and identify such opportune moments to trigger notifications.

These studies show that learning gains can be made with repeated exposure and learning sessions can be broken down into small chunks of quick interactions presented on a mobile device. Mobile applications geared toward learning, however, need to compete for the attention of the user who may have hundreds of other apps installed. This research investigates how to best design a mobile learning application as well as systematic push notification in order to promote prolonged and sustained mobile learning sessions.

6.2 Micro-Learning Sessions Throughout the Day

Learning new skills, knowledge, and practices are often limited by a lack of time, motivation, and resources. Quests, such as learning to program or a foreign language, are daunting tasks that require learners to dedicate large chunks of time on regular basis. According to The Foreign Service Institute (FSI) of the US Department of State[15], it can take up to 750 class hours (30 weeks) for a native

[15] `http://web.archive.org/web/20071014005901/http://www.nvtc.gov/lotw/months/november/learningExpectations.html`

English speaker to become proficient in German. Many language learners simply cannot fulfill such time commitments or lack the motivation to spend such a great deal of time and resources on intensive or immersive instructions.

While most products require learners to dedicate large chunks of time on a regular basis, mobile apps are often designed to support *microlearning* sessions. Microlearning sessions break down a learning task into a series of quick learning interactions, thereby reducing learning units to more manageable chunks that can be completed, for example, during idle moments, such as waiting for the bus or standing in line at the supermarket. Since most people keep their smartphone in reach most of the time, this opens up the possibility to engage in short learning tasks spaced out throughout the day as users are on the go. In addition to enabling people who cannot dedicate large chunks of time to learn a new language, research in psychology found that repetitions are more effective than dedicated long streaks of learning [54].

However, how to best design for such microlearning sessions is still an open question. Due to the anytime-anywhere capabilities of mobile devices, researchers commonly suggest that microlearning is a great tool to convert spare time into something productive [22, 32, 262]. Further, since some people may need nudges to keep learning a language, notifications can be used to turn users' attention to the device. Since notifications recently allowed for interactive content, they can also allow users to quickly deal with short tasks [14], such as reviewing a few words of foreign language vocabulary, hence engaging in microlearning sessions. Yet, it remains unclear whether mobile users prefer learning a new language during short transient moments like during their commute on a train. Also, it remains unclear how such notification-triggered sessions should be designed, specifically, when they should be triggered and how the interactive notifications should look like.

We report on our development of *QuickLearn*, an app to explore and compare different designs of microlearning sessions. Through interactive notifications, the app invites learners to frequently review language content. Word exercises can be attended to directly in the notification drawer or through self-initiated user sessions (see Figure 6.1). We assess the feasibility of notification triggers, presentation modes, and interaction modality with regard to microlearning sessions. In this chapter we present our findings from a mixed method user study: a controlled lab experiment and an in-the-wild study contributing the following:

1. Investigation of context factors of microlearning in mobile settings.

2. Assessment of the feasibility of using interactive notifications for learning sessions, which last shorter than app launches.

3. The comparison of two vocabulary review methods: flashcards vs. multiple choice, of which flashcards tend to be better suited for new word acquisitions.

4. Exploration of using idle moments as an opportunity for learning and memory strengthening.

6.2.1 System Design and Implementation

We developed a vocabulary trainer for Android devices. The application runs on Android phones with OS 4.3 or newer and provides two distinct modes: one allowed users to actively open the app and review vocabulary and the other consisted of a background service that initiated notifications to remind users to review vocabulary. On its first launch, the app shows a consent form explaining what kind of data will be collected and for which purpose. After explicitly stating their consent, users are guided through the permission granting process which consists of giving access to the Android Accessibility Services and allowing the app to access notifications. This is needed to inform the notification schedule algorithm and take into account the users' activities to trigger notifications. In the final step, the app asks the user to specify age, gender, and optionally leave an e-mail address. Users are also asked to indicate their mother tongue (L1), to chose the language they want to learn (L2), and what their current proficiency level in this language is. Since the application supports five languages, we asked users to choose the language they are most comfortable with in case their mother tongue is not available. Once the setup is completed, the background service is started and the app shows the first set of vocabulary.

User-Initiated Learning Sessions

The application can be opened at any time by selecting the application icon. Upon start, it launches an activity that shows words to review in sets of three. After each set, users are asked if they want to continue with the next set of words or to quit the application. The design intended to support the notion of short learning sessions to give users the feeling of completion and accomplishment in minimum time.

System-Initiated Learning Sessions

While the application can be explicitly launched by the user, an additional service running in the background that initiates learning sessions by triggering notifications. These notifications remind the user to review a set of vocabulary. The review can be done directly in the notification area, thereby avoiding the need to open the application and enforcing a context switch. By adding buttons to the notification, users are able to interact with the notification's content, which is why we consider them *interactive* notifications. The app can also be directly launched by clicking on the notification itself.

In Section 3.3 we identified moments, in which people tended to be bored and correlated them with phone usage. In these bored or stimuli-seeking moments, people may be more open to suggestions made by the phone and therefore more open to external interruptions. Since the success of micro-learning strongly depends on peoples' willingness to engage in learning tasks during opportune moments, boredom might be such an opportune state to study language vocabulary. To examine this hypothesis, we applied the model described in Chapter 3.3 and integrated the boredom classifier into the notification trigger algorithm. The model uses features, such as the recency of communications, demographics, or intensity of recent phone usage to estimate whether the phone user is boredom with an accuracy of 74.5% AUCROC. The notification trigger mechanism works as follows: whenever the user turns the phone's screen on, the app service tried to schedule a notification. Therefore, it checked whether a notification was already scheduled or whether the minimum time of 20 minutes since the last posted notification had elapsed. To not disturb users during the night, the notifications are only posted between 11 pm and 7 am. If these conditions were met, a timeout was scheduled with a random delay between 10 seconds and 5 minutes. When the timeout runs out, the classifier estimates the user's boredom state. In case the user was detected to be bored, the notification is posted. If the boredom state is predicted to be negative, the notification is only posted in 1 out of 9 cases. In pilot tests, we found that this makes roughly sure that an equal amount of notifications is posted for each of the two boredom states (bored and non-bored). In case the app is already in the foreground, no notification is posted.

User Interface Design

Words are presented in two display modalities: *flashcards* and *multiple choice*. The modality is randomly assigned when the app is first launched. The flashcard

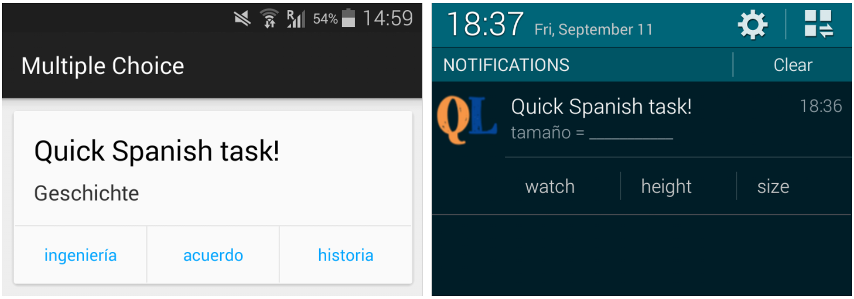


Figure 6.1: *QuickLearn* allows users to engage in microlearning sessions (left) or nudges the user to re-visit vocabulary through system-initiated, interactive notifications (right).

method consists of a series of screens, beginning with the presentation of a foreign noun with no translation (see Figure 6.2). Users can then click "translate" to view the translation and are then instructed to acknowledge whether they had already known the word ("knew it") or had had no previous knowledge of the word ("did not know"). This way, the words that are tagged as previously known by the user can be separated from the words that needed additional practice.

In the multiple choice modality, a foreign word was shown to the user while listing three suggestions for translation (see Figure 6.2): a random order of words including the correct translation as well as two decoy translations randomly retrieved from the vocabulary set. The three options are arranged in a random order. Users then have to choose which translation they believe to be correct. The application provides immediate feedback on whether the selection was correct or wrong. Correctly guessed words are tagged accordingly and have a lower frequency of repetition than those that were guessed wrong.

Word List

We used a vocabulary list consisting of common English nouns originally taken from the British National Corpus[16], similar to Cai *et al.* [38]. The list comprises high-frequency English nouns since in second language learning nouns are typically acquired before verbs and are less context-dependent. We used Google Translate to translate these words into Spanish, French, German, and Arabic. Fi-

[16] `http://www.natcorp.ox.ac.uk`

nally, two native speakers manually went through the list and corrected inaccurate translations and flagged highly ambiguous words. The final word list consists of 476 nouns in each language. The translations are uni-directional, meaning that a word was reviewed by being presented in its foreign translation (L2) before revealing it in the user's comfort language (L1).

Word Scheduling

Vocabulary words are delivered in sets of three, providing obvious stopping points should the user feel inclined to stop. After each set of words, users are asked if they wanted to continue or quit. A mechanism of spacing words for repetition was employed using a variation of the Leitner style schedule [111]: this schedule is based on the principle of spaced repetition. Given that humans exhibit a negatively exponential forgetting curve [90], repetitions need to occur at increasingly spaced intervals. Hence, new words are encountered just as they are about to be forgotten. So, whenever the user reviews a word, it is tagged as either correctly or incorrectly guessed. In the flashcard modality, this is done by admitting to having known the word or not. In the multiple choice modality, this is automatically done by evaluating the given answer. When the app is first launched, the word list is being randomized. With every word review, the app goes sequentially through that list. If a word is seen for the first time and guessed correctly it is appended to the very end of the word list. If this word is guessed incorrectly, it is replicated and spaced throughout the remaining wordlist in positions 4, 8, 16, 32, ... and so on until the end of the word list is reached. Whenever a word is encountered that is represented multiple times in the list, it is simply removed regardless of being guessed correctly or incorrectly, hence making sure that the list does not grow indefinitely.

6.2.2 Controlled User Study

To gain insights into the use of microlearning sessions for language acquisitions, we conducted a user study where we explicitly recruited participants to use the app over a course of two weeks.

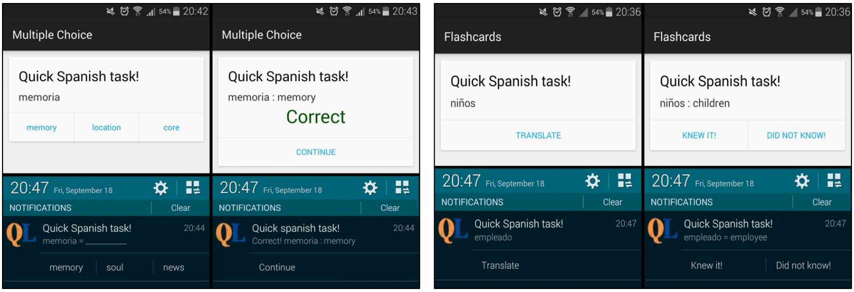


Figure 6.2: The multiple choice modality via app (top) and notification (bottom) showing three possible answers and correct response (left), and the flashcard modality via app and notification displaying initial screen and translation screen (right).

Method

We collected usage data from 17 participants who installed the app under our supervision and returned after each week to fill in a survey and being interviewed about their experiences. Regarding the vocab review methods, we used a repeated measure design with the review method being the independent variable. We counterbalanced the starting condition so that half of the participants started using the app with flashcards, and another half started out with using multiple choice answers. After seven days, the method changed automatically. As dependent variables, we collected app logs, general phone usage data as listed in Table 6.2, and information about the notifications triggered by the app. We assessed language learning effects by administering vocabulary tests after each week. Finally, participants were asked to provide subjective user feedback and comprehensive comments during the final semi-structured interview.

Participants

We initially recruited 19 participants (9 female) through university mailing lists and personal connections with ages ranging between 22 and 44 ($M = 28, SD = 5.2$) years. One participant needed to be excluded from the quantitative analysis due to technical reasons, and another because he had his phone stolen during the study. Most of the remaining 17 participants were students, of which 9 indicated

Language	None	Elementary	Limited	Professional	Full
German	1	2	2	2	0
English	0	0	1	0	0
Spanish	1	3	0	0	0
French	2	2	1	0	0

Table 6.1: Number of participants and their selection of their target language proficiency in the controlled user study.

German to be their first language, 1 French, 2 Arabic, 2 English, and 3 Spanish. Table 6.1 shows an overview of their target languages and their current proficiency levels. Three participants indicated to be currently enrolled in a second language course.

Data Collection

Once the setup was complete, the app started the notification schedule service as described above. It also included a log system where every app interaction was tracked: whenever the app was opened by the user, how it was launched (clicking the application icon, the notification, or direct interaction in the notification), which words were reviewed and with which review method. Further, we collected context parameters, such as time and location. Both sensor and usage data were sent to our servers through a secure connection. To prevent straining users' data plans, the app stored the collected data locally and transmitted only when connected through WiFi. As for whenever a notification was triggered, interacted with, dismissed or ignored, and whenever the app was explicitly launched, we logged the current boredom state as predicted by the classifier.

Procedure

We invited participants to our lab, where we walked them through the purpose of the study and had them fill in an initial survey collecting demographic data as well as their language preferences and previous knowledge. We explained the types of data that would be collected and how this data would be transmitted and stored. Then, we installed the app on their phone, went through the setup procedure and had them complete the first set of words. We explained about the

Phone Context	
ringer mode	mute, vibration, ringer
charging mode	unplugged, charging
battery status	related to usage intensity
display orientation	portrait or landscape mode, orientation changes
light sensor	changes in lightness allow us to derive whether phone is covered (carried in a pocket or taken out)
proximity sensor	phone in pocket
location	GPS data allows the inference of locations visited, and in-place vs. on-the-go states
motion	significant motion sensor, change of position
Phone Usage	
Calls	incoming, outgoing
SMS	incoming, outgoing
Notifications	received, dismissed, ignored, interacted with
Screen	on/off events
Unlocks	phone unlocks
Data usage	upload/download
Applications	applications in foreground, switches, usage duration

Table 6.2: Context data collected from phone sensors.

app's intention to post notifications and explicitly asked users to not feel obliged to answer every incoming app notification, but rather deal with them as found convenient. We counterbalanced the review mode to start with either the flashcard or multiple choice mode. This first setup meeting took about 15 minutes, after which participants used the app for seven consecutive days.

A week later we invited them back to take a vocabulary test and to fill in a survey. The test was dynamically created for each participant and contained only words that fulfilled the following condition: each word listed had been reviewed by the user and had been guessed incorrectly when first encountered. This way we tried to make sure that we tested only new words that were unknown to the user before using the app. The test would list the word in the second language (L2) and ask the participant to type in its translation (L1), thus employing a *recognition* task. We specifically did not list multiple choices or had participants self-assess their performance like in the flashcard review mode because we wanted to test whether the words had actually been memorized irrespective of the review mode. The survey consisted of the eleven 5-point Likert-style questions listed in Table 6.3 asking about the subjective experience with the notifications, the learning in general and its perceived effectiveness. The last part of the survey was a text-based feedback form asking about what participants liked, disliked, and would improve about the app. Depending on the number of words that had to be reviewed, filling in the survey took up to 15 minutes. The filling in of the survey was done under the supervision of an experimenter, in some cases where people could not come in after seven days, we sent them an online link to the survey with the explicit request not to use any auxiliary sources for filling in the vocabulary test. For the next seven days, the app changed its vocabulary review mode to flashcard or multiple choice respectively.

After another seven days of app, usage had passed, we asked participants to come in for a final wrap-up session where we applied another vocabulary test. Words presented in this test fulfilled the same condition described above and were words that had only been encountered in the second week of the study. This way we made sure that each vocabulary test contained words reviewed in one review mode respectively. Finally, participants attended a final semi-structured interview containing 16 questions.

Hence, in total participants came in three times. For the initial setup and the intermediate session after one week we compensated them with 5 EUR each, and another 10 EUR for completing the study after two weeks, accumulating in a total of 20 EUR for their entire participation in the study.

Results

In the following, we report the results with regard to the learning success, notification interaction, user ratings and qualitative feedback as given by participants.

Learning

After two weeks of casual usage, each participant had encountered on average 523 words ($Mdn = 331, SD = 477.14$). Of those, on average 223 ($Mdn = 196, SD = 127.2$) words were unique (not repeat-words) and 55 previously unknown ($Mdn = 39, SD = 53.91$). We categorized words as new or unknown when they had been marked as *"did not know"* or incorrectly answered when encountered for the first time. Participants completed on average 56 learning sessions ($Mdn = 45.5, SD = 31.26$). In total we recorded 557 learning sessions with an average of 10 words reviewed ($Mdn = 3, SD = 19.31$) in 3 word sets per session ($Mdn = 1, SD = 6.43$). A learning sessions took between 1second and 20minutes ($M = 48.8sec, Mdn = 19.2, SD = 1.62min$). In the vocabulary tests, users translated 35 previously unknown words (64%) correctly into the foreign language despite never having had to type an actual word during the study. These 35 words ($Mdn = 20, SD =$

26.3) account for about 18 new words learned per week. The user who was exposed to the most new words (76) translated 32 correctly to L2.

Learning through notifications vs. through the app

Over the course of two weeks, participants reviewed on average 451 words ($Mdn = 257, SD = 473,7$) through the app, and 72 words ($Mdn = 70, SD = 40.3$) directly through notifications. On average, 39 (SD=25.08) learning sessions were completed by actively launching the app and 21 (SD=19.5) sessions by interacting through notifications. A Wilcoxon signed-rank test showed that statistically significant more reviews took place through an explicit app launch than through notification interaction ($Z = -2.510, p = 0.012$). Similarly, significantly more sessions were completed within the app than through notifications ($Z = -2.476, p = 0.013$). Participants spent on average 23.7 seconds ($Mdn = 11.4, SD = 92.9$) per notification session vs. 59.8 seconds ($Mdn = 27.5, SD = 97.3$) per app session. A Wilcoxon signed-rank test showed that notification sessions lasted significantly shorter ($Z = -2.510, p = 0.12$). In notification sessions, participants reviewed between 2 and 15 words ($M = 3.5, Mdn = 3, SD = 1.6$) and in app sessions between 1 and 216 words ($M = 13.3, Mdn = 5, SD = 22.5$) per learning session. Significantly more words were reviewed per app session than per notification session ($Z = -2.746, p = 0.006$).

Learning through flashcards vs. multiple choice

Participants reviewed on average 271 words ($Mdn = 117, SD = 375.3$) with flashcards, and 252 words ($Mdn = 137.5, SD = 255.3$) directly through notifications. On average, 27.8 learning sessions ($Mdn = 22.5, SD = 20.7$) were completed with flashcards and 28.4 ($Mdn = 28.5, SD = 16.2$) sessions with multiple choice. A Wilcoxon signed-rank test showed no statistical significance between the number of words reviewed ($Z = -.784, p = 0.433$), nor between the number of learning sessions completed per condition ($Z = -.550, p = 0.582$). Participants spent on average 51 seconds ($Mdn = 17.8, SD = 115.3$) per flashcard session compared to 46 seconds ($Mdn = 20.5, SD = 75.6$) per multiple choice session. In flashcard sessions, participants reviewed between 1 and 216 words ($M = 10.3, Mdn = 3, SD = 21.4$) and in multiple choice sessions between 1 and 121 words ($M = 10.3, Mdn = 3, SD = 17$) per learning session. A Wilcoxon signed-rank test showed no statistical significance for the average session duration , nor for the average number of words reviewed per condition ($Z = -.863, p = 0.388$). For the flashcard condition participants recalled a median of 10 (3 to 13) previously unknown words, and a median of 5 (1 to 12) in the multiple choice condition. A Wilcoxon signed-rank test yielded no statistically significant differences between conditions, ($Z = -1.023, p = 0.306$), despite a tendency of participants to perform better recalling words learned through flashcards.

Learning Locations

Of in total 556 recorded learning sessions across two weeks, we registered 105 (18.9%) sessions having been completed at home, while 59 (10.6%) having been completed at work, 209 (37.6%) took place at other and 183 (32.9%) at unknown locations. Semantic locations were estimated based on where users spent most of their time during waking/sleep hours. Participants reviewed on average 257 ($SD = 605.5$) words at home, 57 ($SD = 101.2$) at work, and 219 ($SD = 242.5$) in transit (*i.e.* "other"). We applied a Friedman test with a post-hoc analysis using a Wilcoxon signed-rank test with a Bonferroni-corrected significance level set at $p < 0.017$. There was a statistically significant difference in where participants reviewed the most words ($\chi^2 = 11.375, p = 0.003$). The post-hoc analysis revealed a statistically significant difference between the number of words reviewed in transit compared to at work ($Z = -2.937, p = 0.003$), but no significant difference between work and home ($Z = -2.043, p = 0.041$) or home and transit ($Z = -1.448, p = 0.148$). However, there is a tendency towards more words reviewed at home ($MD = 48$) than at work ($MD = 23$) and more reviewed in transit ($MD = 134$) than at home or work. Users learned more words in transit, regardless of the review type (flashcards *vs.* multiple choice).

We further looked at the locations, in which more words were reviewed through notifications or in-app sessions: With regard to interactive notifications there was a statistically significant difference in locations ($\chi^2 = 9.475, p = 0.009$). The post-hoc analysis revealed a statistically significant difference between the number of words reviewed in transit compared to at work ($Z = -2.585, p = 0.01$), but no significant difference between work and home ($Z = -1.884, p = 0.06$) or home and transit ($Z = 1.139, p = 0.255$). There was a tendency towards more words reviewed through notifications at home ($MD = 5, SD = 35.8$) than at work ($MD = 1, SD = 13.6$) and more app reviews in transit ($MD = 18, SD = 32.1$) than at home or work. Similarly, there was a statistically significant difference in where participants used the app directly to review words ($\chi^2 = 11.460, p = 0.003$). The post-hoc analysis revealed a statistically significant difference between the number of words reviewed through the app in transit compared to at work ($Z = -2.844, p = 0.004$), but no significant difference between work and home ($Z = -1.578, p = 0.115$) or home and transit ($Z = -1.306, p = 0.191$). Here, there also was a tendency towards more words reviewed through in-app usage at home

($MD = 13, SD = 507.6$) than at work ($MD = 12, SD = 94$) and more app reviews in transit ($MD = 89, SD = 193$) than at home or work.

Boredom as Trigger

To explore whether people are more open to learning vocabulary when they are bored, we analyzed the data we collected in situations of presumed boredom and non-boredom. We consider the boredom estimation as the independent variable of a quasi-experimental design with two conditions: bored *vs.* normal. The dependent variables were (1) *click-through rate*: the fraction of times that people clicked on the notifications created by QuickLearn, and (2) *words per session*: the number of words learned in a session. First, we cleaned up the data by removing participants who received less than 20 notifications and/or who clicked on less than 5 of them. For the lab study, 15 of the 19 participants passed this filter.

Click-Through Rates In the data set of the lab study, we did not find differences between the fraction of notifications that the participants clicked in each condition. In the *bored* condition, scores ranged from 6% to 87% ($M = 31.2$, $SD = 22.4$). In the *normal* condition, scores ranged from 2% to 88% ($M = 35.2$, $SD = 25.4$). A Levene's test showed that variance of the scores of the two conditions was sufficiently equal to use parametric tests ($F(1,28) = .17, p = .68$). A dependent t-test revealed no significant effect ($t(14) = -.8, p = .43$).

Words Per Session We did not find differences between the average word that the participants learned per session. In the *bored* condition, scores ranged from 0 to 8 words ($M = 6.07$, $SD = 4.33$). In the *normal* condition, scores ranged from 3 to 74 words ($M = 12.87$, $SD = 17.89$). A Levene's test showed that variance of the scores of the two conditions was sufficiently equal to use parametric tests ($F(1,28) = 1.47, p = .20$). A dependent t-test revealed no significant effect ($t(14) = -1.69, p = .11$).

User Ratings

Table 6.3 contains the results from the intermediate and final survey taken after each week of app usage. Questions differed for each week except for the assessment of how participants rated the current learning mode. The ratings were given on a 5-point Likert-style scale with 0 indicating "I strongly disagree" and 4 "I strongly agree". Results show that the app was very well received, in particular, the notifications acting as reminders to review vocabulary ($Mdn = 3, SD = 0.97$) and the interaction possibility through the notification drawer ($Mdn = 3, SD = 1.22$). Fig. 6.3 shows box plots for a selection

Statement after 1st week	Median	SD
I feel the app helped me effectively improve my spanish vocabulary.	3	1
I liked that the app reminded me to review my vocabulary.	3	0.97
I liked learning words directly from the notification window.	3	1.22
I found the notifications well timed.	1	0.82
I found myself dismissing the notifications a lot.	2	1.01
The notifications interrupted my ongoing activities.	1	1.12
I liked being able to spend a few seconds on learning during the day.	4	0.42
I was able to review one set of words in a rather short amount of time.	3	0.73
The app allowed me to squeeze in reviewing vocabulary in between tasks.	3.5	0.84
I opened the app myself whenever I felt bored.	3	1.35
Statement after 2nd week	**Median**	**SD**
I think learning with MultipleChoice was very useful.	2	1.1
I think learning with FlashCards was very useful.	3	1.14
I preferred learning with MultipleChoice.	2	1.72
I preferred learning with FlashCards.	2.5	1.77
During this last week I noticed to get tired of the notifications.	2	1.2
During this last week I did not use the app as much.	2	1.29
I found myself learning with app at home a lot.	1.5	1.08
I found myself learning with app at work a lot.	1	1.4
I found myself learning with the app a lot when when I was on the go.	3	1.01
I found myself reviewing my vocabulary when other people were around.	1.5	0.99
I was very likely to review vocabulary when I found a notification was already there when I checked the phone.	3	1.17
I was very likely to review vocabulary when a notification was triggered while I was using the phone.	3	1.12
The app motivated me to use additional resources (e.g. books, courses, other apps,..) to improve my Spanish.	0.5	0.91

Table 6.3: Subjective user ratings in 5 point Likert-style scales: 0 = strongly disagree, 4 = strongly agree.

of questions participants were asked after the first week. *Effective* refers to whether participants were able to effectively improve their second language skills, *notification-bar* visualizes users' overall affirmation of liking to learn words directly through interactive notifications, thereby not being too *interruptive*. Participants strongly agreed with the statement that they enjoyed being able to spend a few seconds on *learning* during the day ($Mdn = 4, SD = 0.42$). Interaction time was perceived as very low as words could be reviewed in a rather short amount of time ($Mdn = 3, SD = 0.73$) and *in-between* tasks ($Mdn = 3.5, SD = 0.84$).

The novelty effect of introducing the app seems to be limited as participants overly disagreed with the statement that they got tired of notifications ($Mdn = 2, SD = 1.2$), neither did they stop using the app as much ($Mdn = 2, SD = 1.29$). Using the app on the go ($Mdn = 3, SD = 1.01$) seemed to be the preferred mode

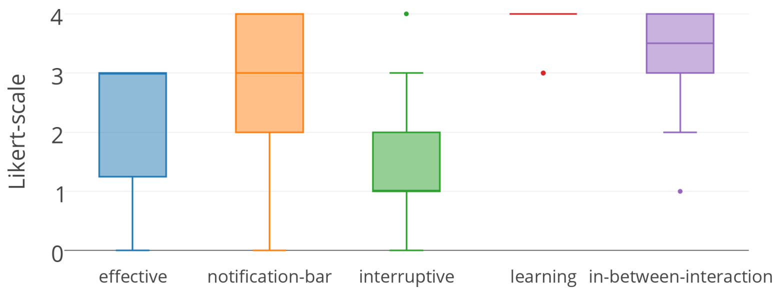


Figure 6.3: User ratings from 0 = strongly disagree to 4 = strongly agree taken from the first week's questionnaire, where we assessed learning *effectiveness*, utility of learning through the *notification bar*, the notifications' *interruptiveness*, *learning* in small chunks throughout the day, and being able to schedule sessions *in-between* tasks.

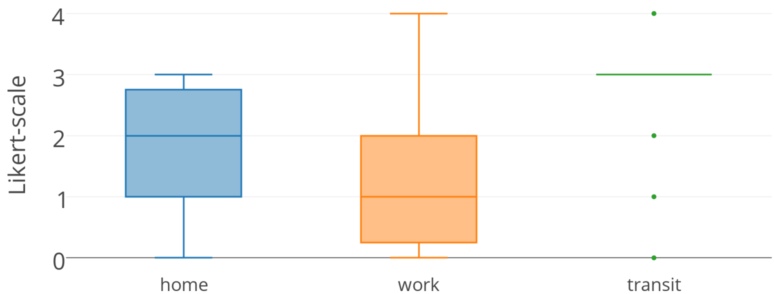


Figure 6.4: Results from subjective user assessment of *where* they found themselves engaging with microlearning sessions.

as compared to at home ($Mdn = 1.5, SD = 1.08$) or at work ($Mdn = 1, SD = 1.4$), as depicted in Fig. 6.4.

There was a statistically significant difference in where participants reported having used QuickLearn ($\chi^2 = 9.333, p = 0.009$). No statistically significant differences were detected between work and home ($Z = -0.677, p = 0.498$), transit and home ($Z = -1.840, p = 0.066$), or transit and work ($Z = -2.271, p = 0.023$). However, there was a tendency of more people indicating transit to be the preferred place where they found themselves reviewing vocabulary.

Qualitative Measures

After each week, participants filled in a survey, which—besides testing vocabulary recall and collecting the subjective user ratings—contained free-form questions, where participants could leave qualitative feedback.

Most participants (N=9) positively commented on the ease of word access and vocab interaction through the notification bar: P1 stated *"I didn't consciously realize that I was learning. Since it was such short periods of time."* P6 commented on the notifications' reminder function: *"I liked that it reminded me to invest some time. A reminder to do something useful."* Eight participants agreed on idle moments where notification interactions were welcome, *e.g.*, while *"fooling around on my phone."* (P1), *"after finishing a task on the phone"* (P7) or *"whenever I was alone or doing something boring."* (P8). Most participant (N=9) explicitly rejected the notion that notifications for microlearning were perceived as a disruption. *"It's not disruptive and allows me to continue easily with what I was doing"* (P14) while another observed *"you don't have to leave the current app, therefore you don't lose the focus"* (P7). Also, it was generally well received to combine pro-active reminders with the possibility to launch the app explicitly: *"I can open it whenever I want, but at the same time, it reminds me to keep learning"* (P16). Further, participants welcomed the brevity of the interactions: *"Takes pretty much no time to learn something new"* (P18). The short interactions seemed to convey a feeling of accomplishment: *"I was motivated to use the app because it was easy and not time-consuming"* (P13) and *"sets of three words give me a sense of achievement"* (P15). Participants found opportune moments for learning when they were, for example, *"pulling out the phone to just check the time and there was a notification there, I would do a couple of words"* (P1). Also here it became clearer that users welcomed notifications while engaged in information consumption rather than actively doing something on the phone: "when the browser was opened and I was reading unspecifically, then it would fit. It did not fit when I was concretely doing something, such as calendar entries" (P10). With regards to the question where participants were most likely to review vocabulary, a majority indicated public transportation and while in-transit to be most fitting, also waiting situations, such as in between breaks at the gym or at the bus stop.

Six participants found themselves explicitly opening the app in waiting situations: *"Mostly when I was on the train or when I was waiting. Wasting time."* (P5), and *"when I am on the lift alone. When I am waiting, queuing for some reason: supermarket, doctor's appointment [...]"* (P8). Five participants welcomed the reminders while they were using the phone (*"While reading emails I was also*

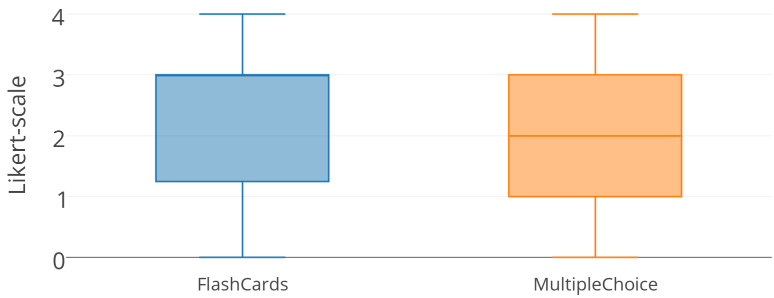


Figure 6.5: Participants' subjective rating of presentation modes: Flashcards vs. MultpleChoice on a 0 - 4 Likert-style scale.

doing some words on the side" (P6)) while others (N=4) preferred reminders to be triggered when they were not paying attention to their phone (*"When I was not using the phone, or right after I was using the phone"* (P10)).

Multiple participants (N=5) complained about the battery drain of the app. Since we collected a great variety of sensor data, this affected battery drainage in a noticeable way. Multiple choice seemed to polarize participants' opinions. Some liked its *"game-like character"* (P10), but a majority (N=9) saw problems with cognates, *i.e.* words that look similar in different languages, and the problem of being able to guess a lot of the words: *"My level of Spanish is not really high, but multiple choice often allows me to guess the answer to words I have never seen before, given my knowledge of French"* (P14) and *"Multiple choice options are certainly not optimal to learn vocabulary, though okay for the quick interactions"* (P16). However, as Fig. 6.5 shows there was barely a difference between learners preferring Flashcards over MultipleChoice. Some participants (N=4) commented negatively on the timing of the notifications, for example when they had just checked their phone only to be reminded to review their vocabulary after putting it away: *"sometimes notifications were delayed and came after I just checked the time"* (P17).

As for app improvements, six participants mentioned the limited choice of vocabulary and how they would like to add custom words or choose their own word topics. As for review modes, four participants requested text entry quizzes, speech output for pronunciation, sentence examples, or adding articles to nouns. One feature consistently requested was a way to track their learning progress.

6.2.3 In-the-wild Study

To increase our body of context sensor data we released *QuickLearn* free to download on Google Play. This part of the study comprises data from users in the wild. We collected sensor data as well as boredom prediction results and collected subjective app feedback through in-app surveys.

Participants

Besides releasing the app on Google Play, we launched a corresponding website and Facebook page to be able to promote the app through social channels. One year after the app release in September 2015, we registered 83 active users, 19 of which being lab-study participants, of which decided to keep using the app for a while beyond the study. While 28 indicated to be female, 55 reported to be male. The reported mean age was 31 ($SD = 11.7$) and according to the most frequent device locales (35 en-US, 9 de-DE, 18 en-GB, 27 other) and timezones, most users were accordingly from the U.S., Germany, and UK. While 20 selected German as their mother tongue, 39 chose English, 11 Arabic, 7 French, and 6 Spanish. As a target language, 25 chose to study English, 23 Spanish, 13 German, 18 French, and 4 Arabic.

Procedure

After downloading the app users went through the setup procedure consisting of agreeing to the terms on the consent form and granting the application access to the Android Accessibility Service as well as to notifications. Users were asked to provide their age, gender, language preference, and current proficiency. The display modality was randomly assigned when the app was first launched. In contrast to the controlled study, the modality changed every three days.

Results

In total, 14,632 notifications were triggered, of which 2765 (18.9%) were interacted with. Further, we registered 1435 app launches resulting in more than 19,000 words reviewed.

Boredom as Trigger

After cleaning up the data similar to the lab study we ended up with 16 participants who provided valid data sets.

Click-Through Rates In the data set of the in-the-wild study, we neither found differences between the click-through rates. In the *bored* condition, scores ranged from 2% to 81% ($M = 20.69$, $SD = 21.91$). In the *normal* condition, scores ranged from 0% to 50% ($M = 18.75$, $SD = 15.92$). A Levene's test showed that variance of the scores of the two conditions was sufficiently equal to use parametric tests ($F(1,30) = .33, p = .57$). A dependent t-test revealed no significant differences between the groups ($t(15) = .38, p = .71$).

Words Per Session In the data set of the in-the-wild study, we neither found differences between the click-through rates. In the *bored* condition, scores ranged from 1 to 49 words ($M = 7.71$, $SD = 11.49$). In the *normal* condition, scores ranged from 2 to 21 words ($M = 6.71$, $SD = 5.02$). A Levene's test showed that variance of the scores of the two conditions was sufficiently equal to use parametric tests ($F(1,32) = .55, p = .46$). A dependent t-test revealed no significant differences between the groups ($t(16) = .49, p = .63$). In summary, we found no evidence to indicate whether phases of boredom are better or worse for microlearning.

Opportune Moments for Language Learning

To be able to analyze context factors that are most opportune for language learning, the *QuickLearn* application collected usage logs as depicted in Table 6.2. From this log data, we elicited 36 features (listed in Table 6.4 and 6.5) from 14,632 instances of notifications sent to 37 different participants. The features were calculated according to a five minutes time window prior to each notification being triggered, *e.g.*, the number of phones unlocks five minutes before the notification was triggered. We had previously filtered out participants who received less than 20 notifications and interacted with less than five of those. By using machine-learning techniques we assessed the importance of these features to predict whether a participant would react to a notification by starting a quick learning session.

To rank the features, we built a model with XGBoost [44], a state-of-the-art gradient boosting regression tree algorithm which has performed exceptionally well in recent competitions[17]. XGBoost creates a ranking of the feature importance

[17] `https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions`

	Context	
Context	charging	Whether the phone is connected to a charger or not
Context	day_of_week	Day of the week (0-6)
Context	hour_of_day	Hour of the day (0-23)
Context	light	Light level in lux measured by the proximity sensor
Context	proximity	Flag whether screen is covered or not
Context	ringer_mode	Ringer mode (silent, vibrate, normal)
Context	semantic_location	Home, work, other, or unknown
Context	is_at_home	User classified as being home
Context	is_at_work	User classified as being at work
	Demographics	
Demographics	age	The participant's age in years
Demographics	gender	The participant's gender
	Last Communication Activity	
Last Activity	time_last_incoming_call_accepted	Time since last incoming phone call was accepted
Last Activity	time_last_incoming_call_denied	Time since last incoming phone call was rejected
Last Activity	time_last_notif	Time since last notification (excluding Borapp probe)
Last Activity	time_last_outgoing_call	Time since the user last made a phone call
Last Activity	time_last_SMS_received	Time since the last SMS was received
Last Activity	time_last_SMS_sent	Time since the last SMS was sent

Table 6.4: List of elicited features related to context, demographics, and time since last communication activity.

when building the model. The importance is defined by the fraction of times that a feature was chosen to be used in a tree. To build the model, we used XGBoost's standard configuration, with one exception: in the standard configuration, the classifier tended strongly towards a positive prediction, while only 18.9% of the notifications had led to learning sessions. Hence, we reduced *scale_pos_weight* by the factor 0.25, thereby increasing the penalty for false positives. The resulting model achieved a precision of 0.43, a recall of 0.712, and an F1-score of 0.526. Compared to the baseline, notifications posted by such a classifier would have resulted in a theoretical increase in the conversion rate by 2.21 times.

Table 6.6 shows the importance of the 10 best predictors as reported by the XGBoost model. We further analyzed the correlations between those ten features and the ground truth via Spearman's Rank correlation. Positive predictions of the classifier had significant, non-negligible correlations with:

- Less time passed since the phone had been last unlocked last. $(r = -0.303, p < 0.001)$

	Usage (related to usage intensity)	
Usage Intensity	battery_drain	Average battery drain in time window
Usage Intensity	time_in_comm_apps	Time spent in communication apps, categorized to none, micro session, and full session
	Usage (related to whether it was triggered externally)	
Usage External	num_notifs	Number of notifications received in time window
Usage External	last_notif	Name of the app that created the last notification
Usage External	last_notif_category	Category of the app that created the last notification
	Usage (related to the user being idling)	
Usage Idling	apps_per_min	Number of apps used in time-window divided by time the screen was on
Usage Idling	num_apps	Number of apps launched in time window prior to probe
Usage Idling	num_unlock	Number of phone unlocks in time window prior to probe
Usage Idling	time_last_notif_access	Time since the user last opened the notification center
Usage Idling	time_last_unlock	Time since the user last unlocked the phone
Usage Idline	num_screen_on	Number of screen on events in time window
	Usage (related to the type of usage)	
Usage Type	screen_orient_changes	Flag whether there have been screen orientation changes in the time window
Usage Type	app_category_in_focus	Category of the app in focus prior to the probe
Usage Type	app_in_focus	App that was in focus prior to the probe
Usage Type	comm_notifs_in_tw	received in the time window prior to the probe
Usage Type	app_category_in_focus_comm	Current app in focus is communication app
Usage Type	most_used_app	Name of the app used most in the time window
Usage Type	most_used_app_category	Category of the app used most in the time window
Usage Type	prev_app_in_focus	App in focus prior to *app_in_focus*

Table 6.5: List of elicited features related to usage intensity, external triggers, idling and type.

- Higher age.
  ($r = 0.144, p < 0.001$)
- More apps used during the last 5 minutes.
  ($r = 0.213, p < 0.001$)
- Less time passed since the last access of the notification center.
  ($r = -0.186, p < 0.001$)
- More screen unlocks during the last 5 minutes.
  ($r = 0.264, p < 0.001$)
- Not having received a notification in the last 5 minutes.
  ($r = 0.109, p < 0.001$)

Feature	Importance
time_last_unlock	0.2163
age	0.1450
time_last_incoming_phonecall_denied	0.0814
time_last_notif	0.0687
num_apps	0.0662
time_last_notification_center_access	0.0662
num_unlock	0.0433
last_notif_category#None	0.0407
time_last_incoming_phonecall_accepted	0.0407
time_last_outgoing_phonecall	0.0382

Table 6.6: Feature importance as reported by the XGBoost model.

Primarily, these results indicate that participants were more likely to engage in quick learning sessions when they had interacted with the phone more recently. This finding is in line with previous work that showed that more intense phone use correlates with boredom and stimulation-seeking (Section 3.3), *ritualistic* phone use [132].

6.2.4 Discussion and Limitations

Combining the two studies presented we provide empirical evidence that using a proactive mobile learning app facilitates microlearning and is an effective way of spacing out learning sessions throughout the day. This is mainly due to users' tendency to review vocabulary on the go. We were able to shed some light on context factors, in which microlearning was found feasible, namely when in transit as well as during idle moments, such as waiting situations. The design of the app has supported this notion of learning on-the-go by providing interactive notifications. Further, we analyzed context factors linking phone usage patterns, such as phone unlocks, app usage, notification access, and screen on/off events to opportune moments for learning. Learning sessions through notifications were found to be shorter than explicit app launches. Therefore, notifications offer great possibilities for engaging users in quick learning tasks that only take up a few seconds. Also, despite a tendency towards flashcards being more effective (as in new words learned and remembered) than multiple choice questions, a statistically significant difference could not be determined. In a mobile setting and with regard to users often finding themselves killing time with the phone positive aspects of MultipleChoice seem to be the game-like mechanism and the pleasure of guessing as stated by study participants. When notifications were triggered out of predicted boredom situations, we were not able to detect a statistically significant difference when compared to non-bored situations. Therefore, the study did not find conclusive evidence whether people are more likely to engage in micro-learning tasks depending on whether they are bored or not. While boredom is characterized as a state in which people seek stimulation—hence might be open for the suggestion to learn vocabulary—it is also characterized as the inability to find stimulation in current activities [89]. While people who are currently experiencing boredom might, in theory, have the time to engage in microlearning, they may not be able to get up the energy or may not expect the stimulation they desire from this mentally demanding task. It may be worthwhile, however, to investigate users' receptiveness of other content types suggested in moments of detected boredom.

By analyzing phone usage context we provide an assessment of opportune moments for scheduling learning sessions. Proactive learning reminders were especially welcomed when the phone had been recently used as the correlations with phone unlock and app usage patterns show. Also, the higher participants' age, the more likely they would engage with learning content. Unfortunately, the nature of the in-the-wild-study and the limited knowledge of our users prevents us from drawing clear causal relationships here. However, by updating the boredom classifier with our innate model we could boost acceptance rates of the app's learning reminders by more than factor 2.

For a one-year in-the-wild study we registered relatively low usage numbers. Besides the fact that we did not actively push the app's popularity, its data settings and therefore privacy implications might have had a deterrent effect. Obviously, our apparatus does not suffice to teach a foreign language to a learner. We rather focused on foreign language vocabulary as a proof-of-concept application. However, due to the well-received notion of microlearning sessions throughout the day, such technology could pose as a complement to existing learning techniques. For example, combined with an e-Reader or audiobook application, new words from articles the user has read could automatically be placed into the learning cue and spaced out in systematic repetitions in order to help users eventually commit them to long-term memory.

6.2.5 Study Conclusion

Microlearning in combination with ubiquitous technologies bears great potential for learners who naturally lack time and motivation to tackle a daunting task, such as learning a foreign language. In our work, we set out to explore different aspects of microlearning as a tool on-the-go, namely presentation mode, automatic reminders, interaction modalities, and learning context. Therefore, we created an Android app, with which users learned on average 18 new words per week by being exposed to about 37 words per day: whether in relatively quick word reviews through the notification bar or through longer sessions through explicit app launches. Our results show that people were more open to engaging in quick learning sessions when they are on-the-go. Both notification interactions and in-app sessions registered more word reviews when people were in transit compared to at home or at work. We see microlearning as a complement to existing learning techniques: throughout the day we read various articles, make notes, and try to remember new information encountered. As part of a greater knowledge management system, small learning units can be provided to users in idle moments. In this work, we showed the feasibility of such an approach, but the possibilities are far greater when, for example, external resources are connected. Collaborative platforms, such as Anki[18], for example, offer a great source for flashcards that can be fed into a microlearning system, such as the one proposed here. Through intelligent scheduling of repetitions and smarter triggers for reminders, such technology can help users make use of idle moments, commit knowledge to long-term memory and therefore increase their personal effectiveness.

While predicted states boredom did not seem to correlate with opportune moments for learning we showed that training a classifier with context information retrieved from the phone's sensors can more than double compliance rates. As discussed, learning activities may be too mentally demanding as to be a preferred activity in states of boredom. To validate and apply the boredom detection algorithm developed in Section 3.3, together with Pielot *et al.* [214], we conducted a study on triggering suggestions of light reading material based on predicted states of boredom.

[18] `http://ankisrs.net/`

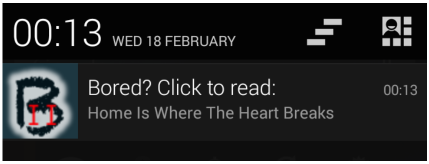


Figure 6.6: The app Borapp2 triggered notifications throughout the day to invite users to read BuzzFeed articles in bored and non-bored states.

6.3 Using Predicted Boredom to Suggest Reading Content

Here, we report on a study we conducted together with Pielot *et al.* [214] in 2015, which followed the initial user study, in which we developed the boredom prediction algorithm described in Section 3.3. The main goal of this study was to assess whether users were more receptive to suggested reading content when predicted to be bored. In contrast to the *QuickLearn* study, in which we suggested learning content, light-weight reading articles were randomly selected from the news app *Buzzfeed*[19]. Therefore, a new version of *Borapp* called *Borapp2* was released on GooglePlay, which suggested reading articles throughout the day.

For this study, 16 users living in Central Europe between 16 and 51 years old ($Mdn = 39$, $M = 36.31$, $SD = 9.37$) were explicitly recruited to install and use the app over the course of at least two weeks. The app was equipped with the boredom detection algorithm, which analyzed the features described in Table 3.1 and 3.2 locally on the mobile device in order to classify users to be bored or non-bored via a Random Forest classifier. Throughout the day the app triggered notifications, which suggested users to open the Buzzfeed news app. The notifications contained the title of the most recent Buzzfeed article and prompted the user to click in order to read the article (see Figure 6.6). Notifications were scheduled with a delay of up to five minutes when the phone screen was turned on in case 30 minutes had passed since the last notification. Before triggering the notification, the app checked whether the user was currently classified as bored or not. To provide a balanced dataset the app made sure to trigger a similar amount of notifications for each boredom condition. The app removed notifications that were ignored by the user for more than five minutes.

[19] `www.buzzfeed.com`

Using the participants' receptiveness with regard to the notifications triggered as the dependent measure, the experiment yielded two measurements: 1) the *click-ratio*, which described the number of notifications clicked in a condition divided by the total number of notifications presented in this condition. And 2) the *engagement-ratio*, which additionally described how long users spent reading the suggested article after opening it. Users were regarded as being engaged when they spent at least 30 seconds with the suggested content.

Over the course of twelve days (the first two days were excluded from the data analysis to exclude a novelty bias) participants had received 941 content suggestions in form of notifications ($M = 60.81$, $SD = 38.27$). Participants were predicted to be bored in 48.0% of these cases. We found that users were more likely to click on suggested content ($Mdn = 20.5\%$ *vs.* $Mdn = 8\%$) and more likely to spend more than 30 seconds engaging with it ($Mdn = 15\%$ *vs.* $Mdn = 4\%$) when they were predicted to be bored. Differences in bored *vs.* non-bored states for both measures in terms of click-ratio ($z = -2.102, p = .018$, $r = -.543$) and engagement-ratio ($z = -2.102, p = .018$, $r = -.511$) were statistically significant with large effect sizes. These results show the feasibility of delivering reading content in opportune moments, namely when predicted to be bored.

6.4 Chapter Summary

In this Chapter, we presented applications that take into account the current user state and context to deliver content suggestions throughout the day. While we built upon the algorithms and methods developed in Chapter 3 and 4 we further explored characteristics of opportune moments for delivering learning and reading content. For the first field study reported in this Chapter, we built a foreign language vocabulary trainer that triggered learning reminders throughout the day. To analyze opportune moments for delivering learning content we collected app usage data along with context information from the phones' sensor data. Further, we collected subjective user feedback through surveys and interviews. Results show that moments, which users spend in transit or idle (*e.g.*, in waiting situations) present opportunities for quick learning sessions, so-called microlearning sessions. Because of the ubiquity of mobile devices such sessions can be triggered throughout the day, allowing users to fill idle times and memorizing information through spaced-out repetition. Analyzing sensor context factors we found that phone usage patterns such as phone unlocks, app usage, notification access, and screen on/off events can be used to predict such opportune moments for learning. Our classifier reached a recall rate of 71.2%.

While we were interested in investigating the link between moments of predicted boredom and a willingness to engage with learning content, we were not able to find any reliable correlation between the two. However, when training a machine-learning algorithm with the context data collected during the study we would be—in theory—able to more than double the compliance rate. This lets us assume that moments of boredom might be different from moments, in which learning activities were appreciated by users.

We further looked at the design of learning triggers by comparing flashcards to multiple choice learning. For both modes, we used interactive notifications which could be handled in the notification drawer rather than requiring the user to explicitly launch the app. This allowed users to engage in brief microinteractions throughout the day, which study participants appreciated. The ease of access to learning content allowed participants to briefly engage with learning tasks in between other activities.

We applied the framework presented in Section 5.3 as follows: 1) we derived our ground truth by observation, *i.e.*, from log data. Therefore, we used instances, in which users either proactively opened the app or clicked the notifications as a willingness to engage in learning activities. Instances, in which notifications were ignored or dismissed we considered as unwillingness to learn. 2) We elicited a set of 36 features in order to 3) train a machine-learning model for predicting opportune moments for delivering learning content.

We also investigated the suggestion of reading content. Therefore, we integrated our boredom prediction algorithm into a mobile app that suggested articles from a popular news app throughout the day. A field study we conducted showed that users were more likely to engage with content suggestions and spent more time reading when predicted to be bored.

These two studies show the overall feasibility of delivering content to users in opportune moments, *i.e.*, moments in which users are open to follow these suggestions and engage in activities related to information intake and learning. With regard to **RQ5** we investigated how such moments could be determined and following our proposed framework for developing cognition-aware systems, we were able to validate our approach on an application level. By triggering learning reminders we analyzed users' receptiveness to content suggestions in order to predict user engagement based on phone usage data. Similar studies could explore the nature of different content types and their respective user context: besides learning foreign vocabulary and reading news articles, application categories, such as games and entertainment apps could be explored. Lee *et al.* [173] recently proposed a context-aware application scheduler, which preloads background applications as they are predicted to be relevant to the user at the time. While proactive recommenders, in general, run the risk of patronizing users, predictive systems can be used to engage and re-engage users with services, contents, and apps. However, transparency and appropriate intervention mechanism should be considered to allow users to control such proactive services.

Detecting opportune moments for content delivery can further be used to instill new habits: through context analysis (*e.g.*, daily wait-time at the bus stop) applications could learn when to successfully trigger certain reminders and hence use such moments for systematically pushing content related to a certain topic or related to a certain skill to the user. Such an approach could lead to instilling new habits in people since habits are triggered by context and therefore become deeply ingrained [279]. An example would be a daily foreign vocabulary repetition taking place every morning at the bus stop. The advantage of habits is that even if the technology seizes to remind its user or is simply not available at the time, the desired behavior could still be successfully triggered.

Chapter 7

Adaptive Reading Interfaces

Reading is an activity mostly taken up for information gain and pleasure. With the rise of the information age, we face a great amount of information on a daily basis. With the introduction of multimedia, information is tailored more broadly to our various senses. However, reading remains the primary channel to consume information. The skill to read and absorb information efficiently has become vital in both the private and professional sector. We specifically focus in this Chapter on reading because of its function in information consumption. The effectiveness of committing information from its perception to memory is highly impacted by our current ability to concentrate. Attention levels are therefore a vital factor for reading activities. While Chapter 6 focused on detecting opportune moments for content delivery, in this Chapter we explore how reading content can be adjusted to support users taking in that information effectively.

To cope with the ever-growing amount of textual information to consume, different techniques have been proposed to increase reading efficiency. Reiner *et al.* [224] recently summarized these efforts concluding with a clear assessment that language skill is at the core of reading speed. In order to increase speed while maintaining high comprehension, readers need to become more skilled language users. Advanced language skills imply increased vocabulary, which allows readers to quickly grasp the gist of a text while being able to comprehend more of the arguments presented. There is a general trade-off between speed and accuracy. If a thorough understanding of a text is not the goal, then deliberately increasing reading speed allows the reader to get through text faster at a moderate comprehension level.

However, comprehension levels also depend on the reader's mental state and ability to focus. Increasing reading speed while trying to maintain high comprehension levels increases cognitive strain. High focus generally supports our information bandwidth, therefore we can process information more effectively in states of high concentration. In contrast, when cognitive capacities run low, more text pushed at us will result in more information remaining unprocessed. This absent-mindedness during information intake leads to information not committed to memory and therefore not retrievable [234].

Currently, people have little awareness and even less control over how to adjust their reading flow. In this Chapter, we investigate how reading UIs can be adjusted in real-time to match the user's cognitive capacities. First, we look at common methods to dynamically display text in order to modulate reading speed. Based on a review of previous work in the field we propose a dynamic reading UI that uses a kinetic stimulus to guide readers' eye movements and speeds up their reading. We evaluate its effectiveness compared to alternative methods with regard to text comprehension and users' mental load. Then, we focus on reading in ubiquitous device environments, where text is increasingly read on wearable devices, such as smartwatches and smart eyewear. With their limited screen sizes, such devices pose a challenge to effective information consumption. Hence, we applied dynamic reading UIs, such as Rapid Serial Visual Presentation (RSVP), to allow information intake on-the-go. User control over reading speed is vital, which is why we explored different interaction concepts for manipulating reading flow. Hence, we investigated both implicit and explicit interaction techniques for reading control.

The research probes presented in this Chapter are examples of how systems can adjust the information bandwidth according to their users' cognitive capacities. This can either be done by providing user controls through explicit interaction techniques or by using implicit bio-signals, such as eye movement data, to adjust information interfaces. Cognition-aware systems can help match the attentional requirements of a reading interface with the user's cognitive state in order to prevent frustration in case of highly complex contents or to avoid boredom in case of simple materials. In this chapter we, therefore, investigate adaptive reading interfaces that facilitate readers' information intake, addressing the following research question:

- **RQ6**: How can information displays and more specifically reading UIs be adjusted in real-time to decrease or increase information bandwidth?

Parts of this chapter are based on the following publications:

- T. Dingler, A. S. Shirazi, K. Kunze, and A. Schmidt. Assessment of stimuli for supporting speed reading on electronic devices. In *Proceedings of the 6th Augmented Human International Conference*, AH '15, pages 117–124, New York, NY, USA, 2015. ACM
- T. Dingler, R. Rzayev, V. Schwind, and N. Henze. Rsvp on the go - implicit reading support on smart watches through eye tracking. In ACM, editor, *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '15, New York, NY, USA, 2016. ACM

7.1 Related Work

Our investigation is based on previous work on reading interfaces, interaction techniques, eye tracking and further inspired by speed reading approaches as they are applied by speed readers and commercially taught.

The Reading Process

Reading is a psychomotor activity during which our eyes move across visually displayed text. During those movements—called *saccades*—vision is mainly suppressed [186]. Hence, information is only acquired during the period of time when the eyes remain fairly still—called *fixations*. Although there are exceptions to this where information can be acquired during these eye movements [255], in most cases our eyes move too quickly across a visual stimulus so that only a blur would be perceived. An average fixation in reading lasts for about 225-250ms [223], but depending on factors such as reading skill, language familiarity, and text complexity fixation duration can vary considerably. The average saccade length is 7-9 letters for readers of English and similar writing systems, which also can significantly differ. Another important component of reading is so-called *regressions*, *i.e.*, saccades in the opposite of the reading direction, which occur about 10-15% of the reading time. Regressions are important indicators of text comprehension (*e.g.*, many regressions occur in particularly difficult texts) and differ from *return sweeps*, which describe larger saccades from right-to-left in order to get from the end of one line to the beginning of the next. The tracking of eye movements can give us insights into people's reading skills and also text difficulty: more difficult text generally leads to longer fixations, shorter saccades, and more regressions. Frequent readers typically read at a rate of 200-400 words per minutes (wpm), but skilled readers are reported to reach much higher reading speeds. These speeds highly depend on a number of factors, such as the reader's language proficiency and level of concentration, but also on the text's complexity and typographical aspects.

Cognitive Effects of Reading

Stanovich and Cunningham [250] identified the *"Matthew effect"* with regard to reading, which describes the concept of a rich-get-richer and poor-get-poorer phenomenon. Hence, poor readers tend to expose themselves to less text than their more skilled peers, thereby increasingly corrupting their reading skill level. They showed that much of what we read directly influences our language skills and the size of our vocabulary. This is due to the fact that written language is lexically much richer than spoken language. So sheer reading volume comes with a higher exposure to rich language and hence enriches readers' language skills. Boosted language skills again contribute to the development of higher cognitive functions like reasoning and judgment [205] and also lead to a greater general knowledge about the world. Pronin *et al.* [218] linked the acceleration of thought processes in a series of experiments to joy-enhancing effects, rapid reading being one of those methods used for accelerating thought. Hence, we started looking into ways to help people read faster and more.

Increasing Reading Efficiency (Speed Reading)

Evelyn Wood, creator of the Evelyn Wood Method [104] and one of the pioneers of speed reading, was supposedly capable of reading 6000 words a minute. Techniques she developed and applied were:

- Reading groups of words rather than single words, therefore, she needed to train her peripheral perception.
- Avoiding involuntary rereading of passages.

- Use of a finger or pointer to trace lines of text while eliminating sub-vocalization (*i.e.*, reading out loud in reader's head).

Wood noticed that the sweeping motion of her hand across a page caught her eyes' attention and helped them move more smoothly. Hence, the hand or finger could be used as a pacer. This insight inspired us to design an equivalent for electronic devices in form of a kinetic stimulus.

Hansen [121] reports on a series of studies on reading comprehension with rapid readers trained in the Evelyn Wood method. Her results showed that rapid readers were superior in comprehension of relational aspects of text and were able to recall significantly more information than normal readers due to the fact that they were able to read the material more than once given a time constraint. In other words, when normal readers and speed readers were given the same amount of time to read a text, the speed readers covered more of the material and their recall indicated comprehension of the gist of the passage. They especially tended to recall more idea clusters than normal readers, but less detail about each idea. Other studies on the Wood method reported that average comprehension levels went down as reading rate increased [40, 93, 116]. Just and Carpenter [150], on the other hand, show that training speed reading can increase the comprehension level of higher level information even on faster reading speeds. Yet, the increases also depend on text types and difficulty. They show that easy texts can be read very fast without loss in comprehension. Together with Rayner *et al.* [224] they give a comprehensive overview of the processes of speed reading. Both come to the conclusion that there is a trade-off between reading speed and text comprehension. In order to maintain high comprehension while reading faster, readers need to practice reading and therefore become more skilled language users. However, Rayner *et al.* also point out that speed reading or skimming text is feasible if a thorough text understanding is not the goal. Hence, moderate comprehension may be acceptable for being able to read through certain texts more quickly.

A series of studies have focused on how to best display text on computer screens and mobile devices: while limited screen size poses a challenge for reading UIs, reading performance has been shown to generally increase with bigger font size, which also affects the readers' subjective preference and lowers levels of perceived difficulty [15, 58]. Reading while performing a secondary task, such as walking, has further shown to decrease reading performance while increasing cognitive load [236, 256]. Also, there is a variety of commercial tools available that facilitate reading on the web: *Readability*[20], for example, used as a browser plugin cleans up web pages and displays content in a more readable manner. Another example is the *BeeLine Reader*[21], which aims at smoother line break transitions by guiding the user's eye through a color code from one line to the next. While these tools may facilitate reading on electronic screens they do not necessarily nudge readers to increase their reading speed.

Rapid Serial Visual Presentation (RSVP)

RSVP, a term coined by Forster [103], is an experimental model for examining temporal characteristics of attention. With this method users focus on visual items being continuously presented in the same place. High information transfer rates are thus possible because the need for saccadic eye movements is eliminated.

For electronic devices RSVP allows space to be traded for time and hence can be used to support information browsing and search tasks on small displays [60]. However, the capability of the human visual system seems to be the limiting factor in the application of RSVP. Presentation rate and the visual similarity of stimuli have been shown to influence the effectiveness of RSVP streams [217]. Subsequent targets are often missed especially when they occur in rapid succession (180-450ms), which Raymond *et al.* [222] described as *attentional blink*. When users are engaged in a secondary task, more stimuli are potentially missed. This is specifically critical for reading activities, where missing parts of a sentence may severely inhibit readers' text comprehension.

Masson [183] reported reading studies using RSVP, in which participants were often able to correctly outline the essence of a passage without necessarily recalling specific words. Schotter *et al.* [241] on the other hand presented findings of how repressing regressions in reading, which RSVP effectively does, hinders text comprehension, especially when dealing with ambiguous sentence structures. Hedin and Lindgren [128] examined reading on mobile devices using RSVP in regard to reading comprehension and efficiency. In a user study, they compared reading with RSVP vs. reading with scrolling using different reading speeds. They found that with RSVP speed and comprehension is high, but that users are generally uncomfortable with the technique.

[20] `http://www.readability.com/`

[21] `http://www.beelinereader.com/`

There are different modes of RSVP as described by Spence [248], however, for reading activities we focus on the sequential presentation of words in one spot, as it was used more recently by the commercial application *spritz*[22] for a text presentation technique on mobile phones and smart watches. Georgiev [110] investigated reading speeds—including via RSVP—on mobile devices and compared reading on a PC screen with reading on paper. Top reading speeds were achieved on computer screens with a font size of 14pt and on paper. In our work, we apply these findings and focus on the effects of RSVP with the facilitation of speed reading in mind.

Gaze-based Reading Interaction

First work on gaze-based interaction focused on the user-computer dialogue in a natural and unobtrusive way [146]. Kern *et al.* [153] investigated the feasibility of using eye tracking to facilitate the resumption of an interrupted task: they developed a system that provided visual placeholders to highlight the last gaze position which allowed users to efficiently switch between tasks. Hansen *et al.* [122] added gaze tracking to smartwatches to allow hands-free interaction through gaze gestures. Dickie *et al.* [66] introduced a platform for sensing eye contact on mobile screens based on an infrared camera system. They further discussed a reading application using RSVP and controlling the reading flow through eye gaze. However, the application was neither implemented nor tested.

Concluding, various reading methods have been proposed to increase reading speed or to accommodate for certain devices' characteristics. While there is a trade-off between reading speed and text comprehension, not every text necessarily needs to be processed in full detail depending on the readers' goals. Also, reading performance depends on the reader's current ability to concentrate: in phases of high cognitive performance, text can be read faster and more effectively processed than in times of low attention levels. Reading with high speed and concentration, therefore, affects the reader's mental load. Awareness of readers' attention and cognitive performance levels can inform reading UIs to dynamically adjust text display and keep the mental load in balance. Hence, we explored the adjustment of reading UIs by dynamically increasing users' reading speed and measuring the effects on cognitive load. We begin with the report of a study to compare dynamic speed reading UIs.

---

[22] `http://www.spritzinc.com/`

7.2 Reading Interfaces to Increase Reading Speed

Naturally, over the course of time people develop their innate reading, skimming and skipping strategies. Speed reading techniques are a much-discussed topic that has gained many followers over the last decades. Studies conducted are often disputed but agree on a natural trade-off between reading speed and comprehension levels. Various techniques are taught by books and seminars that allegedly enable speed reading, such as by Frank [104]. RSVP has been proposed as a reading technique to push a reader through a text by displaying single or groups of words sequentially in one focal point. Recent Web apps have spurred excitement around the prospect of achieving higher speed reading by effectively reducing eye movements (saccades). Other common techniques include the use of a kinetic stimulus (such as a moving hand, pen or finger) to guide a reader consistently across lines of text.

To investigate whether such techniques can be applied to reading on electronic devices and to explore their feasibility, we assessed two stimuli: 1) the RSVP method by using the open source framework *Squirt*[23] and 2) a kinetic stimulus in form of a dynamic text underline effect. Here, we present two consecutive user studies: first, we report on a lab study in which we collected data about eye movements, mental load and comprehension levels of 24 participants and second, we collected subjective feedback data from 108 participants in an online reading experiment.

In this Section, we 1) describe the application of speed reading techniques to reading on electronic devices, 2) introduce a kinetic stimulus in order to actively increase users' reading speed, and 3) investigate the effects of RSVP and the kinetic stimulus on text comprehension, mental load, eye movements and subjective perception.

7.2.1 Stimuli to Support Speed Reading

We implemented two stimuli with the goal of increasing reading speed by guiding the user's eye. We built our prototypes with a combination of HTML5, CSS and JavaScript to make our system not only run in the lab but also on the web to make it accessible to a broad pool of study participants.

[23] `http://squirt.io/`

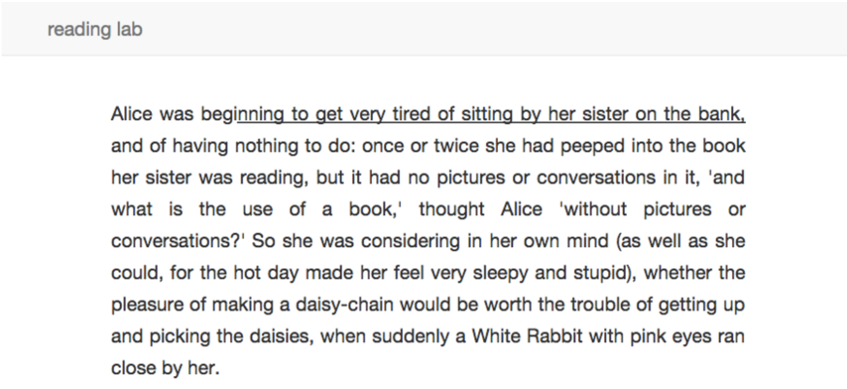


Figure 7.1: Kinetic stimulus: A moving line as underline effect.

The first stimulus is modeled after the idea of swiping the user's finger or a pen across a text in order to keep a constant reading flow [104]. Initially, we had several prototypes of stimuli suggesting a constant motion across the screen. By conducting pilot studies and questioning independent researchers we decided to use the technique of dynamically underlining lines of text as a kinetic stimulus for the user study. Using the HTML5 *canvas* element we implemented this dynamic line placed under a line of text that moves from left to right at a predetermined speed (Figure 7.1). Hereby not the entire line moves, but the left beginning of the line moves to the right end only underlining words that are still up for reading. This is intended to keep the eyes focused on the current line and avoid jumps of the eye between lines as well as regressions. Once the stimulus reaches the end of one line, the next line is underlined and the stimulus starts to run again. The entire text is visible at all times and the stimulus moves from left to right and line to line.

The second stimulus is modeled after the RSVP method where one word is displayed at a time at a focal point. We based our implementation on the javascript open source code of *Squirt*. It basically takes as input a text and a reading speed in words per minute (wpm) and displays this text word by word in the middle of the screen (see Figure 7.2). A blue letter marks the position where the reader's eye is supposed to focus on. This letter roughly marks the first third of the character sequence, since in western cultures the perceptual span is skewed to the right, hence we perceive more letters to the right than to the left of our focus point [144]. The mark should suggest the eyes to keep their focus on this focal area. The script analyzes the text upfront and dynamically assigns a viewing duration

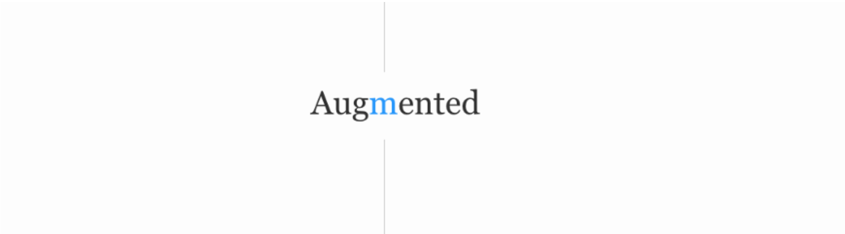


Figure 7.2: RSVP: sequentially displaying words in one focal point.

to each word, hence words with more characters are presented slightly longer. After a slight delay to prepare the reader, the presentation of words is quickly increased to its target speed. The same goes for words at the end of a sentence to indicate the beginning of a new one. The entire text is not displayed upfront and after finishing cycling through the text, the application closes its text window.

After a series of pilot studies where we raised readers' normal reading speed by varying amounts using both stimuli, we settled on an increase to 150%, since it seemed both significant and feasible. To assess the effectiveness of the system we conducted two user studies: 1) a lab study with 24 participants and a stationary eye tracker and 2) an online study to collect subjective feedback on the two speed reading stimuli.

7.2.2 Lab Study with Eye Tracker

First, we carried out a lab study to assess the effect of sped-up kinetic and RSVP stimuli on reading performance of users. We recruited 24 participants (22 male, 2 female) with an average age of 23 ($SD = 2.28$) years. We reached out to potential participants through university mailing lists and social networks. Most participants were students all of which indicated German to be their first language. Five participants indicated previous exposure to speed reading techniques.

Methodology

We designed the study using a between-subject measure design with the reading stimuli as the only independent variable. The stimuli comprised two levels: *kinetic* and *RSVP*. Participants were asked to read four rounds of text with one corresponding stimulus, for each stimulus condition we had 12 participants. As dependent variables, we measured text comprehension and mental load using the NASA-TLX questionnaire [124] after each round of reading. Using the eye tracker we also recorded average fixation durations and number of fixations as well as regressions calculated from saccades.

Apparatus

The study was conducted using the system described above. To record participants' eye movements we used the stationary SMI RED250 eye tracker with a sampling rate of 120 Hz. For study purposes, we integrated the browser-based prototype into a task sequence as defined in the study software of the eye tracker. To ensure the validity of measuring text comprehension we used an official text set from the *TestDaF* institute[24], which focuses on the development and application of tests to assess language proficiency of German as a foreign language. Each text comprised on average 583 (SD=19.8) words and came with a list of ten questions for measuring the readers' text comprehension.

Procedure

After explaining the purpose of the study, the participant was asked to sign the consent form. We then randomly assigned the participant to one of the two conditions. We calibrated the eye tracker and conducted a test to assure the eye tracker worked properly, after which the actual experiment was started. In the initial phase, we provided a text, which participants were asked to read as baseline condition without using any of our stimuli in order to calculate participants' regular reading speed (wpm). In the following, the participant read four texts in four rounds with 150% of her regular wpm rate. At the end of each round, she was asked to answer ten comprehension questions and fill in the NASA-TLX questionnaire. The study, including a short debriefing session, took approximately 60 minutes per participant.

Results

In the following, we present both quantitative and qualitative results.

[24] `https://www.testdaf.de`

Mental Load

We analyzed differences in mental load between the two groups using the RAW-TLX scores. The homogeneity of variances was not violated ($p > .05$). The independent t-test revealed no significant difference in the mental load between the two stimuli in any round ($t(22) = .889, p = .38$). The average mental load across all rounds for the kinetic stimulus was 64.80 ($SD = 4.81$) and for the RSVP was 67.41 ($SD = 4.35$). Further, we investigated how the mental load changed between the first and last round within each stimulus. For the kinetic, the t-test showed that the mental load decreased significantly from the first round ($M = 74.41, SD = 11.18$) to the last round ($M = 61.75, SD = 19.7$) , $t(11) = 3.22, p = .008, r = .70$. The effect size estimate indicates that the change in the mental load created by using the kinetic stimulus was a large and therefore substantial effect. No significant difference were found for the RSVP stimulus ($t(11) = 1.72, p = .11$). The average mental load for the first round was 70.75 ($SD = 8.89$) and for the last round was 63.58 ($SD = 19.30$).

Comprehension

We compared the number of correct answers for each text to assess users' text comprehension. The statistical analysis of the two stimuli revealed no significant difference ($t(22) = .62, p = .40$). The average number of correct answers using the kinetic stimulus was 5.48 ($SD = 1.63$) and using the RSVP stimulus was 5.18 ($SD = 1.72$). Further, we investigated the comprehension scores between the first and fourth round of each stimulus. For both stimuli the t-test revealed that comprehension increased significantly (kinetic stimulus: $t(11) = 2.80, p = .01, r = .65$; RSVP stimulus: $t(11) = 2.75, p = .01, r = .64$). The effect size estimates indicate a large and substantial effect. The average correct answer using the kinetic stimulus increased from 5.25 ($SD = 1.48$) in the first round to 6.08 ($SD = 1.67$) in the final round. Using the RSVP stimulus, the average number of correct answers was increased from 4.58 ($SD = 1.431$) to 6.0 ($SD = 1.65$).

Fixations & Regressions

We further analyzed the fixation and regression information collected by the eye tracker during the study. Fixations can be a measure to assess engagement or difficulties in extracting information [149]. The average number of fixations for the kinetic stimulus was 381.31 ($SD = 63.22$) and for the RSVP stimulus was 119.73 ($SD = 66.98$). Such a significant difference between the two stimuli was expected due to the nature of RSVP where the user focuses on a single point instead of moving eyes across a text. Further, we assessed the number of fixations and the average fixation duration between the first and fourth round of each stimulus. The t-test revealed no significant differences neither for the kinetic (number of fixations: $t(11) = -.07, p = .95$, fixation duration: $t(11) = -1.50, p = .16$), nor for RSVP (number of fixations: $t(11) = -.08, p = .94$, fixation duration: $t(11) = .49, p = .63$) stimulus. For the kinetic stimulus' first round, the number of fixations was 379 ($SD = 66.4$) with an average fixation duration of 241.25 ms ($SD = 29.82$), for the fourth round, the number of fixations was 379.92 ($SD = 66.1$) with an average fixation duration of 248.9 ms ($SD = 36.40$). During the first round of the RSVP condition, the number of fixations was 118 ($SD = 78.2$) with an average fixation duration of 1044.07 ms ($SD = 352.22$). In the fourth round, the number of fixations was 119.42 ($SD = 49.1$) with an average fixation duration of 996.1 ms ($SD = 385.5$).

We also evaluated the regression information collected while reading the texts using the kinetic stimulus between first and fourth round. We define regressions as eye movements opposite to the reading direction. While regressions are negligible when using RSVP, in normal reading they generally indicate re-reading of words or entire sentences and hence slow down the reading process overall. The t-test revealed the regression decreased significantly from the first round ($M = 16.42, SD = 11.15$) to the last round ($M = 11.25, SD = 7.1$), $t(11) = 2.877, p = .01, r = .65$. The effect size reveals that using the kinetic stimulus has a substantial effect on eye regressions while reading a text.

Qualitative Assessment

Taking a look at scan path visualizations of the eye tracking data of participants using different stimuli we notice some interesting differences. In scan path visualizations each eye fixation is represented by a circle. The longer the duration of the fixation, the bigger the circle's radius. A line represents a saccade between two fixations. Figure 7.3 shows the scan path of a participant freely reading a text without any of our two stimuli present. Fixation durations are quite variable and a number of line jumps as well as regressions can be noticed. Figure 7.4 depicts the scan path of that same participant directly thereafter when using the kinetic stimulus at 150% of her initially measured reading speed. Fixations seem more widely spread, which is probably due to the nature of the kinetic stimulus moving across all text. Figure 7.5 shows the scan path of a participant reading with the RSVP stimulus. Far fewer, but longer fixations on the central focus are presented.

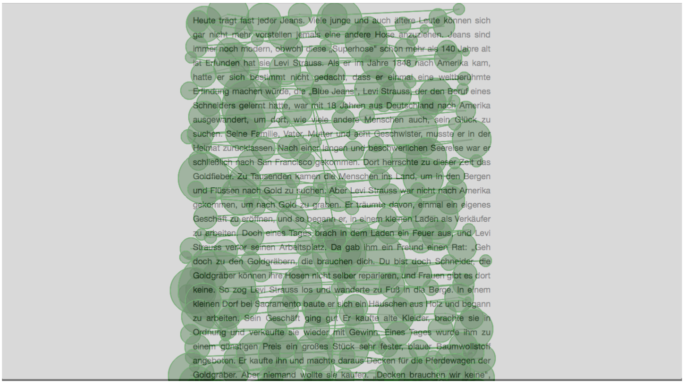


Figure 7.3: Scanpath visualization of reading freely.

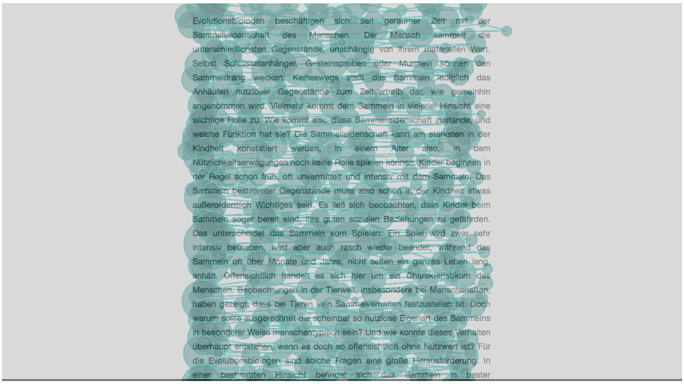


Figure 7.4: Scanpath visualization of reading with the kinetic stimulus.

Discussion

Looking at the significantly decreasing strain on mental load after using the kinetic stimulus for a while, we conclude a strong learning effect while using this method. The same applies to the significant increases in comprehension. Initially, users seem challenged by having a stimulus dictate them where to read and at which speed, but they adjust relatively quickly, *i.e.*, over the course of four trials. We further observed significantly fewer regressions after using the kinetic stimulus for a while. The eyes seem to adjust to following the stimulus and regressions are effectively reduced. The fact that, in accordance with fewer regressions, we

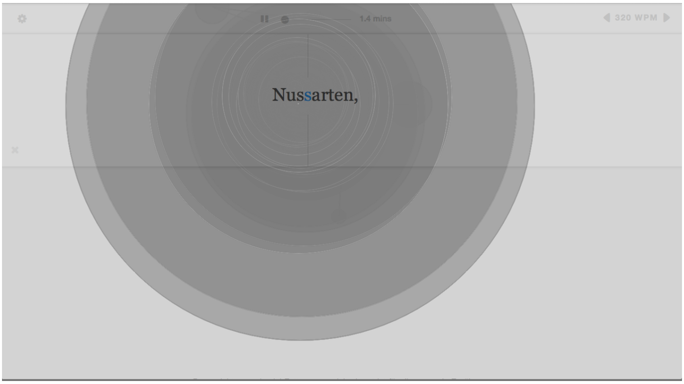


Figure 7.5: Scanpath visualization of reading with the RSVP stimulus.

also measured increasing comprehension levels contradicts findings of Schotter *et al.* [241], which is probably due to the nature of the text types used. Schotter *et al.* deliberately focused on ambiguous sentence structures in their studies. When designing such studies it is crucial to not only pay attention to the nature of the texts used, but also to the difficulty of assessing text comprehension. We opted for a solution in which the text came with pre-defined comprehension tests as they are used in assessing language proficiency. Hence, they require not an only literal translation, but also transfer skills. When reviewing the test results we realized these type of questions were anything but trivial. However, we are positive that considering certain text types, comprehension goals and given that users take into account a practice phase, using such stimuli for reading is feasible.

7.2.3 Online Study

To collect more in-depth data on user perception aspects of the presented stimuli we conducted an online study targeted at a broader audience. A total of 108 participants (72 male, 34 female, 2 without gender indication) between 12 and 69 years old ($M = 27.9, SD = 8.3$) fully completed the study. We disseminated the study across university mailing lists, social networks and personal contacts in other research facilities across U.S. and Europe. *German* was indicated to be the first language by 78 study participants, 30 reported it to be *English*. Participants' background ranged from students to engineers, business-related professions, and lawyers.

Methodology

The study followed a repeated-measure design, so all participants were exposed to both conditions: RSVP and kinetic as independent variables. For each stimulus participants were asked to read a short paragraph of text comprising 284 words each. As dependent variables, we collected subjective feedback in form of a Likert-style scale and free-text survey. The texts used for the study were taken from literature (*"The Trial"* by F. Kafka) and from a popular blog to ensure text diversity. We provided both English and German versions, self-selected by participants. Average time to complete the online study was M=8.5 ($SD = 4.8$) minutes, variations depending on individuals' baseline reading speed and the time they took for filling in the surveys.

Apparatus

We used again the basic web implementation of the stimuli as described above. Additionally, we used a server running PHP as the backend to collect and store the corresponding data. We designed a survey to be filled in after each condition. In the consent form, we collected basic demographic information, while a final survey was used to collect general feedback.

Procedure

On the landing page of the online study participants were asked to select their preferred language suggesting, *i.e.*, English and German. After, they were redirected to a consent form explaining the background of the study and where they could enter their demographics including age, gender, and profession. The first step comprised the assessment of general reading speed: participants were asked to fully read a paragraph of 521 words taken from *"Alice in Wonderland"* by L. Carroll. Completion time was measured from when they clicked on *'Start'*, on which the actual text appeared until they clicked on *'Finished'*. From this measurement, we calculated their reading speed in Words per Minute (WPM), which was taken as baseline speed. From there we randomly assigned both stimulus and text order and increased participants' reading speed by 150% to a maximum of 600 wpm. This cap seemed necessary to ensure that especially the RSVP condition did not completely overburden readers who might have read the initial text for speed assessment in overly rapid manner.

In the next step, for each stimulus one paragraph of text was read and five questions were answered using a 5-point Likert-style scale (1=strongly disagree, 5= strongly agree). Questions targeted subjective perceptions of text *comprehension*, reading *comfort*, *exhaustion* by reading, support in *speeding* up reading rates and perceived reading *duration*. After the last condition participants were directed to a final survey with three free-text questions where they could state their general preference for any of the stimuli. The questions aimed at eliciting what type of texts participants could imagine using the presented reading techniques for and on what kind of devices. In the last text box, we asked for general feedback and comments.

Results

Study participants on average reached 386 wpm ($SD = 194$) during the speed assessment task. Hence, for the study, we increased the stimuli speed on average to about 512 wpm ($SD = 107$). For 42 participants (38.9%) we capped the reading speed at 600 wpm.

Figure 7.6 comprises the results of the list of questions we asked after each stimulus condition. A Wilcoxon Signed-Rank test revealed that there is a significant difference in rating of subjective text comprehension ($Z = -2.18, p = .03$), speed reading support ($Z = -4.15, p = .0001$), reading comfort ($Z = -3.04, p = .002$), and perceived duration ($Z = -4.09, p = .001$). In these four aspects, RSVP was rated higher than the kinetic stimulus. The rating of exhaustion showed no significant difference.

In the final survey, 42.6% participants indicated to prefer the RSVP stimulus, 36.1% preferred kinetic, whereas 21.3% indicated no preference.

Text & Device Types

For the final survey of the study participants were asked to fill in for which types of text and devices they could imagine using these stimuli. For each stimulus different types of text and devices could be named. In total, we collected 106 recommendations for text types and 121 recommendations for device types. To identify the preferred types of text and devices for each stimulus, two researchers independently analyzed the free texts provided by the participants and categorized them. Then, the researchers crosschecked their categories and agreed on a set of text types and device types. In total four types of text, namely: *short texts*, *books*, *technical literature* and *news* were derived. The top three types of text participants mentioned for the kinetic stimulus being useful for were: books (30%), technical literature (23%), and news (15%). For RSVP, short texts were ranked highest (31%), then books (25%) and news (20%).

Regarding device types that may be fit for using the two stimuli for, 6 types were identified in total: *smartwatch*, *smartphone*, *tablet*, *e-reader*, *head-mounted display* and *PC/laptop*. The top three device types mentioned for the kinetic stimulus were e-reader (47%), tablet (45%) and PC/laptop (34%). The devices mentioned to be fit for RSVP were smartphones (68%), smartwatches (27%) and tablets (17%).

Qualitative Feedback

Looking at the general comments and feedback participants left in form of free text, we get a broader picture of subjective perception of using such sped-up stimuli for reading. Many participants doubted to have fully understood the texts but also stated that for unimportant text little comprehension may be acceptable. Others indicated having lost the context because of a moment of inattention or due to the fast speed of the stimulus. At least with the kinetic stimulus, some were able to catch up again, but with the costs of having missed some details in between. Many participants found the kinetic stimulus initially confusing while one of the problems with RSVP seemed to be a lack of sense of how far into the text the reader already was and how much more there was to come. One participant stated *"the surrounding sentence is missing"*, which shares the general assessment of others in the difficulty to put the single words into the overall context. However, one participant stated that sequential reading turns boring stories into interesting ones because she found *"fun in being challenged not to miss the context of a single word"*.

One participant expressed her desire to go through different paragraphs with different reading speeds as to use her imagination in crucial parts of a story. There were quite some complaints about the lack of manual control: *"The kinetic method would be reasonable if it had some sort of 'pause' functionality"*. One comment stated *"Speed reading is not about reading every word."* Similarly, another one stated that there needs to be the chance to get an oversight of the text while reading. Further, the techniques should allow skipping entire sentences or paragraphs. We also had participants who complained about the slowness of the stimulus, since we had a cap on 600 wpm for study purposes.

Numerous ideas to improve on the stimuli were brought forward as well. For example, the kinetic stimulus should be used for types of text where readers are inclined to easily digress (*e.g.*, mandatory texts/emails at work), especially for *"long passages I just need to get through"*. Another participant imagined the idea of using a kinetic stimulus for collective reading as well as on public displays. Further, it could be useful on large screens with lots of text as some sort of *"reading guide"*. Other application scenarios for the kinetic stimulus included highlighting and re-reading of important sentences/take-aways. RSVP, on the other hand, could be used for one-line ad screens at the bottom of cellphone screens or for displaying stock prices. Some participants mentioned that difficult words would need longer display time than others. One participant stated RSVP to be potentially useful for proofreading text.

Many comments focused on the issue of the speed of the stimuli in particular as well as increasing reading speed in general. One participant indicated the kinetic stimulus being *"too fast compared with my comfortable speed zone"*. A great number of users mentioned that they would appreciate using the stimuli at a slower speed. Some felt especially pressured by the RSVP stimulus as if *"being in a challenge"*. To conclude, one participant stated *"reading is more than simply a speedy transfer of data. Any 'quality reading' - at whatever speed - requires that the reader first understand the reading, next remember it, then analyze or intellectualize it from various reference points - in other words, think about the reading"*. Whereas another one summarized his comments with *"Reading should be for fun, and not a race"*.

Discussion

As can be seen from the general comments left by participants, the stimuli triggered some mixed feelings in user perception. Whereas in terms of measures the RSVP stimulus was clearly preferred to the kinetic stimulus, comments mostly revolved around the feasibility of the kinetic approach, given some sort of user control. However, findings seem to convey various application scenarios for both stimuli. In that sense, the kinetic stimulus was generally preferred to be used for rather long passages that require a certain amount of concentration. The RSVP stimulus, on the other hand, seems to be more suitable for short texts.

Further, users seem to prefer the kinetic stimulus when reading on sufficiently large displays as compared to RSVP, which they find feasible for small displays like on smartwatches or smartphones. Tablets seem to be the type of device that splits the categories: in case of larger available screen estate, RSVP was preferred; in case of smaller, the kinetic technique. This indicates that reading using two complementary devices as demonstrated by Piazza *et al.* [210] can make sense.

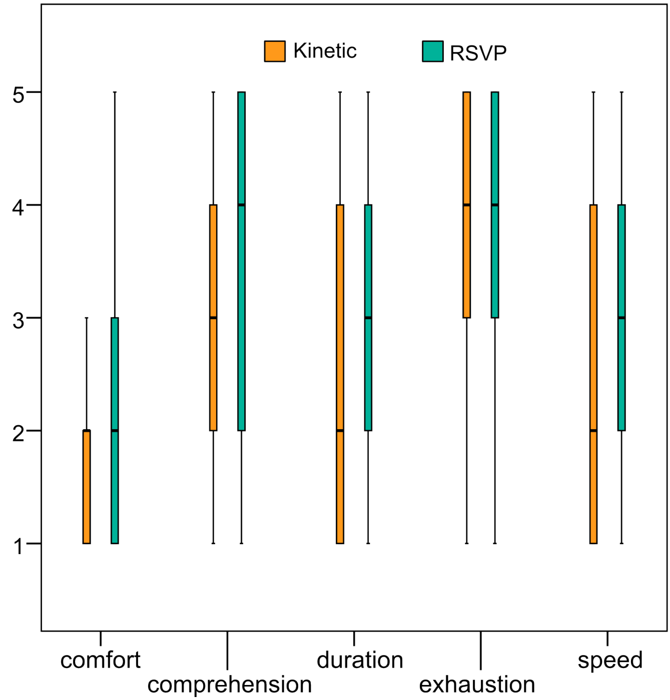


Figure 7.6: Subjective feedback on perceptions of text *comprehension*, reading *comfort*, *exhaustion* by reading, support in *speeding* up reading, and perceived reading *duration* using the kinetic and RSVP stimuli.

Speed was based on one single reading pass to determine the initial wpm rate. One participant stated he *"didn't pay much attention to the text during the speed assessment and hence skipped some parts"* which lead to a very high reading speed throughout the rest of the study. Further, these reading techniques are not suitable for all types of readings. It is generally difficult to decide on an appropriate selection of text passages. As the comments of some participants show, there are great differences in text complexity or readers' background knowledge for that matter, as certain paragraphs were already known by some participants. Further, not all of the participants were native German or English speakers and thus were overwhelmed by being pressured to increase reading speed in a foreign language.

7.2.4 Implications

We set out with the goal to assess the feasibility of applying speed reading techniques to reading activities on electronic devices. Our qualitative analysis shows that the act of enforcing higher reading speeds seems to have an alienating effect on users, especially when reading naturally comes with the idea of pleasure and relaxation. However, despite initial discomfort, we observed quick learning effects that lowered mental load and increased comprehension rates, which naturally go down when initially using such techniques. These findings encourage us to look into how such techniques can be further developed to lower the entry barrier for users to, slowly, but not drastically, pick up and increase their regular reading speed. Therefore, we identified a number of issues that need to be considered when designing such techniques:

Control

As we have learned from the qualitative feedback, it is crucial for users to remain in control over their choice of reading technique as well as over their current reading speed. Hence, possibilities to stop, start and pause stimuli need to be provided. Further, speed reading does not necessarily imply reading word by word. Hence, stimuli should take into account the users' inclination to skip words or entire paragraphs.

Retaining Oversight

What comes naturally to conventional books is a challenge for electronic reading interfaces: a sense of size, position, and oversight. When reading through an actual book, the page location and arrangement conveys a feeling of the whereabouts of the reader in the story. In ebooks, for example, this intuition is lost, even more so when using the RSVP stimulus as we have seen. Since a single eye fixation provides a view of the world that is roughly elliptical (about 200 degrees wide and 130 degrees high) [144], we can us this knowledge to adjust the number of words being displayed at once and make entire word groups graspable with one fixation, instead of displaying only one word at a time. Further, for text comprehension it seems crucial to be aware of the context of a word, sentence, even of the entire paragraph and hence adequate features should be provided.

Context-dependent Variability of Techniques

Different types of text require the reader to use different reading speeds. In such cases reading speed can depend on various factors, such as linguistic complexity or density of content, but also on the reading goals. In the first case, extracting language semantics could help as well as using reading models, such as [227], to dynamically adjust the speed of the stimulus. In the second case, readers should have the means to either push through a text or be allowed to skip certain parts where 100% comprehension may not be required, but skimming is acceptable. Further, device types should be taken into account since display size is an important factor for the choice of stimuli.

Taking into Account User Diversity

Reading is a highly complex psychomotor skill. There is a great variety of factors that influence reading performance, such as the reader's general background knowledge, familiarity with the language or with the type of text, and also eye mobility, attention span and current level of fatigue. Learning about users' reading habits can yield great adaption variability. Also, taking into account bio-feedback to dynamically pause stimuli or adjust reading speed may be feasible. Oliveira and Guimarães [204] presented a tool to assess mental workload from EEG signals and adjust reading parameters, such as text size, contrast and presentation speed in real-time. In case of high mental workload, text presentation can be slowed down to reduce discomfort and on the other hand accelerated to make use of available mental resources. Further, eye tracking can be used to take bio-feedback into account and dynamically pause stimuli or adjust reading speed in real-time. As one study participant stated: *"the kinetic [stimulus] could be a lot better if it were combined with gaze detection – this would allow me to flip back and re-process a sentence that I had missed. [...] It could auto adjust the speed, rather than just plowing on regardless"*.

7.2.5 Study Conclusion

In this study, we investigated reading UIs that can be adapted in real-time to support information intake and assessed effects on cognitive load. Such adaptions can take place according to the user's current state of concentration, for example, by balancing text throughput with the user's current cognitive capacities. We evaluated two approaches to increase reading speed on electronic devices by applying a kinetic and an RSVP stimulus to text. Therefore, we implemented an animated line that moves through the text as well as an RSVP stimulus, each moving at 150% of the reader's regular speed. In two user studies, we collected quantitative and qualitative data on the effects and feasibility of such stimuli. Despite users being initially alienated by the approach, results show quick learning effects in adjusting reading speed, lowering mental load and increasing text comprehension levels. We concluded with a set of design guidelines for applications using such reading techniques: therefore users should be able to control their speed and mode of reading and be allowed to retain oversight. Readers' individual preferences and reading goals should be taken into account as well as the different types of text and particularities of devices. By combining eye tracking with natural language processing, systems can detect the reader's skill level, automatically assess the peculiarities of text types and adjust the text display accordingly. Hence, such systems can offer different reading strategies to facilitate reading tasks according to the user's current cognitive state and reading goal.

In the next Section, we focus on the notion of controlling the reading flow through eye tracking. Eye movements are strong indicators of user interest and cognitive activity. Hence, we present another application of cognition-aware systems that process eye movement data and adjust the reading UI in real-time to support information intake.

7.3 Implicit Reading Support Through Eye Tracking

As we have assessed in the previous Section, reading with RSVP requires high user attention since words are flashed in rapid sequence. This makes reading while performing an additional task, such as walking, challenging. Even small distractions cause users to briefly look up from the screen and hence lose part of the currently displayed sentence.

To avoid missing vital text parts, controls are required to pause and resume the flashing of words. RSVP is typically controlled using explicit user input. However, pressing a button or performing a gesture requires time, attention and accuracy. Explicit input can, therefore, be challenging for short and frequent interruptions. Looking up from the screen while walking to adjust the walking

Figure 7.7: RSVP reading interface with four visual markers augmenting the smartwatch to detect when user's eye gaze is directed at the watch.

direction, for example, can take less than a second. For such brief interruptions, we created a system, which allows users to implicitly control RSVP text presentation. Therefore, we augmented readers with a mobile eye tracker to detect when the reading flow is interrupted and pause the sequence of words accordingly. In case of longer interruptions, the text backtracks to the beginning of the sentence to restore the reader's context. Hence, we aim to combine the advantages of normal reading, which allows readers to freely switch their focus between text and environment, with the advantage of RSVP requiring very little screen space.

In this Section, we present and evaluate a system that uses RSVP to display text on a smartwatch and allows users to explicitly or implicitly control the text presentation. In the following, we report on a user study with 15 participants, in which we investigated the advantages of implicit eye gaze control.

7.3.1 Explicit and Implicit RSVP control

We developed a prototype that enables users to explicitly or implicitly control RSVP text presentation. The system consists of a smartwatch with a touchscreen and an eye tracker. Therefore, we implemented an Android Wear RSVP application based on an open source framework[25] for the Motorola Moto 360 smartwatch. The Moto 360 runs Android (version 6.0.1) and has a circular 320×290 pixel 1.56″ touchscreen display with a 46 mm diameter. The application displays text,

[25] `https://github.com/andrewgiang/SpritzerTextView`

which is shown word by word with a specially colored letter serving as a focal point for the reader's eyes to focus on (see Figure 7.7). The number of words shown per minute (*e.g.*, 200 wpm) can be freely selected from the application's setting.

Users explicitly control the text flow through the touchscreen of the smartwatch: reading can be paused and resumed through simple taps (Figure 7.8 (left)). To allow readers restore the reading context, the text presentation goes back to the beginning of the current sentence in case of longer pauses (*i.e.*, longer than 5 seconds). We use an eye tracker to enable implicit control over the text presentation (Figure 7.8 (right)). For tracking the user's eye movements, we use the Pupil Pro eye tracker by Kassner *et al.* [151], which comes with a 3d-printed frame and a software package for calibration, gaze detection, and surface registration. The modular eye tracker consists of a 120 Hz head-mounted monocular camera with a resolution of 640×480 pixels and a world camera with 30 Hz which delivers the video stream in FullHD. We attached four visual markers, each with a size of 20 mm × 20 mm, to the bezels of the smartwatch screen (see Figure 7.7), which allow the software to determine whether the eye gaze is directed inside or outside of the marked rectangle, *i.e.*, users look at the smartwatch display or not. The gap between markers constitutes 14 mm (vertically) and 64 mm (horizontally) so that the diameter of the registered virtual surface corresponds to the diameter of the watch interface. The eye tracker communicates with the smartwatch using the Pupil Capture software[26], which entails the Pupil Server plugin for broadcasting eye gaze data over a network. The Android Wear application on the smartwatch receives the data stream and determines whether the user is currently looking at the watch or not. Hence, the reading flow pauses when the user looks away and automatically resumes when the user looks back at the watch. In case of a longer reading pause (longer than five seconds), the reading position is further reset to the beginning of the current sentence.

7.3.2 User Study

To compare explicit with implicit control of RSVP reading flow we conducted a user study where participants read texts with the system described above. We hypothesized that an adaptive reading interface taking into account the user's eye gaze would lead to higher text comprehension and reading confidence.

[26] `https://github.com/pupil-labs/pupil`

Study Design

We employed a repeated-measures design with the RSVP control modality being the independent variable resulting in two conditions: 1) manual control through touch interaction (*tap*) and 2) implicit control through eye gaze (*eye gaze*). For each condition, we introduced a secondary task: while reading on the smartwatch, participants were asked to monitor words displayed on a desktop monitor in front of them. Shown words were either countries or city names. If a word was a country name, participants had to press a button on the keyboard. City names had to be ignored. Words were shown for 10 to 15 seconds. As dependent variables we measured 1) overall task completion time, 2) comprehension scores, 3) tracked eye movements (*i.e.*, reading pauses), 4) errors on the secondary task, *i.e.*, number of countries missed, and 5) measured mental load using a NASA TLX questionnaire after each condition.

Procedure

After welcoming and introducing participants to the purpose of the study we asked them to sign a consent form and recorded demographic data. We then introduced the RSVP reading interface on the smartwatch and set up the mobile eye tracker, which required a brief calibration for each participant. We explicitly asked to keep the arm wearing the watch rather still so that it remained in the camera view of the mobile eye tracker. After they had familiarized themselves with the interface and the available controls, we assigned participants to a starting condition. We then asked them to read while completing the secondary task, whereas we instructed them to treat both tasks as equally important. After each text, we administered a 10-item comprehension test. For each condition, participants read two texts, after which they completed a NASA TLX questionnaire. We counterbalanced both the sequence of conditions via latin-square as well as the allocation between tasks and reading texts. All texts were taken from a collection used for English as a Second Language (ESL) learners, an adaption from [220]. These texts came with a predefined set of ten comprehension questions per text and comprised of average 548 words ($SD = 2.87$). Since participants read two texts per condition, the maximum comprehension score per condition was 20. The study took about 50 minutes per participant, which was concluded with a final questionnaire.

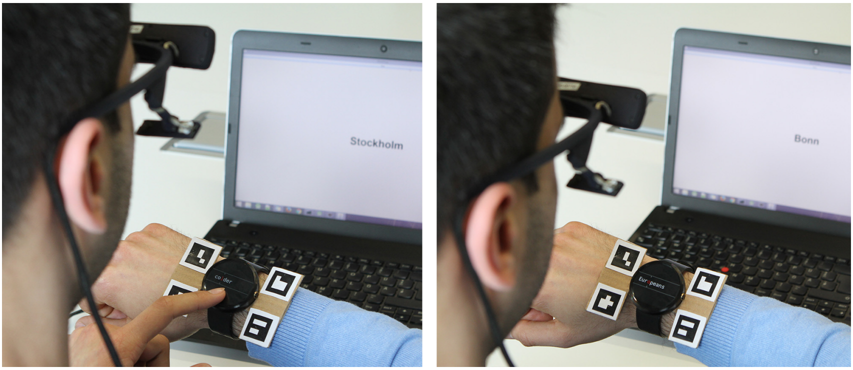


Figure 7.8: Participant reading a text on a smartwatch while monitoring city or country names on a secondary display. Left: the reading flow is controlled explicitly through touch. Right: the reading flow is controlled implicitly through eye gaze.

Participants

We recruited 15 participants (11 males, 4 females) through a university mailing list. With a mean age of 26.5 years ($SD = 3.5$) most had a background in IT or were university students. All reported English to be their second language, 3 were wearing contact lenses (20%), 5 glasses (33.3%). 8 of them indicated to be familiar with the RSVP reading technique (53.3%), 8 were wearing watches on a regular basis (53.3%), none of which were smartwatches. Participants were rewarded 10 EUR for taking part in the study.

Results

Each participant read in total four texts, two in each condition. Table 7.1 summarizes the descriptive system measurements.

Objective Measures

A Wilcoxon Signed-Ranks Test revealed that the median of the comprehension scores for the eye gaze condition ($Md = 18$) was significantly higher than for the tap condition ($Md = 16$), $Z = 74.5, p = .041$. Thus, participants had a better

	explicit interaction			implicit interaction		
	Md	M	SD	Md	M	SD
Number of pauses	8	13.5	11.73	29	30.9	6.14
Percentage of missed countries	0	14.02	18.54	6.25	9.84	12.59
Num. of mistakes on secondary task	0	0.6	0.74	1	0.93	1.22
Task completion times (second)	297.2	313.3	46.4	352.4	354.1	45.3
Comprehension Scores	16	15.86	1.88	18	17.33	1.67

Table 7.1: System measurements for both conditions: explicit (touch) and implicit (eye gaze).

understanding of the read text when reading with implicit control using the eye tracker. Another Wilcoxon Signed-Ranks Test also revealed that the median of the average number of pauses for the tap condition ($Md = 8$) was significantly lower than for the eye gaze condition ($Md = 29$), $Z = 105.0, p = .001$. In sum, participants made 9 mistakes on the secondary task, *i.e.*, cities were selected instead of countries, during the tap condition ($Md = 0$) and 14 mistakes during the eye gaze condition ($Md = 1$). We found no statistically significant differences between the percentage share of missed countries while tapping ($M = 14.019, SD = 18.539$) vs. while using eye gaze ($M = 9.842, SD = 12.599$), $t(14) = 1.549, p = .144, r = .842$. There was a statistically significant difference between task completion times for the tap condition (in seconds: $M = 313.3, SD = 46.4$) and the eye gaze condition ($M = 354.1, SD = 45.3$), $t(14) = -2.383, p = .032, r = .044$, where participants took more time when reading with implicit control over the text presentation.

Subjective Feedback

As for the subjective assessment through the Nasa TLX questionnaire, a student t-test showed no statistically significant difference between the tap condition ($M = 10.278, SD = 2.564$) and eye gaze condition ($M = 10.289, SD = 2.293$), $t(14) = -.014, p = .989, r = .190$. Thus, we found no effect of the control mechanism on the perceived mental load. In the final questionnaire, most participants reported the frequent taps to pause the text to be annoying and hence preferred the support through eye gaze. However, they felt a lack of control, since the implicit pauses entailed no noticeable feedback.

7.3.3 Discussion and Limitations

Comprehension was higher for implicit than for touch interaction. Therefore, eye gaze interaction seems to be less distracting than when explicitly having to control the reading flow, which confirms, our hypothesis. While the average number of pauses was also higher in the eye gaze condition, switches between tasks seem to have been done with ease: participants reported in the explicit interaction condition that they sometimes looked up without pausing the text flow in order to take a brief glance at the secondary screen. Hence, they compromised on text comprehension, whereas the eye gaze interaction implied a pause during a quick glance. This is also what contributed to the overall increase in task completion time. The monitoring task we employed as a secondary task was meant as a distraction task. Because of the eye tracker's sensitivity to large head motions, we refrained from having participants perform a walking task. With more robust trackers we would like to test our hypothesis in the context of a navigational and therefore spacial task. Although we designed the eye gaze interaction with RSVP in mind, a similar approach could be taken to control the flow of automatic scrolling through text displayed on a conventional reading interface. Other reading techniques as proposed in the previous chapter could also benefit from eye gaze tracking to dynamically adjust reading speed, for example. Further, we envision front-facing cameras soon to be able to detect the user's eye gaze which will render the mobile eye tracker obsolete and allow for more widely applied eye gaze interactions, also for reading UIs.

7.3.4 Study Conclusion

For investigating the adjustment of reading UIs through implicit user feedback we built a working prototype that controls the reading flow in RSVP by tracking the user's eye gaze. When reading on mobile devices screen size is limited and distractions from the environment can hinder the reading flow and negatively affect comprehension. Context- and cognition-awareness systems can adjust their reading UIs to compensate for this effects: by automatically pausing the reading flow when users are detected to be distracted we are able to facilitate information intake. In a user study, we compared the implicit interaction through eye gaze with explicit user controls. Results show that gaze interaction is more effective than having users explicitly pause and resume reading through touch. Hence, tracking and processing eye movements in real-time can be used to create more effective reading UIs that support reading in distractive environments.

7.4 Chapter Summary

In this Chapter, we explored reading as an application scenario for using awareness of users' cognitive capacities to support information intake. Since information is consumed to great parts in the form of text, we investigated how reading interfaces can be adjusted according to readers' cognitive capacities. Therefore we explored dynamic adjustments of reading UIs that allow readers to regulate their reading speed and reading flow. An increase in reading speed generally implies an increase in cognitive demand and compromises on text comprehension. Such compromises can be acceptable if a text does not necessarily need to be processed in full detail, but rather with regard to its gist. In phases of high user alertness more cognitively demanding tasks are feasible, hence an adaptive reading UI can invite readers to read at faster rates.

To investigate the relationship between reading speed, imposed cognitive demand, and text comprehension we conducted two user studies, in which we assessed the use of two different reading methods: 1) we developed a kinetic stimulus in form of a moving line that guides the user's eyes across text and 2) we used RSVP to show words sequentially at a focal point. Both methods superimpose reading speed on the reader and can dynamically be adjusted. In a controlled lab study, we assessed these methods with regard to increased reading speed, namely 150% of readers' normal reading rates. We found substantial learning effects for perceived mental load and text comprehension: when reading with our kinetic stimulus mental load was initially high when it was introduced, but was significantly decreased over the course of four short texts. Similarly, comprehension rates went up over time for both the kinetic and the RSVP condition. Since eye movement patterns (*i.e.*, number and duration of fixations) remained mainly the same, the improvements were likely due to cognitive adjustments. In the second study, we collected subjective feedback through a web experiment, which also revealed that users initially felt challenged by being dictated how to read. While RSVP was seen as appropriate for reading a rather short text, the kinetic stimulus could be used for longer text where a larger display is required. We derived a set of design implications for adaptive reading UIs that comprise recommendations for providing readers with flow and speed controls, allowing them to retain oversight of text position, providing a variability of reading methods and speeds that also depend on the text to read, and taking into account individual user characteristics, such as preferences, reading skill, and current cognitive capacity.

Eye movements are strong indicators of user interest and cognitive activity. Hence, in the second part of this Chapter, we addressed the implicit control of reading flow based on eye gaze. Therefore, we equipped readers with a mobile eye tracker and assessed their ability to read on smartwatches while being engaged in a secondary task. While reading when being on-the-go, we are exposed to a distracting environment and possibly need to navigate that environment at the same time. Due to the rapid flashing of visual stimuli when reading with RSVP, single words can easily be missed and even more so when the reader is engaged in a secondary task, such as walking. Hence, we used a mobile eye tracker to inform the reading UI when the reader focuses on the text or elsewhere. The interface then pauses, resumes or resets the text dynamically in order to support the reader. In a lab study, we investigated the effects of gaze-aware reading UIs on text comprehension, secondary task performance, and reading time. Compared to explicit controls provided by a touch interface, implicit adjustment of the interface lead to higher comprehension rates while performance on the secondary task was comparable. The implicitly controlled interface introduced more pauses in the reading flow compared to the explicit condition, in which readers often did not pause the text to quickly deal with a secondary task and hence missed some keywords resulting in lower comprehension rates. The study showed the feasibility of using eye gaze to inform reading UIs about readers' focus and provide in-situ reading support.

Concluding, there is a general trade-off between reading speed and comprehension. In this Chapter, we showed that speed reading stimuli can be used during reading to enforce higher reading speed, which mentally strains the reader. To answer **RQ6** we developed a set of reading UIs that allow the dynamic control of reading speed and reading flow in real-time. Cognition-awareness can be used in systems to proactively regulate these parameters according to the user's current cognitive state. The goal of matching reading mode with the current cognitive state text is to allow readers to read more efficiently (or more leisurely) by balancing the cognitive load, reading throughput and retention.

IV

Conclusion and Future Work

Chapter 8

Conclusion and Future Work

In this thesis, we investigated how ubiquitous technologies can be used to help people deal with information intake and learning tasks through cognition-awareness. Through a series of lab studies and field experiments, we explored human attention and user context while interacting with ubiquitous technologies and how awareness of users' cognitive context can be applied to a range of applications. In the following, we provide a summary of the research presented, outline the contribution with regard to the research questions, and conclude with an outlook on future research directions.

8.1 Summary

As knowledge grows at an ever-increasing pace, people face the constant challenge to consume information and acquire new knowledge. While ubiquitous computing devices are deeply ingrained in our lives and therefore provide information access anytime and anywhere, our attention is generally limited. One of the main reasons for failing to remember a piece of information is what Schacter [234] describes as *absent-mindedness* at the point of encountering that information. When attention is not focused, information is not encoded correctly and therefore not memorized. Hence, attention is a crucial piece of effective information intake and processing. In Chapter 3 we started out by investigating people's attention in a technological context. Posing the question "*How can users' attentiveness be quantified across* *the day and reliably predicted from phone usage patterns?*" (**RQ1**) we conducted a field study collecting data on mobile messaging behavior and related phone usage. Results showed that people are highly attentive to messages throughout the day and that phases of inattentiveness are generally brief. However, the fact that people spend a lot of attention to their devices does not say much about why they turn to their phones and in what kind of attentional states. We were especially interested in idle moments, in which users engage with their mobile device actively seeking stimulation. Hence, we investigated the question "*Does boredom measurably affect phone usage and which usage characteristics are most prevalent in such states?*" (**RQ2**). Assuming that people would be more receptive to engage with content suggestions in such moments of boredom, we conducted a field study using *Experience Sampling* and collected corresponding phone usage data. By training a machine-learning model we were able to predict phases of boredom with 82.9% accuracy and identified usage characteristics, such as the recency of communication activities, phone usage intensity, time of day, and demographic information to give most prevalent indications for predicting phases of boredom. The prediction model elicited would be the basis for further content suggestion studies described in Chapter 6.

The engagement studies we conducted, show how attentional states are linked to people's mobile device activities. In the next step, we focused the user's situational context to support information intake. Chapter 4 considered context in the form of content-awareness: this investigation aimed to answer the question "*How can awareness of the content which the user is currently exposed to be used to augment the user experience?*" (**RQ3**). Tools that proactively provide additional information and explanations to complement the content can support users building out multiple associations and therefore helps to memorize that content. Such tools or applications need to be able to detect the content consumed in real-time and across device boundaries. Hence, we built a system to augment the TV experience that taps into the stream of subtitles, extracts keywords, and provides additional information by retrieving content from Wikipedia and Google searches. These contents are displayed on the user's second-screen phone app and turn out to be highly relevant, thereby enhancing the subjective TV experience.

Having explored user context in terms of current attentiveness and content consumed, in a next step we focused on more regular patterns of attention as they are dictated by our internal body clock. These diurnal fluctuations impact the users' levels of alertness and cognitive performance in the course of the day. Being aware of these circadian rhythm can inform technologies to adapt information selection and presentation in ways to match users' current processing capacities. To answer the question "*How can technology be used to elicit the user's circadian* *rhythm of attention and cognitive performance?*" (**RQ4**) we adapted a sequence of three cognitive tasks to a mobile context. With data from a mobile field study, we were able to recreate a general model of people's attentional fluctuations across the day. We open-sourced the system in order to allow future psychology experiments to be conducted in-the-wild, and which we see as a starting point for building cognition-aware systems. We envision such cognition-awareness to help build applications that schedule tasks during respective phases of the day and adjust interface complexity and information bandwidth to match the user's current state. Hence, in phases of high concentration complex information could be efficiently displayed, whereas in phases of low concentration complexity could be removed to prevent information overload and frustration. Besides the open-source toolkit, we presented a conceptual framework based on our learning from the attention elicitation study based on phone usage data. The framework describes a three-step approach to building algorithms that detect cognitive states of interest: (1) it maps out how to collect ground truth data, (2) elicit meaningful features, and (3) train machine-learning models that can be integrated into applications. To validate this approach we focused in the next step on concrete application scenarios of cognition-aware systems.

This final part of the thesis (Chapter 6 and 7) presents a series of research probes that explore the application of systems that take into account users' current attentiveness and cognitive capacities. In Chapter 6 we report on two studies in which we explored content recommendations in opportune moments. To answer the question "*How can opportune moments for content delivery be used to foster information intake and learning?*" **(RQ5)** we built two mobile systems, with which we investigated characteristics of such moments for learning and reading. While we used our previously developed boredom prediction algorithm to trigger content suggestions, we did not find sufficient evidence for a correlation between boredom and opportune moments of learning. However, based on collected phone usage data and ground truth in the form of compliance rates to learning triggers, we trained another prediction model and elicited indicating features. When suggesting entertaining reading content during phases of predicted boredom, we noted users were more open and more engaged with the content than during non-bored phases.

Besides the identification of opportune moments for delivering intake, we also explored how content can be adjusted in real-time to match levels of high or low cognitive load. The higher the user's current cognitive capacities, the more prospective applications can take advantage of an increased information bandwidth. In Chapter 7 we, therefore, explored adaptive reading UIs, with which we investigated the interplay between reading speed and text comprehension. To answer the question "*How can information displays and more specifically reading UIs be adjusted in real-time to increase or decrease information bandwidth?*" (**RQ6**) we investigated so-called speed reading interfaces, both of which allow the UI to superimpose reading speed on the user. Hence, we applied RSVP to reading as it had been suggested as a feasible reading technique on small-screen displays. We further introduced a kinetic stimulus that helped guide the reader's eyes across text thereby effectively increasing reading speed. In order to maintain text comprehension, a high user focus is required, which is why such reading techniques are not always feasible. Awareness of concentrated states can help users pick a corresponding reading strategy to optimize their information intake in a given state. Since during reading activities tracking eye gaze can inform systems about the user's focus of attention, we investigated implicit eye gaze interaction in the context of RSVP reading. In a lab study, we found implicit eye gaze interaction to lead to higher text comprehension than explicit touch controls for pausing and resuming reading flow.

8.2 Contribution and Results

In the course of this thesis, we followed an experimental approach to answer six research questions that were targeted at investigating ways to use ubiquitous computing devices that take into account their users' cognitive states to support effective information intake and retention. In the following we will briefly summarize the five key contributions made:

8.2.1 Quantification and Assessment of Users' Attentiveness

Our study on mobile messaging behavior (see Section 3.2) showed that people are highly attentive to messaging, namely for 73.5% of their waking hours, which constitutes for more than twelve hours per day. Phases of inattentiveness are typically brief and only last for a few minutes, which renders delay strategies for notifications or alerts feasible. Such strategies can, therefore, be applied to avoid user disruption during those inattentive phases, while the risk that important information might be missed is negligible. There are numerous cases where alerts are not required to bring back user attention to the device. In idle moments people turn to their phones seeking stimulation. To detect such phases of boredom we conducted a field study (see Section 3.3), which allowed us to build a prediction algorithm capable of distinguishing bored *vs.* non-bored with an accuracy of close to 83%. Our examination of corresponding phone usage data showed that detecting states of boredom is feasible by looking at the most prominent features, such as the recency of communication, usage intensity, time of day, and demographic information. Phone awareness of such moments can be used for benign interventions, such as sending content recommendations in form of reading or learning materials.

8.2.2 Tools for Researching and Building Cognition-aware Systems

Based on our studies in Chapter 3 we derived a framework in Section 5.3 with which researchers can build cognition-aware systems in the context of ubiquitous computing. The framework can be used to relate cognitive states to technology usage patterns by combining user assessments or inferences with machine learning. By collecting ground truth together with context sensor data (step 1), extracting features from that sensor data (step 2) and training and applying classifiers (step 3) we can build machine models based on usage correlations to detect and predict user behavior and states. We demonstrated how ground truth collection by *Experience Sampling* (Section 3.3) and observation (Section 6.2) results in rich data sets. We further presented a mobile system that adapted standard tests from previous studies in psychology and deployed them in a mobile field study. We showed the effectiveness of using mobile tasks to assess people's cognitive capacities throughout the day and elicit their circadian rhythm of alertness. Over the course of a single week, we were able to elicit a general model of people's circadian rhythm of alertness and performance, which in previous works took either significantly longer or was significantly more cumbersome to attain. Awareness of users' diurnal performance variations across the day can help inform systems about the user's current cognitive state. To help researchers build such cognition-aware systems we released our system as an open source project.

8.2.3 Identification and Characterization of Opportune Moments for Content Delivery

By applying the proposed framework we conducted a series of studies (described in Chapter 6) to investigate the nature of opportune moments for information intake in two ways: vocabulary learning and reading activities. We showed the feasibility of mobile devices promoting microlearning through proactive reminders whose scheduling can be trained based on observed phone usage patterns. Microlearning tasks are especially well received when people are on the go and face idle moments, such as waiting in line for the supermarket. As for finding reading slots during the day mobile devices can not only be used to remind us to work on our reading list, but also select reading content based on the current user context. Awareness of current time available and phone usage patterns can help match user situation and reading material. In a related study (see Section 6.3) we used our boredom detection algorithm to successfully recommend random reading articles to users in states of boredom. In such predicted states users were shown to be significantly more likely to open and engage with suggested content on their mobile phones. Hence, we validated the application of the proposed framework for building cognition-aware systems and its feasibility for gaining insights into the relationship of different cognitive states and related user activities.

8.2.4 Implications for Adaptive Reading

Our studies about dynamically adapting reading UIs have shown that the readers' information intake can be influenced by using reading stimuli. Besides the established method of RSVP, we introduced a kinetic stimulus which allows readers to keep the entire text line in view while providing eye guidance (Section 7.2). There is a trade-off between speed and comprehension, but if a thorough understanding of a text is not the goal, increasing reading speed allows the reader to get through text faster at a moderate comprehension level. Reading speed has an effect on the reader's mental load. Cognitive load, in turn, depends on the user's current cognitive capacities. Hence, reading interfaces with cognition-awareness could adjust their cognitive demand to optimize for information bandwidth between interface and user. Therefore, we proposed implicit and explicit ways to control reading speed: implicitly through eye tracking, which we showed to be more feasible than explicit user control through touch (Section 7.3).

8.3 Limitations

Throughout the course of our research we have been taking an empirical approach by conducting lab studies (*e.g.*, to assess mental load and text comprehension) and in-the-wild deployments (to take user context in the real world into account). The lab studies (*e.g.*, in Section 7.2) were necessary to isolate the effects of our probes on the user. But with regard to cognition-awareness and daily information intake external validity would be limited in lab environments due to artificial settings and limited time constraints. This is the reason why we took most of our prototypes to the wild by conducting field studies. By doing so we gave up a certain amount of control over our variables and causal relationships that might exist between user behavior and our measures, which we might not be able to detect. However, since most of our probes require users to exhibit everyday usage patterns, this approach was vital.

To infer cognitive states we correlated user behavior with the ground truth we collected. This ground truth may not always reflect the complexity of cognitive interactions, but is rather a simplified abstraction. Relying on user's self-reports can introduce users' biases and misconceptions. Observing usage patterns and making inferences from user activity may be confounded by external influences. On the other hand, the correlations we identified may not be directly resulting from the respective cognitive states, but features may rather become indicative when combined with others. Researchers applying our proposed framework should be aware of these potential pitfalls. However, as our validation studies showed, application of prediction models might still work despite possibly confounded underlying relations. Also, correlation does not imply causation, so we have to be careful when making statements about the exact relationships between features.

Device usage is also rarely the sole predictor of cognitive states. People's focus is influenced by a number of aspects, such as sleep, general fitness, nutrition or environmental factors. However, the nature of our applied research approach is more focused on the application layer, rather than claiming to uncover psychological truths. For systems to become cognition-aware, monitoring context factors have shown to be sufficient for indicating states of attention and receptiveness.

We created prediction models based on general usage patterns, not based on individuals' data. This may be sufficient for proof-of-concept prototyping, but future models and algorithms should consider continuously learning from the user's behavior in order to adapt to individual and possibly changing usage patterns.

While the applications presented in part III are linked in theory to cognition-aware system triggers, a fully integrated system for eliciting cognitive states to adjusting the UI, for example, requires thorough design and evaluation. This includes privacy implications that such sensitive data collection implies. Further, by adjusting UIs to cognitive state such adjustments may change the user state in return. The investigation of this feedback loop—and the corresponding search for a "sweet spot"—between user and application (as suggested by Schmidt [238]) will need to be picked up by future work for which this thesis provides a starting point.

8.4 Future Work

This thesis provides a set of tools for future research on cognition-aware systems. While we focused on their applications for supporting information intake and learning, we came across a number of application scenarios and use cases these might be used for. Also, during the course of our research on building these tools and applications, we identified several additional research challenges which are beyond the scope of this thesis. In the following, we will lay out how future research and development can be continued immediately, mid-term, and more long-term.

Research projects that can be **immediately** picked up concern the application of the tools provided by our work for more general validation purposes, but also for applying them to new application scenarios. The conceptual framework, for example, can be applied to elicit and predict a range of user behavior and states. While we mainly focused on attention and receptiveness, future experiments could focus on relating phone usage patterns to different emotions and affects. Mottelson and Hornbaek [191] showed the feasibility of eliciting positive affect from smartphone sensor data linked to movement. Using experience sampling or an analysis of current writing style (using language processing) could provide systems with further insights into the user's current mood and emotional state. While emotions impact memorization [12] users may be more inclined to certain activities in negative *vs.* positive emotional states. The same accounts for quiet or aroused states. A related research question concerns the relationship between people's individual chronotypes and the emotional patterns that unfold across the day. If such patterns can reliably be elicited, systems aware of users' affects can schedule activities across the day, recommending to engage in communication tasks, for example, in phases of positive affect. The range of possible applications should be explored in more detail. While we focused on content suggestions in the form of learning and reading content, receptiveness to suggestions to go for a run or call a friend, for example, might be inherently different. Further implications for the advertising industry should be investigated: since in our attention economy commercial services compete for being noticed, receptive states are worth exploring for their economic value. Advertisers might be willing to pay higher prices for ad impressions triggered during attentive rather than inattentive states.

The mobile toolkit we presented in Section 5.2 can also be the basis for future investigations into a) how to apply awareness of user alertness and cognitive performance in different applications but also on b) how to elicit these circadian rhythms in more implicit ways. Performance measures, such as reaction time and error rates when trying to inhibit or follow a stimulation can be integrated in gaming applications, where the player's performance in a game such as "Whack-a-Mole"[27] could be used to elicit the circadian rhythm of alertness. The spatial task in the form of multiple object tracking, which we applied, can be found in games, such as "Air Commander"[28], hence different game designs and mechanics should be explored to record such performance measures throughout the day. The feasibility of this approach would need to be tested and mechanisms to be applied that invite players to play the game during different times of the day rather only when they feel like it and explicitly launch the game.

Research projects with a more **mid-term** time horizon (*ca.* 3-4 years) could focus on how to integrate a network of devices that record activities and derive cognitive states. Arising challenges include how an ecosystem of devices could be realized that applies cognition-awareness, communicates beyond device boundaries and splits up learning tasks across devices and times of the day. For example, keywords could be extracted from a foreign language article that the user reads on a tablet while commuting to work in the morning. The smartphone then invites the user to review these words in opportune moments throughout the day. These repetitions can be spaced out over time so that these words are transmitted to long-term memory. Pronunciation exercises are then scheduled at night when users sit in front of a laptop equipped with a microphone and where they can speak freely. Such interweaved systems entail a networked system architecture including wearables, phones, and stationary devices and require research in context sensing, cross-device task sharing, application scenarios, and social acceptability. Such an integrated research project provides the basis for a ubiquitous personal assistance system that accompanies users throughout their day and helps them with scheduling pending tasks according to their situational and cognitive context. Future research on this topic further entails disruption management and how such a personal device ecosystem can help users manage interruptions throughout the day, stay focused, but communicate interruptibility in opportune moments.

[27] `https://play.google.com/store/apps/details?id=com.feasy.game.MoleWhack`

[28] `https://play.google.com/store/apps/details?id=com.gunrose.aircommander`

As we pointed out in our ethical considerations, making sure we provide reliable privacy and security features, is essential for the long-term acceptance of such systems. Initially, rule-based learning rather than black-box machine learning could help foster user acceptance. If users can relate to why certain recommendations are triggered at certain times and can further confirm or dismiss their usefulness, these systems can learn from this feedback loop and increase user comfort. Future research that focuses on privacy and security should investigate how to make cognition-aware systems secure, transparent, and socially acceptable. By applying concepts, such as privacy by design and homomorphic encryption users might be more willing to grant technologies comprehensive access to their personal data, habits, and behavioral patterns.

As for **long-term** time horizon (10 years) we see as main challenge the wide proliferation of the technologies and vision proposed in this thesis to be highly dependent on user acceptance. While applications like adjustable reading UIs are already commercially being spread by companies, such as Spritz[29], a cognition-aware ecosystem of devices is much more privacy invading. Location sharing, for example, has been a highly controversial topic and continues to be so to this day, but the advantages are increasingly dominating the concerns: we share our location for efficient navigation, for area explorations, being able to meet up with friends, or being alerted about potential dangers in our surroundings. Cognition-aware systems face similar challenges and need to be proven useful first.

While we provide first investigations in applying such tools to learning and reading tasks, a fully developed system that comprehensively supports users with their information tasks throughout the day, could boost an entire society's productivity level. Beyond awareness of circadian rhythms of cognitive performance, research could address the question of how technologies can stage interventions through awareness and thus induce favorable cognitive states. Just like people drink a cup of coffee when they start to feel tired, technologies could detect fatigue and proactively help users to change their current state. This is similar to what meditation or mindfulness practices already propose in order to increase general focus or wind down after a long day of work. A study by Solberg *et al.* [246] has shown that meditation can help obtain physical and mental stability and boost the performance of athletes in shooting competitions. More recently, systems have been proposed that support users' mental preparedness [107] or induce creative states [1], but systems could go beyond single use cases and schedule a variety of interventions throughout the day to help users get to and remain in highly

[29] `http://spritzinc.com/`

productive states. Such potentially invasive systems are clearly controversial and hence a research plan would need to include assessments of long-term effects on their users including behavior changes and potential health impact. Chronically high levels of mental load have been shown to cause various health problems, such as stress, depression, and burnout [258]. Future research in this direction would need to look at the long-term feasibility and implications of such systems on people's levels of health, productivity, and comfort.

8.5 Final Remarks

When developing tools for augmenting cognitive processes the question comes up whether technology enhances our cognitive aptitude or helps to build and improve innate skills. The difference between these two notions becomes apparent when the tool is removed. Does the skill persist?

Clark and Chalmers [48] coined the term *active externalism* describing the active role our environment plays for our cognitive processes. They argue that cognition is not limited to the physical boundaries of our skull. External objects play a crucial role in cognitive processes, such as memory retrieval, linguistic processes, or skill acquisition. For example, we use our fingers to augment our working memory in calculations or use pen and paper to perform multiplications. While the brain is performing operations it delegates some of its workload to its external environment. Kirsh and Maglio [155] demonstrated how performing actions in the world can lead to quicker solutions of certain cognitive and perceptual problems than performing them mentally. They showed how physically rotating a shape by 90°, for example, could be done in about 100 ms, plus 200 ms required to press a respective button. In contrast, mentally rotating the shape took about 1000 ms. They distinguish *pragmatic actions*—where the world is altered because some physical change is desirable for its own sake—from *epistemic actions*, in which the world is altered to aid and augment cognitive processes, such as recognition or search, *i.e.*, to understand the world. In the rotating shape example that would mean a person gains knowledge about the world (does the shape fit an appropriate slot?) by pragmatic action (physically rotating the shape) quicker than through epistemic action (rotating the shape in the head). Hence, tools and technologies augment our cognitive processes and our understanding of the world. Obviously such knowledge gain remains even if the tool is removed. In other tasks, such as performing a tricky calculation, a calculator becomes coupled with the person as a tool. According to Clark and Chalmers [48] such coupling is considered an externally augmented cognitive process, which is sufficient if the tools (the relevant capacities) are generally available when they are needed: "*In effect, they are part of the basic package of cognitive resources that I bring to bear on the everyday world. These systems cannot be impugned simply on the basis of the danger of discrete damage, loss, or malfunction, or because of any occasional decoupling: the biological brain is in similar danger, and occasionally loses capacities temporarily in episodes of sleep, intoxication.*" [48] Hence, extended cognition is considered to be a core cognitive process, not an add-on extra.

The increasing spread of ubiquitous computing devices supports this notion. Devices are portable, ingrained in our everyday life and therefore near-constantly available. Hence, the functions and support they provide become part of our everyday cognitive processes including looking up information, noting something down, and sharing it. The notion of extended cognition is taken to the next level by some of the ideas and prototypes presented in this thesis. Context- and cognition-aware systems support the user in-situ according to current abilities and aptitudes. Proactive recommendations and content suggestions as described in Chapter 6 match content to the user's real-time processing capabilities and therefore facilitate information intake and processing. Seamless support may become so ingrained in our everyday life that the cognitive boost we receive through them becomes self-evident and may only be noticed when the tools err or fail. But since these technologies are built on people's innate circadian rhythm and productive phases, we could argue that even if the technology failed them, it might succeed in instilling productive habits. So even if the content was no longer pushed proactively, formed habits and awareness might cause users to actively seek out activities that match their current cognitive state. Similarly, for the adaptive reading, UIs presented, enhanced reading skill is applicable also for offline reading. The kinetic stimulus, for example, has the potential to school eye movements which endure even if the stimulus is absent. Also, one could expect that by using such tools, the overall reading volume is increased over time, which has been shown to lead to an increase in vocabulary and therefore advance word processing capabilities [250], and innate cognitive skill.

V

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] S. Abdullah, M. Czerwinski, G. Mark, and P. Johns. Shining (blue) light on creative ability. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 793–804, New York, NY, USA, 2016. ACM.

[2] S. Abdullah, E. L. Murnane, M. Matthews, M. Kay, J. A. Kientz, G. Gay, and T. Choudhury. Cognitive rhythms: unobtrusive and continuous sensing of alertness using a mobile phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 178–189. ACM, 2016.

[3] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. In *International Symposium on Handheld and Ubiquitous Computing*, pages 304–307. Springer, 1999.

[4] R. F. Adler and R. Benbunan-Fich. Self-interruptions in discretionary multitasking. *Computers in Human Behavior*, 29(4):1441–1449, 2013.

[5] R. Agarwal and E. Karahanna. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly*, pages 665–694, 2000.

[6] P. E. Agroudy, T. Machulla, R. Rzayev, T. Dingler, M. Funk, A. Schmidt, G. Ward, and S. Clinch. Impact of reviewing lifelogging photos on recalling episodic memories. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pages 1014–1019, New York, NY, USA, 2016. ACM.

[7] J. Allan. Automatic hypertext construction. Technical report, Cornell University, 1995.

[8] K. Angerbauer, T. Dingler, D. Kern, and A. Schmidt. Utilizing the effects of priming to facilitate text comprehension. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1043–1048, New York, NY, USA, 2015. ACM.

[9] D. L. Ashbrook. Enabling mobile microinteractions. 2010.

[10] D. Avrahami and S. E. Hudson. Responsiveness in instant messaging: predictive models supporting inter-personal communication. In *Proc. CHI '06*. ACM, 2006.

[11] D. Avrahami and S. E. Hudson. Responsiveness in instant messaging: predictive models supporting inter-personal communication. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 731–740. ACM, 2006.

[12] A. Baddeley, A. Eysenck, and M. Anderson. Memory, 2009.

[13] H. P. Bahrick and E. Phelphs. Retention of spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(2):344, 1987.

[14] N. Banovic, C. Brant, J. Mankoff, and A. Dey. Proactivetasks: The short of mobile device use sessions. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services*, MobileHCI '14, pages 243–252, New York, NY, USA, 2014. ACM.

[15] L. Barnard, J. S. Yi, J. A. Jacko, and A. Sears. An empirical comparison of use-in-motion evaluation scenarios for mobile computing devices. *International Journal of Human-Computer Studies*, 62(4):487–520, 2005.

[16] L. A. Barnett and S. W. Klitzing. Boredom in free time: Relationships with personality, affect, and motivation for different gender, racial and ethnic student groups. *Leisure Sciences*, 28(3):233–244, 2007.

[17] D. Basaraba, P. Yovanoff, J. Alonzo, and G. Tindal. Examining the structure of reading comprehension: do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing*, 26(3):349–379, 2013.

[18] E. B. Basoglu and O. Akdemir. A Comparison of Undergraduate Students' English Vocabulary Learning: Using Mobile Phones and Flash Cards. *Turkish Online Journal of Educational Technology - TOJET*, 9(3):1–7, July 2010.

[19] A. Battestini, V. Setlur, and T. Sohn. A large scale study of text-messaging use. In *Proc. MobileHCI '10*, pages 229–238. ACM, 2010.

[20] J. S. Beaudin, S. S. Intille, E. M. Tapia, R. Rockinson, and M. E. Morris. Context-sensitive microlearning of foreign language vocabulary on a mobile device. In *European Conference on Ambient Intelligence*, pages 55–72. Springer, 2007.

[21] J. B. Begole, J. C. Tang, R. B. Smith, and N. Yankelovich. Work rhythms: analyzing visualizations of awareness histories of distributed groups. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 334–343. ACM, 2002.

[22] C. BenMoussa. Workers on the move: New opportunities through mobile commerce. *Stockholm mobility roundtable*, pages 22–23, 2003.

[23] C. M. Bishop et al. Pattern recognition and machine learning, vol. 1springer. *New York*, (4):12, 2006.

[24] R. Bixler and S. D'Mello. Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In *Proc. IUI '13*. ACM, 2013.

[25] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proc. MM '14*. ACM, 2014.

[26] A. Bogomolov, B. Lepri, and F. Pianesi. Happiness recognition from mobile phone data. In *Proc SOCIALCOM '13*. IEEE Computer Society, 2013.

[27] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. Falling asleep with angry birds, facebook and kindle: A large scale study on mobile application usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 47–56, New York, NY, USA, 2011. ACM.

[28] A. A. Borbély. A two process model of sleep regulation. In *Human Neurobiology*, volume 1, pages 195–204. Springer-Verlag, 1982.

[29] S. A. Brasel and J. Gips. Enhancing television advertising: same-language subtitles can improve brand recall, verbal memory, and behavioral intent. *Journal of the Academy of Marketing Science*, 42(3):322–336, 2014.

[30] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[31] B. Brown, M. McGregor, and D. McMillan. 100 days of iphone use: Understanding the details of mobile device use. In *Proc MobileHCI '14*. ACM, 2014.

[32] P. A. Bruck, L. Motiwalla, and F. Foerster. Mobile Learning with Micro-content: A Framework and Evaluation. Bled, Slovenia, June 2012.

[33] A. Bulling and D. Roggen. Recognition of visual memory recall processes using eye movement analysis. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 455–464. ACM, 2011.

[34] A. Bulling, J. A. Ward, and H. Gellersen. Multimodal recognition of reading activity in transit using body-worn sensors. *ACM Transactions on Applied Perception (TAP)*, 9(1):2, 2012.

[35] A. Bulling, J. A. Ward, H. Gellersen, and G. Troster. Eye movement analysis for activity recognition using electrooculography. *IEEE transactions on pattern analysis and machine intelligence*, 33(4):741–753, 2011.

[36] A. Bulling and T. O. Zander. Cognition-aware computing. *IEEE Pervasive Computing*, 13(3):80–83, July 2014.

[37] K. Busemann and F. Tippelt. Second screen: Parallelnutzung von fernsehen und internet. *Media Perspektiven*, 7:408–416, 2014.

[38] C. J. Cai, P. J. Guo, J. Glass, and R. C. Miller. Wait-learning: Leveraging Conversational Dead Time for Second Language Education. CHI EA '14, pages 2239–2244, New York, NY, USA, 2014. ACM.

[39] S. K. Carpenter and E. L. DeLosh. Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, 19(5):619–636, 2005.

[40] R. P. Carver. *Sense and nonsense in speed reading*. Revrac Publications, 1971.

[41] C. Castillo, G. De Francisci Morales, and A. Shekhawat. Online matching of web content to closed captions in intonow. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 1115–1116, New York, NY, USA, 2013. ACM.

[42] N. Cavus and D. Ibrahim. m-Learning: An experiment in using SMS to support learning new English language words. *British Journal of Educational Technology*, 40(1):78–91, Jan. 2009.

[43] G. C. Cawley and N. L. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, 11:2079–2107, Aug. 2010.

[44] T. Chen and C. Guestrin. XGBoost: a scalable tree boosting system. In *Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, 2016.

[45] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[46] Y.-L. Chuang, C.-W. Liao, W.-S. Chen, W.-T. Chang, S.-H. Cheng, Y.-C. Zeng, and K.-H. Chan. Use second screen to enhance tv viewing experiences. In *Cross-Cultural Design. Methods, Practice, and Case Studies*, pages 366–374. Springer, 2013.

[47] K. Church and R. de Oliveira. What's up with whatsapp? comparing mobile instant messaging behaviors with traditional sms. In *Proc. MobileHCI '13*. ACM, 2013.

[48] A. Clark and D. Chalmers. The extended mind. *analysis*, 58(1):7–19, 1998.

[49] S. Cohrs. Sleep disturbances in patients with schizophrenia. *CNS drugs*, 22(11):939–962, 2008.

[50] S. Consolvo and M. Walker. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing*, 2(2):24–31, 2003.

[51] L. Cruickshank, E. Tsekleves, R. Whitham, A. Hill, and K. Kondo. Making interactive tv easier to use: Interface design for a second screen approach. *The Design Journal*, 10(3):41–53, 2007.

[52] M. Csikszentmihalyi. The flow experience and its significance for human psychology. 1988.

[53] A. Csomai and R. Mihalcea. Linking documents to encyclopedic knowledge. *Intelligent Systems, IEEE*, 23(5):34–41, 2008.

[54] W. L. Cull et al. Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, 14(3):215–235, 2000.

[55] E. Cutrell, M. Czerwinski, and E. Horvitz. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. 2001.

[56] M. Czerwinski, E. Horvitz, and S. Wilhite. A diary study of task switching and interruptions. In *Proc. CHI '04*. ACM, 2004.

[57] E. Dane. Paying attention to mindfulness and its effects on task performance in the workplace. *Journal of Management*, 2010.

[58] I. Darroch, J. Goodman, S. Brewster, and P. Gray. The effect of age and font size on reading text on handheld computers. In *IFIP Conference on Human-Computer Interaction*, pages 253–266. Springer, 2005.

[59] T. H. Davenport and J. C. Beck. *The attention economy: Understanding the new currency of business*. Harvard Business Press, 2013.

[60] O. de Bruijn and R. Spence. Rapid serial visual presentation: a space-time trade-off in information presentation. In *Advanced visual interfaces*, pages 189–192, 2000.

[61] D. Dearman and K. Truong. Evaluating the Implicit Acquisition of Second Language Vocabulary Using a Live Wallpaper. CHI '12, pages 1391–1400, New York, NY, USA, 2012. ACM.

[62] F. N. Dempster. Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, 79(2):162, 1987.

[63] A. K. Dey. Context-aware computing: The cyberdesk project. In *Proceedings of the AAAI 1998 Spring Symposium on Intelligent Environments*, pages 51–54, 1998.

[64] A. K. Dey. *Providing architectural support for building context-aware applications*. PhD thesis, Georgia Institute of Technology, 2000.

[65] A. K. Dey, K. Wac, D. Ferreira, K. Tassini, J.-H. Hong, and J. Ramos. Getting closer: an empirical investigation of the proximity of user to their smart phones. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 163–172. ACM, 2011.

[66] C. Dickie, R. Vertegaal, C. Sohn, and D. Cheng. eyelook: Using attention to facilitate mobile media consumption. In *Proc. of UIST '05*, pages 103–106, New York, NY, USA, 2005. ACM.

[67] D. F. Dinges and J. W. Powell. Microcomputer analyses of performance on a portable, simple visual rt task during sustained operations. *Behavior research methods, instruments, & computers*, 17(6):652–655, 1985.

[68] T. Dingler. Cognition-aware systems as mobile personal assistants. In ACM, editor, *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '15 Adjunct, New York, NY, USA, 2016. ACM.

[69] T. Dingler, P. E. Agroudy, H. V. Le, A. Schmidt, E. Niforatos, A. Bexheti, and M. Langheinrich. Multimedia memory cues for augmenting human memory. *IEEE MultiMedia*, 23(2):4–11, Apr 2016.

[70] T. Dingler, P. E. Agroudy, G. Matheis, and A. Schmidt. Reading-based screenshot summaries for supporting awareness of desktop activities. In *Proceedings of the 7th Augmented Human International Conference 2016*, AH '16, pages 27:1–27:5, New York, NY, USA, 2016. ACM.

[71] T. Dingler, T. Bagg, Y. Grau, N. Henze, and A. Schmidt. ucanvas: A web framework for spontaneous smartphone interaction with ubiquitous displays. In *Human-Computer Interaction–INTERACT 2015*, pages 402–409. Springer, 2015.

[72] T. Dingler, A. Bexheti, E. Niforatos, and F. Alt. Workshop on mobile cognition: Using mobile devices to enhance human cognition. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '15, pages 970–973, New York, NY, USA, 2015. ACM.

[73] T. Dingler and S. Brewster. Audiofeeds: A mobile auditory application for monitoring online activities. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1067–1070, New York, NY, USA, 2010. ACM.

[74] T. Dingler, M. Funk, and F. Alt. Interaction proxemics: Combining physical spaces for seamless gesture interaction. In *Proceedings of the 4th International Symposium on Pervasive Displays*, PerDis '15, pages 107–114, New York, NY, USA, 2015. ACM.

[75] T. Dingler, C. Giebler, U. Kunze, T. Würtele, N. Henze, and A. Schmidt. Memory displays: Investigating the effects of learning in the periphery. In *Proceedings of the 5th ACM International Symposium on Pervasive Displays*, PerDis '16, pages 118–123, New York, NY, USA, 2016. ACM.

[76] T. Dingler and N. Henze. That's the dog from my wedding – algorithms for memory shaping. In *Proceedings of the CHI Workshop on Designing Technology for Major Life Events, 2014*, 2014.

[77] T. Dingler, K. Kunze, E. Niforatos, C. Gurrin, I. Giannopolos, A. Dengel, and K. Kise. Wahm 2016: 3rd workshop on ubiquitous technologies for augmenting the human mind. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pages 1010–1013, New York, NY, USA, 2016. ACM.

[78] T. Dingler, J. Lindsay, and B. N. Walker. Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. *Proceedings of the 14th International Conference on Auditory Display, Paris, France*, pages 1–6, 2008.

[79] T. Dingler and M. Pielot. 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*, MobileHCI '15, pages 1–5, New York, NY, USA, 2015. ACM.

[80] T. Dingler, R. Rzayev, V. Schwind, and N. Henze. Rsvp on the go - implicit reading support on smart watches through eye tracking. In ACM, editor, *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '15, New York, NY, USA, 2016. ACM.

[81] T. Dingler, A. Sahami Shirazi, and N. Henze. There is more to wellbeing than health data – holistic lifelogging through memory capture. In *Proceedings of the CHI Workshop on Beyond Quantified Self: Data for Wellbeing, 2014*, 2014.

[82] T. Dingler and A. Schmidt. A real-time bidding platform for public displays. In *Proceedings of the 4th International Symposium on Pervasive Displays*, PerDis '15, pages 245–246, New York, NY, USA, 2015. ACM.

[83] T. Dingler and A. Schmidt. Peripheral displays to support human cognition. In *Peripheral Interaction*, pages 167–181. Springer International Publishing, 2016.

[84] T. Dingler, A. Schmidt, K. Kunze, M. Langheinrich, N. Davies, and N. Henze. Wahm 2014: workshop on ubiquitous technologies for augmenting the human mind. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1339–1345. ACM, 2014.

[85] T. Dingler, A. Schmidt, and T. Machulla. Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance. volume 1, pages 47:1–47:15, New York, NY, USA, Sept. 2017. ACM.

[86] T. Dingler, A. S. Shirazi, K. Kunze, and A. Schmidt. Assessment of stimuli for supporting speed reading on electronic devices. In *Proceedings of the 6th Augmented Human International Conference*, AH '15, pages 117–124, New York, NY, USA, 2015. ACM.

[87] T. Dingler, D. Weber, M. Pielot, J. Cooper, C.-C. Chang, and N. Henze. 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, pages 28:1–28:12, New York, NY, USA, 2017. ACM.

[88] I. DIS. 9241-210: 2010. ergonomics of human system interaction-part 210: Human-centred design for interactive systems. *International Standardization Organization (ISO). Switzerland*, 2009.

[89] J. D. Eastwood, A. Frischen, M. J. Fenske, and D. Smilek. The unengaged mind: Defining boredom in terms of attention. *Perspectives on Psychological Science*, 7(5):482–495, 2012.

[90] H. Ebbinghaus. *Memory: A contribution to experimental psychology*. Number 3. University Microfilms, 1913.

[91] D. Edge, S. Fitchett, M. Whitney, and J. Landay. MemReflex: Adaptive Flashcards for Mobile Microlearning. MobileHCI '12, pages 431–440, New York, NY, USA, 2012. ACM.

[92] D. Edge, E. Searle, K. Chiu, J. Zhao, and J. A. Landay. MicroMandarin: Mobile Language Learning in Context. CHI '11, pages 3169–3178, New York, NY, USA, 2011. ACM.

[93] E. Ehrlich. Opinions differ on speed reading. *NEA Journal*, 52:43–44, 1963.

[94] R. Farmer and N. D. Sundberg. Boredom proneness: The development and correlates of a new scale. *J. Pers. Asses*, 50(1):4–17, 1986.

[95] O. Fenichel. *On the psychology of boredom*. Columbia University Press, 1951.

[96] N. Fernandez Garcia, J. Arias Fisteus, and L. Sanchez Fernandez. Comparative evaluation of link-based approaches for candidate ranking in link-to-wikipedia systems. *Journal of Artificial Intelligence Research*, pages 733–773, 2014.

[97] D. Ferreira, J. Goncalves, V. Kostakos, L. Barkhuus, and A. K. Dey. 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. ACM, 2014.

[98] D. Ferreira, V. Kostakos, and A. K. Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2:6, 2015.

[99] J. E. Fischer, C. Greenhalgh, and S. Benford. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proc. MobileHCI '11*. ACM, 2011.

[100] C. D. Fisherl. Boredom at work: A neglected concept. *Human Relations*, 46(3):395–417, 1993.

[101] A. Fleury, J. S. Pedersen, M. Baunstrup, and L. B. Larsen. Interactive tv: Interaction and control in second-screen tv consumption. In *10th European Interactive TV Conference*, pages 104–107, 2012.

[102] J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang. Predicting human interruptibility with sensors. *ACM Trans. Comput.-Hum. Interact.*, 12(1):119–146, Mar 2005.

[103] K. I. Forster. Visual perception of rapidly presented word sequences of varying complexity. *Perception & Psychophysics*, 8(4):215–221, 1970.

[104] S. D. Frank. *Remember Everything You Read: The Evelyn Wood 7-Day Speed Reading & Learning Program*. HarperCollins, 1992.

[105] R. Fuller. Buckminster. critical path. *St, Martin^{1}s Press, New York*, 1981.

[106] M. Funk, T. Dingler, J. Cooper, and A. Schmidt. Stop helping me - i'm bored!: Why assembly assistance needs to be adaptive. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and*

Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers, UbiComp/ISWC '15 Adjunct, pages 1269–1273, New York, NY, USA, 2015. ACM.

[107] K. Futami, T. Terada, and M. Tsukamoto. Success imprinter: A method for controlling mental preparedness using psychological conditioned information. In *Proceedings of the 7th Augmented Human International Conference 2016*, page 11. ACM, 2016.

[108] D. Geerts, R. Leenheer, D. De Grooff, J. Negenman, and S. Heijstraten. In front of and behind the second screen: Viewer and producer perspectives on a companion app. In *Proceedings of the 2014 ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '14, pages 95–102, New York, NY, USA, 2014. ACM.

[109] C. Gentry. *A fully homomorphic encryption scheme*. PhD thesis, Stanford University, 2009.

[110] T. Georgiev. Investigation of the user's text reading speed on mobile devices. In *Proceedings of the 13th International Conference on Computer Systems and Technologies*, pages 329–336. ACM, 2012.

[111] R. Godwin-Jones. Emerging technologies from memory palaces to spacing algorithms: approaches to secondlanguage vocabulary learning. *Language, Learning & Technology*, 14(2):4, 2010.

[112] N. Goel, H. Van Dongen, and D. Dinges. Circadian rhythms in sleepiness, alertness, and performance. *Principles and practice of sleep medicine*, 5:445–55, 2011.

[113] T. Goetz, A. C. Frenzel, N. C. Hall, U. E. Nett, R. Pekrun, and A. A. Lipnevich. Types of boredom: An experience sampling approach. *Motivation and Emotion*, 38(3):401–419, 2014.

[114] E. B. Goldstein, M. Ritter, and G. Herbst. *Wahrnehmungspsychologie*, volume 2. Spektrum Akademischer Verlag Heidelberg, 2002.

[115] Google. The new multi-screen world:understanding cross-platform consumer behavior. 2012.

[116] R. G. Graf. Speed reading: Remember the tortoise. *Psychology Today*, 7(7):112–13, 1973.

[117] D. Green and J. Swets. Signal detection theory and psychophysics. 1966. *New York*, 888:889, 1966.

[118] B. M. Gross. *The managing of organizations: The administrative struggle*, volume 2. [New York]: Free Press of Glencoe, 1964.

[119] Q. Guo, E. Agichtein, C. L. A. Clarke, and A. Ashkan. In the mood to click? towards inferring receptiveness to search advertising. In *Proc. WI-IAT '09*. IEEE Computer Society, 2009.

[120] D. E. Hannula and C. Ranganath. The eyes have it: hippocampal activity predicts expression of memory in eye movements. *Neuron*, 63(5):592–599, 2009.

[121] D. M. Hansen. A discourse structure analysis of the comprehension of rapid readers. 1975.

[122] J. P. Hansen, F. Biermann, E. Møllenbach, H. Lund, J. S. Agustin, and S. Sztuk. A gazewatch prototype. In *Proc. of MobileHCI '15*, pages 615–621, New York, NY, USA, 2015. ACM.

[123] S. G. Hart. 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, 2006.

[124] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[125] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[126] A. Hayati and F. Mohmedi. The effect of films with and without subtitles on listening comprehension of efl learners. *British Journal of Educational Technology*, 42(1):181–192, 2011.

[127] J. Healey, L. Nachman, S. Subramanian, J. Shahabdeen, and M. Morris. Out of the lab and into the fray: towards modeling emotion in everyday life. In *International Conference on Pervasive Computing*, pages 156–173. Springer, 2010.

[128] B. Hedin and E. Lindgren. A comparison of presentation methods for reading on mobile phones. *Distributed Systems Online, IEEE*, 8(6):2–2, 2007.

[129] J. J. Heisz and D. I. Shore. More efficient scanning for familiar faces. *Journal of Vision*, 8(1):9–9, 2008.

[130] J. M. Hektner, J. A. Schmidt, and M. Csikszentmihalyi. *Experience sampling method: Measuring the quality of everyday life*. Sage, 2007.

[131] F. Herzberg, B. Mausner, and B. B. Snyderman. *The motivation to work*, volume 1. Transaction publishers, 2011.

[132] A. Hiniker, S. N. Patel, T. Kohno, and J. A. Kientz. Why would you do that? predicting the uses and gratifications behind smartphone-usage behaviors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 634–645, New York, NY, USA, 2016. ACM.

[133] D. Hintze, R. D. Findling, S. Scholz, and R. Mayrhofer. 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. ACM, 2014.

[134] Y. Hochberg and Y. Benjamini. Controlling the False Discovery Rate : a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[135] W. A. Hofstra and A. W. de Weerd. How to assess circadian rhythm in humans: a review of literature. *Epilepsy & Behavior*, 13(3):438–444, 2008.

[136] G. Holmes, A. Donkin, and I. H. Witten. Weka: A machine learning workbench. In *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 357–361. IEEE, 1994.

[137] E. Horvitz, J. Apacible, and M. Subramani. Balancing awareness and interruption: Investigation of notification deferral policies. In *User Modeling 2005*, pages 433–437. Springer, 2005.

[138] E. Horvitz, P. Koch, R. Sarin, J. Apacible, and M. Subramani1. Bayesphone: Precomputation of context-sensitive policies for inquiry and action in mobile devices. In *Proc UM '05*, 2005.

[139] G. Inc. Gartner says worldwide wearable devices sales to grow 18.4 percent in 2016. `http://www.gartner.com/newsroom/id/3198018`, 2016. Last accessed: November 16, 2017.

[140] S. Intille, C. Haynes, D. Maniar, A. Ponnada, and J. Manjourides. μema: Microinteraction-based ecological momentary assessment (ema) using a

smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1124–1128. ACM, 2016.

[141] S. T. Iqbal and B. P. Bailey. Effects of intelligent notification management on users and their tasks. In *Proc. CHI '08*. ACM, 2008.

[142] S. T. Iqbal and B. P. Bailey. Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Trans. Comput.-Hum. Interact.*, 17(4):15:1–15:28, Dec 2010.

[143] S. T. Iqbal and E. Horvitz. Disruption and recovery of computing tasks: field study, analysis, and directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 677–686. ACM, 2007.

[144] D. E. Irwin. Fixation location and fixation duration as indices of cognitive processing. *The interface of language, vision, and action: Eye movements and the visual world*, pages 105–134, 2004.

[145] S. Ishimaru, T. Dingler, K. Kunze, K. Kise, and A. Dengel. Reading interventions: Tracking reading state and designing interventions. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pages 1759–1764, New York, NY, USA, 2016. ACM.

[146] R. J. K. Jacob. What you look at is what you get: Eye movement-based interaction techniques. In *Proc. of CHI '90*, pages 11–18, New York, NY, USA, 1990. ACM.

[147] N. Janssen and H. A. Barber. Phrase frequency effects in language production. *PLoS ONE7*, 2012.

[148] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton. Measuring and defining the experience of immersion in games. *International journal of human-computer studies*, 66(9):641–661, 2008.

[149] M. A. Just and P. A. Carpenter. Eye fixations and cognitive processes. *Cognitive psychology*, 8(4):441–480, 1976.

[150] M. A. Just and P. A. Carpenter. *The psychology of reading and language comprehension.* Allyn & Bacon, 1987.

[151] M. Kassner, W. Patera, and A. Bulling. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Proc. of UbiComp '14 Adjunct*, New York, USA, 2014. ACM Press.

[152] M. Kay, K. Rector, S. Consolvo, B. Greenstein, J. O. Wobbrock, N. F. Watson, and J. A. Kientz. Pvt-touch: adapting a reaction time test for touchscreen devices. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 248–251. IEEE, 2013.

[153] D. Kern, P. Marshall, and A. Schmidt. Gazemarks: Gaze-based visual placeholders to ease attention switching. In *Proc. of CHI '10*, pages 2093–2102, New York, NY, USA, 2010. ACM.

[154] E. Keuleers, M. Brysbaert, and B. New. Subtlex-nl: A new measure for dutch word frequency based on film subtitles. *Behavior research methods*, 42(3):643–650, 2010.

[155] D. Kirsh and P. Maglio. On distinguishing epistemic from pragmatic action. *Cognitive science*, 18(4):513–549, 1994.

[156] N. Kleitman. Studies on the physiology of sleep. *American Journal of Physiology–Legacy Content*, 66(1):67–92, 1923.

[157] N. Kleitman. Studies on the physiology of sleep: Viii. diurnal variation in performance. *Am J Physiol*, 104:449–456, 1933.

[158] N. Kleitman, S. Titelbaum, and P. Feiveson. The effect of body temperature on reaction time. *American Journal of Physiology–Legacy Content*, 121(2):495–501, 1938.

[159] J. Knittel. Automatic Content Generation for Second Screen Apps Using Subtitles to Enhance the Learning Experience. published.

[160] J. Knittel and T. Dingler. 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, pages 93–103, New York, NY, USA, 2016. ACM.

[161] A. T. Korucu and A. Alkan. Differences between m-learning (mobile learning) and e-learning, basic terminology and usage of m-learning in education. *Procedia-Social and Behavioral Sciences*, 15:1925–1930, 2011.

[162] G. Kovacs. Smart subtitles for language learning. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 2719–2724, New York, NY, USA, 2013. ACM.

[163] L. Kreitzman and R. Foster. *The Rhythms Of Life: The Biological Clocks That Control the Daily Lives of Every Living Thing*. Profile books, 2011.

[164] T. Kubitza, N. Pohl, T. Dingler, and A. Schmidt. Webclip: a connector for ubiquitous physical input and output for touch screen devices. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 387–390. ACM, 2013.

[165] T. Kubitza, N. Pohl, T. Dingler, S. Schneegass, C. Weichel, and A. Schmidt. Ingredients for a new wave of ubicomp products. *IEEE Pervasive Computing*, (3):5–8, 2013.

[166] T. Kubitza, A. Schmidt, N. Pohl, D. Petrelli, T. Dingler, and N. Dulake. Tools and methods for creating interactive artifacts. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction*, pages 385–388. ACM, 2014.

[167] K. Kunze, T. Dingler, N. Henze, K. Kise, and Y. Sato. 2nd workshop on ubiquitous technologies to augment the human mind: Towards the knowledge log. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, UbiComp '15, pages 1213–1216, New York, NY, USA, 2015. ACM.

[168] K. Kunze, S. Sanchez, T. Dingler, O. Augereau, K. Kise, M. Inami, and T. Tsutomu. The augmented narrative: Toward estimating reader engagement. In *Proceedings of the 6th Augmented Human International Conference*, AH '15, pages 163–164, New York, NY, USA, 2015. ACM.

[169] M. Langheinrich. Privacy by design-principles of privacy-aware ubiquitous systems. In *International conference on Ubiquitous Computing*, pages 273–291. Springer, 2001.

[170] M. Langheinrich, A. Schmidt, N. Davies, and R. José. A practical framework for ethics: the pd-net approach to supporting ethics compliance in public display studies. In *Proceedings of the 2nd ACM International Symposium on Pervasive Displays*, pages 139–143. ACM, 2013.

[171] J. Lave. *Cognition in practice: Mind, mathematics and culture in everyday life*. Cambridge University Press, 1988.

[172] H. V. Le, S. Clinch, C. Sas, T. Dingler, N. Henze, and N. Davies. Impact of video summary viewing on episodic memory recall - design guidelines for video summarizations. In ACM, editor, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, volume 34 of *CHI '16*, New York, NY, USA, 2016. ACM.

[173] J. Lee, K. Lee, E. Jeong, J. Jo, and N. B. Shroff. Context-aware application scheduling in mobile systems: What will users do and not do next? In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 1235–1246, New York, NY, USA, 2016. ACM.

[174] A. Lenhart. Teens, smartphones & texting. Technical report, PewResearch Internet Project, 2012.

[175] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proc. MobiSys '13*. ACM, 2013.

[176] L. Lischke, T. Dingler, S. Schneegass, A. Schmidt, M. van der Vaart, and P. Wozniak. Parallel exhibits: Combining physical and virtual exhibits. In *NODEM 2014 CONFERENCE & EXPO ENGAGING SPACES*, 2014.

[177] S. P. Liversedge and J. M. Findlay. Saccadic eye movements and cognition. *Trends in cognitive sciences*, 4(1):6–14, 2000.

[178] C. Makris, Y. Plegas, and E. Theodoridis. Improved text annotation with wikipedia entities. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 288–295, New York, NY, USA, 2013. ACM.

[179] G. Mark, S. Iqbal, M. Czerwinski, and P. Johns. Capturing the mood: facebook and face-to-face encounters in the workplace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1082–1094. ACM, 2014.

[180] G. Mark, S. Iqbal, M. Czerwinski, and P. Johns. Focused, aroused, but so distractible: Temporal perspectives on multitasking and communications. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 903–916. ACM, 2015.

[181] G. Mark, S. Iqbal, M. Czerwinski, P. Johns, and A. Sano. Email duration, batching and self-interruption: Patterns of email use on productivity and stress. In *Proceedings of CHI*, 2016.

[182] G. Mark, S. T. Iqbal, M. Czerwinski, and P. Johns. Bored mondays and focused afternoons: The rhythm of attention and online activity in the workplace. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 3025–3034, New York, NY, USA, 2014. ACM.

[183] M. E. Masson. Conceptual processing of text during skimming and rapid sequential reading. *Memory & Cognition*, 11(3):262–274, 1983.

[184] A. Mathur, N. D. Lane, and F. Kawsar. Engagement-aware computing: modelling user engagement from mobile contexts. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 622–633. ACM, 2016.

[185] A. Matic, M. Pielot, and N. Oliver. Boredom-computer interaction: Boredom proneness and the use of smartphone. In *Proc. UbiComp '15*. ACM, 2015.

[186] E. Matin. Saccadic suppression: a review and an analysis. *Psychological bulletin*, 81(12):899, 1974.

[187] S. B. Merriam, R. S. Caffarella, and L. M. Baumgartner. *Learning in adulthood: A comprehensive guide*. John Wiley & Sons, 2012.

[188] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[189] H. Mitterer and J. M. McQueen. Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS one*, 4(11):e7785, 2009.

[190] T. H. Monk and J. Carrier. A parallelism between human body temperature and performance independent of the endogenous circadian pacemaker. *Journal of biological rhythms*, 13(2):113–122, 1998.

[191] A. Mottelson and K. Hornbæk. An affect detection technique using mobile commodity sensors in the wild. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 781–792. ACM, 2016.

[192] E. L. Murnane, S. Abdullah, M. Matthews, M. Kay, J. A. Kientz, T. Choudhury, G. Gay, and D. Cosley. Mobile manifestations of alertness: Connecting biological rhythms with patterns of smartphone app use. In *Proceedings of the 18th International Conference on Human-Computer Interaction with*

Mobile Devices and Services, MobileHCI '16, pages 465–477, New York, NY, USA, 2016. ACM.

[193] W. E. Nagy and P. A. Herman. Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. *The nature of vocabulary acquisition*, 19:35, 1987.

[194] T. Nakata. English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL*, 20(01):3–20, 2008.

[195] A. Nandakumar and J. Murray. Companion apps for long arc tv series: Supporting new viewers in complex storyworlds with tightly synchronized context-sensitive annotations. In *Proceedings of the 2014 ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '14, pages 3–10, New York, NY, USA, 2014. ACM.

[196] I. S. Nation. *Learning vocabulary in another language*. Ernst Klett Sprachen, 2001.

[197] T. Neate, M. Jones, M. Evans, and R. Bbc. Mediating Attention for Second Screen Companion Content. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3103–3106, New York, NY, USA, 2015. ACM.

[198] U. Neisser. *Cognitive psychology*. Appleton-Century-Crofts, New York, NY, USA, 1967.

[199] B. New, M. Brysbaert, J. Veronis, and C. Pallier. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(04):661–677, 2007.

[200] J. Nielsen. The digital consumer. *Online Resource by The Nielsen Company*, pages 1–28, 2014. Last accessed: November 16, 2017.

[201] A. Nijholt and D. Tan. Playing with your brain: brain-computer interfaces and games. In *Proceedings of the international conference on Advances in computer entertainment technology*, pages 305–306. ACM, 2007.

[202] B. O'Conaill and D. Frohlich. Timespace in the workplace: Dealing with interruptions. In *Conference companion on Human factors in computing systems*, pages 262–263. ACM, 1995.

[203] D. Odijk, E. Meij, and M. de Rijke. Feeding the second screen: Semantic linking based on subtitles. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 9–16, Paris, France, France, 2013. Le Centre de Hautes Etudes Internationales D'Iinformatique Documentaire.

[204] I. Oliveira and N. Guimarães. A tool for mental workload evaluation and adaptation. In *Proceedings of the 4th Augmented Human International Conference*, AH '13, pages 138–141, New York, NY, USA, 2013. ACM.

[205] D. R. Olson. Intelligence and literacy: The relationships between intelligence and the technologies of representation and communication. *Practical intelligence*, pages 338–360, 1986.

[206] A. Oulasvirta, T. Rattenbury, L. Ma, and E. Raita. Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing*, 16(1):105–114, 2012.

[207] A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti. Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 919–928, New York, NY, USA, 2005. ACM.

[208] H. E. Pashler. *Attention*. Psychology Press, 1998.

[209] V. Pejovic and M. Musolesi. Interruptme: Designing intelligent prompting mechanisms for pervasive applications. In *Proc. UbiComp '14*. ACM, 2014.

[210] T. Piazza, M. Fjeld, G. Ramos, A. Yantac, and S. Zhao. Holy smartphones and tablets, batman!: mobile interaction's dynamic duo. In *Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction*, pages 63–72. ACM, 2013.

[211] M. Pielot. Large-scale evaluation of call availability prediction. In *Proc. UbiComp '14*, 2014.

[212] M. Pielot, K. Church, and R. de Oliveira. An in-situ study of mobile phone notifications. In *Proc. MobileHCI '14*, 2014.

[213] M. Pielot, R. de Oliveira, H. Kwak, and N. Oliver. Didn't you see my message? predicting attentiveness to mobile instant messages. In *Proc. CHI '14*. ACM, 2014.

[214] M. Pielot, T. Dingler, J. S. Pedro, and N. Oliver. 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, pages 825–836, New York, NY, USA, 2015. ACM.

[215] M. Pielot, H. Kwak, R. de Oliveira, and N. Oliver. Didn't you see my message? predicting attentiveness in mobile instant messaging. In *Proc. CHI '14*, 2014.

[216] B. Poppinga, W. Heuten, and S. Boll. Sensor-based identification of opportune moments for triggering notifications. *Pervasive Computing*, 13:22–29, 2014.

[217] M. C. Potter and E. I. Levy. Recognition memory for a rapid sequence of pictures. *Journal of experimental psychology*, 81(1):10, 1969.

[218] E. Pronin, E. Jacobs, and D. M. Wegner. Psychological effects of thought acceleration. *Emotion*, 8(5):597, 2008.

[219] Z. W. Pylyshyn and R. W. Storm. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial vision*, 3(3):179–197, 1988.

[220] E. Quinn and I. S. P. Nation. *Speed reading: A course for learners of English*. Oxford University Press, 1974.

[221] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman. Activity recognition from accelerometer data. In *AAAI*, volume 5, pages 1541–1546, 2005.

[222] J. E. Raymond, K. L. Shapiro, and K. M. Arnell. Temporary suppression of visual processing in an rsvp task: An attentional blink? *Journal of Experimental Psychology: Human perception and performance*, 18(3):849, 1992.

[223] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

[224] K. Rayner, E. R. Schotter, M. E. Masson, M. C. Potter, and R. Treiman. So much to read, so little time how do we read, and can speed reading help? *Psychological Science in the Public Interest*, 17(1):4–34, 2016.

[225] J. Read. *Assessing vocabulary*. Cambridge university press Cambridge, 2000.

[226] J. L. Redondoio Garcia, L. De Vocht, R. Troncy, E. Mannens, and R. Van de Walle. Describing and contextualizing events in tv news show. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 759–764, New York, NY, USA, 2014. ACM.

[227] E. D. Reichle, K. Rayner, and A. Pollatsek. The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(04):445–476, 2003.

[228] R. Revlin. *Cognition: Theory and practice*. Palgrave Macmillan, 2012.

[229] S. Robertson, C. Wharton, C. Ashworth, and M. Franzke. Dual device user interface design: Pdas and interactive television. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, pages 79–86, New York, NY, USA, 1996. ACM.

[230] S. Rosenthal, A. K. Dey, and M. Veloso. Using decision-theoretic experience sampling to build personalized mobile phone interruption models. In *Proc. Pervasive '11*. Springer-Verlag, 2011.

[231] A. Sahami Shirazi, N. Henze, T. Dingler, K. Kunze, and A. Schmidt. Upright or sideways?: Analysis of smartphone postures in the wild. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, pages 362–371. ACM, 2013.

[232] A. Sahami Shirazi, N. Henze, T. Dingler, M. Pielot, D. Weber, and A. Schmidt. Large-scale assessment of mobile notifications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 3055–3064, New York, NY, USA, 2014. ACM.

[233] S. Sanchez, T. Dingler, H. Gu, and K. Kunze. Embodied reading: A multisensory experience. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 1459–1466, New York, NY, USA, 2016. ACM.

[234] D. L. Schacter. The seven sins of memory: insights from psychology and cognitive neuroscience. *American psychologist*, 54(3):182, 1999.

[235] W. B. Schaufeli, M. Salanova, V. González-Romá, and A. B. Bakker. The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness studies*, 3(1):71–92, 2002.

[236] B. Schildbach and E. Rukzio. Investigating selection and reading performance on a mobile phone while walking. In *Proc. of MobileHCI '10*, pages 93–102, New York, NY, USA, 2010. ACM.

[237] R. Schleicher, N. Galley, S. Briest, and L. Galley. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51(7):982–1010, 2008.

[238] A. Schmidt. *Ubiquitous computing-computing in context*. PhD thesis, Lancaster University, 2003.

[239] A. Schmidt, M. Beigl, and H.-W. Gellersen. There is more to context than location. *Computers & Graphics*, 23(6):893 – 901, 1999.

[240] S. Schneegass, B. Pfleging, T. Dingler, and A. Schmidt. Interaction spaces: Interactive spatial areas to control smart environments. In *Mensch & Computer*, pages 333–336, 2013.

[241] E. R. Schotter, R. Tran, and K. Rayner. Don't believe what you read (only once) comprehension is supported by regressions during reading. *Psychological science*, page 0956797614531148, 2014.

[242] A. Sellen, Y. Rogers, R. Harper, and T. Rodden. Reflecting human values in the digital age. *Communications of the ACM*, 52(3):58–66, 2009.

[243] G. D. M. Serugendo, A. A. V. Castro, D. Konstantas, L. Liang, D. Zheng-yu, G. Thyagaraju, U. P. Kulkarni, A. R. Yardi, G. H. Cánepa, A. J. Molina, et al. Activity-based computing. *IEEE Pervasive Computing*, 7(2):58–61, 2008.

[244] W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460, 2015.

[245] A. Smith and J. L. Boyles. The rise of the "connected viewer". *Pew Internet & American Life Project*, 2012.

[246] E. Solberg, K. Berglund, O. Engen, O. Ekeberg, and M. Loeb. The effect of meditation on shooting performance. *British journal of sports medicine*, 30(4):342–346, 1996.

[247] C. Speier, J. S. Valacich, and I. Vessey. The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences*, 30(2):337–360, 1999.

[248] R. Spence. Rapid, serial and visual: a presentation technique with potential. *Information visualization*, 1(1):13–19, 2002.

[249] L. R. Squire. Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of cognitive neuroscience*, 4(3):232–243, 1992.

[250] K. E. Stanovich and A. E. Cunningham. What reading does for the mind. *Journal of Direct Instructions*, 1(2):137–149, 2001.

[251] E. Stuyven, K. Van der Goten, A. Vandierendonck, K. Claeys, and L. Crevits. The effect of cognitive load on saccadic eye movements. *Acta psychologica*, 104(1):69–85, 2000.

[252] A. Trusty and K. N. Truong. Augmenting the web for second language vocabulary learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 3179–3188, New York, NY, USA, 2011. ACM.

[253] E. Tulving. Episodic and semantic memory 1. *Organization of Memory. London: Academic*, 381(4), 1972.

[254] T. Tuomi, C. L. Nagorny, P. Singh, H. Bennet, Q. Yu, I. Alenkvist, B. Isomaa, B. Östman, J. Söderström, A.-K. Pesonen, et al. Increased melatonin signaling is a risk factor for type 2 diabetes. *Cell metabolism*, 2016.

[255] W. R. Uttal and P. Smith. Recognition of alphabetic characters during voluntary eye movements. *Perception & Psychophysics*, 3(4):257–264, 1968.

[256] K. Vadas, N. Patel, K. Lyons, T. Starner, and J. Jacko. Reading on-the-go: a comparison of audio and hand-held displays. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pages 219–226. ACM, 2006.

[257] A. Van Cauwenberge, G. Schaap, and R. Van Roy. "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, 2014.

[258] G. van Daalen, T. M. Willemsen, K. Sanders, and M. J. van Veldhoven. Emotional exhaustion and mental health problems among employees doing "people work": The impact of job demands, job resources and family-to-work conflict. *International archives of occupational and environmental health*, 82(3):291–303, 2009.

[259] H. P. Van Dongen and D. F. Dinges. Circadian rhythms in fatigue, alertness, and performance. *Principles and practice of sleep medicine*, 20:391–9, 2000.

[260] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[261] B. M. Velichkovsky and J. P. Hansen. New technological windows into mind: There is more in eyes and brains for human-computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 496–503. ACM, 1996.

[262] M. Virvou and E. Alepis. Mobile educational features in authoring tools for personalised tutoring. *Computers & Education*, 44(1):53–68, 2005.

[263] S. J. Vodanovich. On the possible benefits of boredom: A neglected area in personality research. *Psychology and Education: An Interdisciplinary Journal*, 2003.

[264] B. N. Walker, J. Lindsay, A. Nance, Y. Nakano, D. K. Palladino, T. Dingler, and M. Jeon. Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 55(1):157–182, 2013.

[265] R. Want, A. Hopper, V. Falcao, and J. Gibbons. The active badge location system. *ACM Transactions on Information Systems (TOIS)*, 10(1):91–102, 1992.

[266] S. Webb. The effects of repetition on vocabulary knowledge. *Applied linguistics*, 28(1):46–65, 2007.

[267] D. Weber, A. Voit, T. Dingler, M. Kallert, and N. Henze. Assessment of an unobtrusive persuasive system for behavior change in home environments. In *Proceedings of the 5th ACM International Symposium on Pervasive Displays*, PerDis '16, pages 245–246, New York, NY, USA, 2016. ACM.

[268] M. Weiser. The computer for the 21st century. *Scientific american*, 265(3):94–104, 1991.

[269] M. Weiser and J. S. Brown. The coming age of calm technology. In *Beyond calculation*, pages 75–85. Springer, 1997.

[270] E. Weissinger, L. L. Caldwell, and D. L. Bandalos. Relation between intrinsic motivation and boredom in leisure time. *Leisure Sciences*, 14(4):317–325, 1992.

[271] K. Westhoff and C. Hagemeister. *Konzentrationsdiagnostik*. Pabst Lengerich, 2005.

[272] R. A. Wever. *The circadian system of man: results of experiments under temporal isolation*. Springer Science & Business Media, 2013.

[273] C. D. Wickens. Processing resources and attention. *Multiple-task performance*, 1991:3–34, 1991.

[274] J. William. *Principles of psychology*. Hardpress Limited, 2013.

[275] A. L. Wilson. The promise of situated cognition. *New directions for adult and continuing education*, 1993(57):71–79, 1993.

[276] T. D. Wilson, D. A. Reinhard, E. C. Westgate, D. T. Gilbert, N. Ellerbeck, C. Hahn, C. L. Brown, and A. Shaked. Just think: The challenges of the disengaged mind. *Science*, 345(6192):75–77, 2014.

[277] K. Wolf, Y. Abdelrahman, D. Schmid, T. Dingler, and A. Schmidt. Effects of camera position and media type on lifelogging images. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, MUM '15, pages 234–244, New York, NY, USA, 2015. ACM.

[278] K. Wolf, S. Schneegass, N. Henze, D. Weber, V. Schwind, P. Knierim, S. Mayer, T. Dingler, Y. Abdelrahman, T. Kubitza, M. Funk, A. Mebus, and A. Schmidt. Tuis in the large: Using paper tangibles with mobile devices. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1579–1584, New York, NY, USA, 2015. ACM.

[279] W. Wood and D. T. Neal. A new look at habits and the habit-goal interface. *Psychological review*, 114(4):843, 2007.

[280] H.-Y. Wu, M. Rubinstein, E. Shih, J. V. Guttag, F. Durand, and W. T. Freeman. Eulerian video magnification for revealing subtle changes in the world. 2012.

[281] S. Yoon, S.-s. Lee, J.-m. Lee, and K. Lee. Understanding notification stress of smartphone messenger app. In *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14*, CHI EA '14, pages 1735–1740, New York, NY, USA, 2014. ACM.

VI

Appendix

Additional User Study Documents

- Consent form used for the controlled *QuickLearn* study (Section 6.2).
- An example RECALL Ethics Worksheet for the controlled *QuickLearn* study (Section 6.2) following the ethics process [170], which we applied for all user studies under RECALL.
- Consecutive screenshots of the kinetic stimulus as it was used in our user study(Section 7.2).

Universität Stuttgart

Study: QuickLearn
Participant ID: _______

Informed Consent

Thank you for your participation in our study.

Please read the following information carefully. A copy is available through the Quick Spanish application for future reference.

Experiment:
Promoting Language-Learning Through Smartphone Application
Identifying opportune moments for language learning.

Description:
Volunteering in this two-week user study involves using a language learning application on your Android smartphone. The study is dedicated to understanding optimal moments and different interface designs for language acquisition. For the two-week duration of the study, various aspects of your phone usage is being processed. We guarantee that your data will be stored securely and anonymously.

Participation is completely voluntary and you are free to drop out of this study at any point in time.

Compensation:
You will be compensated with 5 € at each of the three visits during the study. An additional 5 € will be given at the final visit if you attend all three sessions resulting in a total of 20€.

Experimenters:
Tilman Dingler, Universität Stuttgart, tilman.dingler@vis.uni-stuttgart.de
Chun-Cheng Chang, University of Washington, changcc@uw.edu
Jennifer Cooper, University of Minnesota, cooperj@umn.edu

Please do not hesitate to contact us if you have questions or concerns.

Thanks!

Study: QuickLearn

Participant ID: _______

Universität Stuttgart

Consent Form

❏ I have read and understood the information above.

❏ I have understood the purpose of this study and I agree to participate.

❏ I have understood that I can cancel my participation at any point in time.

Participant ID ________________________ (filled in by the experimenter)

Signatures

Participant __________________________ Date ______________

Experimenter __________________________ Date ______________

RECALL Ethics Worksheet

This worksheet documents a particular experiment within RECALL that is covered by one or more *Study Process Templates (SPTs)*. It is used for documentation purposes and forms part of the corresponding deliverable in which this experiment took place. It helps researchers within RECALL to ensure that all personal information collected and processed in RECALL is treated in accordance with the project's ethical guidelines and the feedback from its ethical advisory board (EAB).

Study Information

4.1 Study Title

Give a concise title to your study that describes the particular problem you are investigating in your experiment and/or field study.

Detecting Opportune Moments of Learning through Language Acquisition Application

4.1 Brief Description

Describe your experiment in a few sentences (no more than 3-4 sentences). The description should include the problem you are trying to address and the methods you are planning to use.

We investigate whether a previously established boredom detection algorithm can be leveraged to promote productivity on smartphones (e.g., language learning) during periods of predicted boredom. Additional phone usage will be logged to determine whether factors outside of boredom lead to opportune moments of learning. Users of the app will be asked to engage in a language-learning task that teaches foreign language words in two methods; flashcards and multiple-choice. We will measure the number of words that users learn at the close of each week during a controlled two-week study and additionally release the app on GooglePlay for a broader audience.

4.2 Planned Duration

How long (start, finish, duration) do you plan to run these experiments and/or field studies for?

10.09.15 – Sep.'16

4.3 Work Package

What RECALL work package(s) does this work fall in?

WP 2, WP 4, WP 5

Project Staff

4.4 Principle Researcher, Institution

Who are the principle researchers responsible for this experiment/study (incl. institution)?

Tilman Dingler (University of Stuttgart)

4.5 Other Staff (Project Members)

Name all personnel involved in this set of experiments/field studies. This may also include project members from other institutions,

Dominik Weber (University of Stuttgart)
Niels Henze (University of Stuttgart)

4.6 External Staff

Name all personnel involved in this set of experiments/field studies. This may also include project members from other institutions, as well as external researchers.

Chun-Cheng Chang (University of Washington)
Jennifer Cooper (University of Minnesota)

Aims and Methods

4.7 Goals and Research Questions

What are the goals of this research effort? What research questions do you hope to be able to answer with this set of experiments? Please be as specific as possible.

The goal of the study is to evaluate the vocab learning gains made in each of the language learning application delivery methods (i.e., flashcard, multiple choice) during different timing conditions (i.e., boredom detection, random). The additional condition of user-initiated usage will be compared to an existing boredom detection algorithm in an attempt to uncover additional measures of opportune moments for microlearning based on when the user chooses to engage in the language-learning application. We hope to answer the question of whether periods of boredom lead to learning opportunities or if there are other conditions that lead to these opportunities. These questions will be answered in two ways, one by pushing notifications to users during periods of predicted boredom, we can see whether they participate in learning at higher rate than during other random times; and two, log data from phone usage should reveal patterns that will dictate the best times to provide opportunities for productivity for the user.

4.8 Envisioned Methods (Planned Experiments and Studies)

Describe the set of experiments and/or field studies that you are planning to conduct within the scope of this research effort. If there is an order to the experiments, enumerate them in order. Otherwise list them as bullet points in no particular order. Be as specific as possible (e.g. instead of "interviews" write "online interviews through SurveyMonkey" or "Student focus groups on campus"…

System-Initiated: Repeated-measures design, where each student will be exposed to each of the four independent variables:

1) Boredom-detected: flashcard
2) Boredom-detected: multiple-choice
3) Random: flashcard
4) Random: multiple-choice

User-initiated: Correlational analyses will be conducted and the results will contribute to the "ground truth" for opportune moments of learning.

4.9 Study Type and ID

Select the type of study and if appropriate enter a reference to the guidelines used. If the experiment does not fit into one of the existing RECALL categories select "Other" and contact the coordinator in order to verify if additional EAB input is needed.

Document ID: Click here to enter text. (see RECALL Wiki)

☐ Lifelogging Experiment (see **guidelines**)

☒ Controlled Volunteer Studies (see **guidelines**)

☐ Public Trials (see **guidelines**)

☐ Others (describe and contact coordinator in case additional ethics advice req.):

Click here to enter text.

4.10 Adequacy of Methods

Briefly explain why the experimental methods indicated above are adequate for the research questions

- A Controlled study because we want to manipulate certain variables while simultaneously measuring the learning gains during specified time periods
- An in-the-wild study to collect feedback from a broader user base necessary to train a new machine-learning algorithm for classifying opportune moments for learning

4.11 Subjects and Recruitment Process (if any)

Who are the subjects of your research, and how are you planning to recruit them? Be as specific as possible. Note that you might not require recruitment if you are simply planning observations in public.

- The experimenters personally recruit subjects and release the app on GooglePlay

4.12 **Data to be Collected**

What information are you collecting in your experiments? Try to be as inclusive and specific as possible, listing all potential data types that you might be interested in collecting. If you are not collecting identifiable information from your subjects, state this here, but still list the (anonymous) data that you are planning to collect.

- Quiz scores
- General log data from cell phone
- Subjective feelings (survey/interview)
- Demographics

Risks and Precautions

4.13 **Potential Risks**

Try to envision the risks that your data collection might pose to your data subjects. What if the data you collected would wind up on the Internet, together with your subjects' real name and contact info? Could they suffer problems at work if their employers would find this information? What if hackers would be able to break into your system and steal your data collection? Would they be able to commit criminal acts with this information?

- Limited risk for participants since we gather mostly uncritical data
- Data collection is anonymized

4.14 **Planned Precautions**

Given the risks you identified in question 4.13, what precautions will be taken by you and your team in order to prevent data leakage? Try to be as specific as possible. You can also refer to your description of data storage in the next section.

- Anonymized data collection.
- Double consent form: 1) in-person consent, 2) in-app consent

Data Storage and Processing

4.15 **Storage Locations**

Describe where each of the data enumerated under 3.6 will be stored. Also consider potential backup processes.

- Local file servers
- Regular backups

4.16 **Access Control**

Explain how access to the data is regulated. Try to be as specific as possible. Use the RECALL Guide to Secure Storage Document for guidance, but make sure to describe your actual implementation of these guidelines here.

- Only researchers directly involved in the project have access to data

4.17 **Data Processing**

Describe how the collected data will be used. What kind of statistics will be assembled, what kind of qualitative information extracted, what kind of information combined?

- Classic data evaluation of qualitative key performance indicators.

4.18 **Data Anonymization**

If some or all of your data will be anonymized or pseudonymized, explain how you do this. What algorithms will be used for the anonymization, and what guarantees do they offer? How are pseudonymous identifiers generated and where is lookup information (if any) for those pseundonymizers kept?

- each participant is assigned an anonymized hash number (Details see: http://tilmanification.org/quicklearn.html)

4.19 **Data Retention**

How long will you keep personally identifiable information (PII)?

☒ No PII will be kept
☐ 6 months after study finishes
☐ 6 months after results have been first published*
☐ 6 months after RECALL finishes
☐ Others (describe): Click here to enter text.

**Or within 6 months after RECALL finishes, if results are not published until end of the project*

Ethical Checklist

Use the checklist below for a quick overview of the ethical issues in your planned experiment.

	Aspect	Yes	No
1.	**Informed Consent Form Needed?** *If yes, attach. If no, justify on attached sheet.*	☒	☐
2.	**Deception Used?** *If yes, justify on attached sheet.*	☐	☒
3.	**Private Information collected?** *(c.f. to 4.13. for list of data items)*	☒	☐
4.	**Subjects Remunerated?** *If yes, write amount here:*	☐	☒
5.	**Involvement of Children, Patients, People with Cognitive Disorders?** *If you answer yes to this question, you must explicitly consult with the EAB!*	☐	☒
6.	**Internal IRB Review Needed?** *If yes, attach feedback after obtained*	☐	☒

Worksheet Versioning

Provide date, authorship, and changes made *to this specific worksheet* in the history table below.

Date	Author	Comment
12.07.15	Jennifer Cooper	v0.1
14.07.15	Tilman Dingler	v0.2
10.09.15	Tilman Dingler	v.0.3

RECALL February 6th 2014 B3. Ethical Issues

ETHICAL ISSUES TABLE

	YES	PAGE
Informed Consent		
1. Does the proposal involve children?	NO	
2. Does the proposal involve patients?	NO	
3. Does the proposal involve persons not able to give consent?	NO	
4. Does the proposal involve adult healthy volunteers?	YES	
Biological research		
5. Does the proposal involve Human Genetic Material?	NO	
6. Does the proposal involve Human biological samples?	NO	
7. Does the proposal involve Human biological data collection?	NO	
8. Does the proposal involve Human Embryos?	NO	
9. Does the proposal involve Human Foetal Tissue or Cells?	NO	
10. Does the proposal involve Human Embryonic Stem Cells?	NO	
Privacy		
11. Does the proposal involve processing of genetic information or personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)	NO	
12. Does the proposal involve tracking the location or observation of people without their knowledge?	NO	
Research on Animals		
13. Does the proposal involve research on animals?	NO	
14. Are those animals transgenic small laboratory animals?	NO	
15. Are those animals transgenic farm animals?	NO	
16. Are those animals cloned farm animals?	NO	
17. Are those animals non-human primates?	NO	
Research Involving Developing Countries		
18. Is any part of the research carried out in countries outside of the European Union and FP7 Associated states?	NO	
Dual Use		
19. Does the research have direct military application	NO	
20. Does the research have the potential for terrorist abuse	NO	
ICT Implants		
21. Does the proposal involve clinical trials of ICT implants?	NO	
(IF NONE) I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL		

hcilab | reading lab

I realized recently that what one thinks about in the shower is more important than I'd thought. I knew it was a good time to have ideas. Now I'd say it's hard to do a really good job on anything you don't think about in the shower. Everyone who's worked on difficult problems is probably familiar with the phenomenon of working hard to figure something out, failing, and then suddenly seeing the answer a bit later while doing something else. There's a kind of thinking you do without trying to. I'm increasingly convinced this type of thinking is not merely helpful in solving hard problems, but necessary. The tricky part is, you can only control it indirectly. Most people have one top idea in their mind at any given time. That's the idea their thoughts will drift toward when they're allowed to drift freely. And this idea will thus tend to get all the benefit of that type of thinking, while others are starved of it. Which means it's a disaster to let the wrong idea become the top one in your mind. You can't directly control where your thoughts drift. If you're controlling them, they're not drifting. But you can control them indirectly, by controlling what situations you let yourself get into. That has been the lesson for me: be careful what you let become critical to you. Try to get yourself into situations where the most urgent problems are ones you want to think about. You don't have complete control, of course. An emergency could push other thoughts out of your head. But barring emergencies you have a good deal of indirect control over what becomes the top idea in your mind.

© hcilab 2014

hcilab | reading lab

I realized recently that what one thinks about in the shower is more important than I'd thought. I knew it was a good time to have ideas. Now I'd say it's hard to do a really good job on anything you don't think about in the shower. Everyone who's worked on difficult problems is probably familiar with the phenomenon of working hard to figure something out, failing, and then suddenly seeing the answer a bit later while doing something else. There's a kind of thinking you do without trying to. I'm increasingly convinced this type of thinking is not merely helpful in solving hard problems, but necessary. The tricky part is, you can only control it indirectly. Most people have one top idea in their mind at any given time. That's the idea their thoughts will drift toward when they're allowed to drift freely. And this idea will thus tend to get all the benefit of that type of thinking, while others are starved of it. Which means it's a disaster to let the wrong idea become the top one in your mind. You can't directly control where your thoughts drift. If you're controlling them, they're not drifting. But you can control them indirectly, by controlling what situations you let yourself get into. That has been the lesson for me: be careful what you let become critical to you. Try to get yourself into situations where the most urgent problems are ones you want to think about. You don't have complete control, of course. An emergency could push other thoughts out of your head. But barring emergencies you have a good deal of indirect control over what becomes the top idea in your mind.

© hcilab 2014

hcilab | reading lab

I realized recently that what one thinks about in the shower is more important than I'd thought. I knew it was a good time to have ideas. Now I'd say it's hard to do a really good job on anything you don't think about in the shower. Everyone who's worked on difficult problems is probably familiar with the phenomenon of working hard to figure something out, failing, and then suddenly seeing the answer a bit later while doing something else. There's a kind of thinking you do without trying to. I'm increasingly convinced this type of thinking is not merely helpful in solving hard problems, but necessary. The tricky part is, you can only control it indirectly. Most people have one top idea in their mind at any given time. That's the idea their thoughts will drift toward when they're allowed to drift freely. And this idea will thus tend to get all the benefit of that type of thinking, while others are starved of it. Which means it's a disaster to let the wrong idea become the top one in your mind. You can't directly control where your thoughts drift. If you're controlling them, they're not drifting. But you can control them indirectly, by controlling what situations you let yourself get into. That has been the lesson for me: be careful what you let become critical to you. Try to get yourself into situations where the most urgent problems are ones you want to think about. You don't have complete control, of course. An emergency could push other thoughts out of your head. But barring emergencies you have a good deal of indirect control over what becomes the top idea in your mind.

© hcilab 2014

Tilman Dingler

Cognition-Aware Systems To Support Information Intake And Learning

Knowledge is created at an ever-increasing pace putting us under constant pressure to consume and acquire new information. Information gain and learning, however, require time and mental resources. While the proliferation of ubiquitous computing devices, such as smartphones, enables us to consume information anytime and anywhere, technologies are often disruptive rather than sensitive to the current user context. While people exhibit different levels of concentration and cognitive capacity throughout the day, applications rarely take these performance variations into account and often overburden their users with information or fail to stimulate.

This work investigates how technology can be used to help people effectively deal with information intake and learning tasks through cognitive context-awareness. By harvesting sensor and usage data from mobile devices, we obtain people's levels of attentiveness, receptiveness, and cognitive performance. We subsequently use this cognition-awareness in applications to help users process information more effectively.

Through a series of lab studies, online surveys, and field experiments we follow six research questions to investigate how to build cognition-aware systems. Awareness of user's variations in levels of attention, receptiveness, and cognitive performance allows systems to trigger appropriate content suggestions, manage user interruptions, and adapt User Interfaces in real-time to match tasks to the user's cognitive capacities. The tools, insights, and concepts described in this book allow researchers and application designers to build systems with an awareness of momentary user states and general circadian rhythms of alertness and cognitive performance.

Universität
Stuttgart