Automatic Text Simplification Tools to Provide Reading Assistance for Deaf and Hard-of-Hearing Individuals

Oliver Alonzo
oa7652@rit.edu

Follow this and additional works at: https://repository.rit.edu/theses

Recommended Citation

This Dissertation is brought to you for free and open access by the RIT Libraries. For more information, please contact repository@rit.edu.
Automatic Text Simplification Tools to Provide Reading Assistance for Deaf and Hard-of-Hearing Individuals

by

Oliver Alonzo

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Computing and Information Sciences

B. Thomas Golisano College of Computing and Information Sciences

Rochester Institute of Technology
Rochester, New York
April, 2023
Automatic Text Simplification Tools to Provide Reading Assistance for Deaf and Hard-of-Hearing Individuals

by
Oliver Alonzo

Committee Approval:

We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in this dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computing and Information Sciences.

Dr. Matt Huenerfauth  Date
Dissertation Advisor

Dr. Kristen Shinohara  Date
Dissertation Committee Member

Dr. Garreth Tigwell  Date
Dissertation Committee Member

Dr. Raja Kushalmagar  Date
Dissertation Committee Member

Dr. Sharon Mason  Date
Dissertation Defense Chairperson

Certified by:

Dr. Pengcheng Shi  Date
Ph.D. Program Director, Computing and Information Sciences
Automatic Text Simplification Tools to Provide Reading Assistance for Deaf and Hard-of-Hearing Individuals

by

Oliver Alonzo

Submitted to the
B. Thomas Golisano College of Computing and Information Sciences Ph.D. Program in Computing and Information Sciences in partial fulfillment of the requirements for the
Doctor of Philosophy Degree at the Rochester Institute of Technology

Abstract

Automatic Text Simplification (ATS) consists of computing techniques to rewrite text to make it simpler to read and understand, and prior research has explored its use to provide reading assistance to various user groups who may benefit from reading support. Prior literacy research has identified great diversity in literacy skills among DHH adults, and prior computing accessibility research has identified benefits from providing ATS-based reading assistance to other user groups. Thus, while ATS holds promise to benefit DHH adults, little research has investigated the benefits from, interests in and preferences for such technologies among DHH adults. Thus, this work aims to investigate the use of ATS-based reading assistance to support DHH adult readers. This dissertation is organized in three parts centered around three major contributions: requirement- and interest-gathering, methodological research, and design and evaluation work.

Part I of this dissertation focuses on the requirements and interests of DHH adults on ATS-based reading assistance tools. As a relatively new technology, it is still unknown whether DHH adults would be interested in such technologies, what purposes they would be interested in them for, and what personal or technical characteristics may affect their interests. Furthermore, with various approaches to ATS available (e.g. lexical or syntactic approaches), it is unknown which ones would be of interest and/or benefit to DHH adults.
Thus, we first explore perspective of DHH adults on ATS-based reading assistance tools.

Then, Part II of this dissertation focuses on methodological work on how to evaluate the output of ATS technologies with DHH adults. Prior work on the evaluation of linguistic technologies has identified challenges when evaluating their technologies among DHH adults with varying literacy skill levels. However, it remains unknown if such challenges are also present when evaluating the output of ATS systems among DHH adults. Considering that there are various aspects of the output to evaluate, including its complexity or readability, and its overall quality or fluency, we investigate which metrics can reliably evaluate those characteristics of ATS output among this user group.

After establishing interests and evaluation methods for various characteristics of simplified texts, Part III of this dissertation focuses on the design of these tools, and includes evaluations of such tools with DHH adults. We first investigate an exploration of the design space to identify which configurations of the various design parameters may provide more usability to DHH adults, and to understand the rational for their preferences. Using the preferred settings identified, we then conduct an evaluation to investigate the DHH users’ preferences and interactions with a system that provides the various capabilities of ATS technologies using realistic output from existing systems.
Acknowledgments

I would like to thank my Ph.D. advisor Dr. Matt Huenerfauth for all his support throughout this dissertation, and for having believed in my potential to become a productive researcher in our field. I have learned so much from him not only as a researcher, but also as a mentor who always valued my contributions, and was always available to provide guidance for both my research and career. I could not have asked for a better advisor.

I am grateful to the members of my dissertation committee Dr. Kristen Shinohara, Dr. Garreth Tigwell and Dr. Raja Kushalnagar for taking the time to review my work and provide constructive feedback. Their insightful comments have definitely made the work presented in this dissertation stronger. I would also like to thank our program director Dr. Pengcheng Shi for his support through my Ph.D. studies.

I would also like to express my gratitude to the numerous co-authors of the publications that resulted from the work in this dissertation, and our research assistants and collaborators who supported this work. In addition to Dr. Matt Huenerfauth, these include Dr. Lisa Elliot, Dr. Abraham Glasser, Dr. Jessica Trussell, Dr. Sooyeon Lee, Dr. James Mallory, Dr. Wei Xu, Mounica Maddela, Matthew Seita, Akhter Al Amin, Becca Dingman, Matthew Watkins, Velvet Howland, and numerous others. I am grateful for my internship mentors Mike Griffin, Dr. Ding Li and Dr. Valentina Shin whose guidance helped me to grow as a researcher and taught me to look at problems from different angles; for all the feedback and support received from my colleagues at the Center for Accessibility and Inclusion Research (CAIR); and to my colleagues at Creighton University, especially Dr. Dave Reed who served as the initial inspiration to pursue this degree and trusted me to start my career as an Assistant Professor during the last year of my program.

Last, but not least, I would like to thank my family and friends for their continued love, support and belief in me. And I will be forever grateful to Kimberly for her unconditional love, and for always being there for and with me.

This material has also been supported by the National Science Foundation under award No. 1822747.
Me gustaría dedicarle esta tesis a mi papá. Con tanto entusiasmo que sugerías que podría un día ser un gran doctor, nunca nos imaginamos que se podría lograr siguiendo mi curiosidad por la tecnología. Tu ejemplo y confianza infalible en mí continúan y continuarán siendo inspiración cada día.
## Contents

1 Introduction .................................................. 1
   1.1 Focus of This Dissertation .............................. 2
   1.2 Research Questions Investigated in This Dissertation .... 4
       1.2.1 Part I: Needs, Interests and Requirements for ATS Among DHH Computing Professionals .... 4
       1.2.2 Part II: Methods for Evaluating ATS among DHH Adults ................................. 5
       1.2.3 Part III: Design and Evaluation of ATS-based Reading Assistance Tools ............ 6
   1.3 Overview of This Dissertation ............................ 8

2 Background ...................................................... 11
   2.1 Using Automatic Text Simplification as Reading Assistance .............................. 11
   2.2 Diversity in Literacy Skill Among DHH Readers ........................................ 13
   2.3 Automatic Text Simplification versus Translation ........................................ 14

Part I: Requirement and Interest Gathering ..................... 15

Prologue to Part I .................................................. 16

3 Background and Related Work ........................................ 18
   3.1 Need for Domain-Specific User Research ............................... 18
   3.2 DHH Readers' Reading Habits .......................................... 19
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3.4 Mitigating Social Accessibility Issues</td>
<td>57</td>
</tr>
<tr>
<td>5.4 Discussion</td>
<td>58</td>
</tr>
<tr>
<td>5.5 Limitations and Future Work</td>
<td>60</td>
</tr>
<tr>
<td>5.6 Conclusion</td>
<td>61</td>
</tr>
<tr>
<td>Epilogue to Part I</td>
<td>62</td>
</tr>
<tr>
<td>Part II: Methods for Evaluating ATS Tools Among DHH Adults</td>
<td>65</td>
</tr>
<tr>
<td>Prologue to Part II</td>
<td>66</td>
</tr>
<tr>
<td>6 Background and Related Work</td>
<td>69</td>
</tr>
<tr>
<td>6.1 Evaluating Complexity of Simplified Texts</td>
<td>70</td>
</tr>
<tr>
<td>6.1.1 Metrics Employed in Evaluations of Complexity</td>
<td>70</td>
</tr>
<tr>
<td>6.1.2 User Groups Involved in Manual Evaluations of Complexity</td>
<td>71</td>
</tr>
<tr>
<td>6.2 Evaluating Fluency of Automatically Simplified Texts</td>
<td>72</td>
</tr>
<tr>
<td>6.3 Evaluating Other Linguistic Technologies Among DHH Adults</td>
<td>73</td>
</tr>
<tr>
<td>6.4 Limitations and Future Work Needed</td>
<td>74</td>
</tr>
<tr>
<td>7 Metrics for Evaluating Complexity</td>
<td>76</td>
</tr>
<tr>
<td>7.1 Hypotheses Evaluated in This Study</td>
<td>77</td>
</tr>
<tr>
<td>7.2 Method</td>
<td>77</td>
</tr>
<tr>
<td>7.2.1 Reading Stimuli</td>
<td>77</td>
</tr>
<tr>
<td>7.2.2 Metrics</td>
<td>79</td>
</tr>
<tr>
<td>7.2.3 English Reading literacy</td>
<td>82</td>
</tr>
<tr>
<td>7.2.4 Data Collection</td>
<td>82</td>
</tr>
<tr>
<td>7.2.5 Data Analysis</td>
<td>84</td>
</tr>
<tr>
<td>7.3 Results</td>
<td>86</td>
</tr>
</tbody>
</table>
CONTENTS

7.3.1 Discriminative Ability (H1) ........................................ 86
7.3.2 Literacy Bias (H2) .................................................. 91
7.4 Discussion ............................................................. 91
  7.4.1 Discriminative Ability (H1) ....................................... 91
  7.4.2 Literacy Bias (H2) ................................................ 97
7.5 Limitations and Future Work ....................................... 98

8 Metrics for Evaluating Fluency ................................... 100
  8.1 Hypotheses Evaluated In This Study .............................. 100
  8.2 Method .................................................................. 101
    8.2.1 Reading Stimuli ................................................ 101
    8.2.2 Stimuli Generation Procedure ............................... 102
    8.2.3 Metrics .......................................................... 106
    8.2.4 Data Collection Procedure ................................... 109
    8.2.5 Participants ..................................................... 110
  8.3 Results .................................................................. 110
    8.3.1 H1: Discriminative Ability .................................... 113
    8.3.2 H2: Literacy bias ................................................. 118
  8.4 Discussion ............................................................. 119
    8.4.1 H1: Discriminative Ability .................................... 119
    8.4.2 H2: Literacy Bias ............................................... 121
  8.5 Limitations and Future Work ...................................... 122
  8.6 Conclusion ............................................................ 123
CONTENTS

Epilogue to Part II

125

Part III: Design and Evaluation of ATS-based Reading Assistance Tools

Prologue to Part III

129

9 Background and Related Work

9.1 User Interface for Text Simplification

9.2 User Autonomy

10 Lexical Simplification and Autonomy

10.1 Introduction

10.2 Method

10.3 Prototype

10.3.1 Simplifications

10.4 Preliminary Study

10.4.1 Materials

10.4.2 Data Collection Procedure

10.4.3 Recruitment and Participants

10.4.4 Results and Discussion

10.5 Experimental Study

10.5.1 Materials

10.5.2 Data Collection Procedure

10.5.3 Recruitment and Participants

10.5.4 Results

10.6 Discussion

10.6.1 RQ3.2: Providing Simplifications On-Demand

132

133

135

136

137

139

140

140

142

142

143

145

147

148

149

149

151

152
10.6.2 Tradeoffs between On-Demand Prototypes ...................................... 153
10.7 Limitations and Future Work ............................................................... 154
10.8 Conclusion ......................................................................................... 156

11 Exploring The Design Space ................................................................. 157
  11.1 Research Questions ........................................................................... 158
  11.2 The Design Space ............................................................................ 159
    11.2.1 Prototype ................................................................................ 162
    11.2.2 Implementation ...................................................................... 165
  11.3 Method ............................................................................................ 166
    11.3.1 Materials ............................................................................... 167
    11.3.2 Procedure ............................................................................. 169
  11.4 Participants ..................................................................................... 170
  11.5 Data Analysis .................................................................................. 170
  11.6 Results ............................................................................................ 171
    11.6.1 Parameter Preferences (RQ3.3.1) .......................................... 172
    11.6.2 Participants’ Rationale for Preferences (RQ 3.3.2) ............. 173
  11.7 Discussion ....................................................................................... 179
    11.7.1 Visual Strategies and User Interface .................................... 181
  11.8 Limitations and Future Work ........................................................... 182
  11.9 Conclusion ....................................................................................... 184

12 Evaluation of an ATS Tool .................................................................... 185
  12.1 Research Questions .......................................................................... 185
  12.2 Prototype ........................................................................................ 186
  12.3 Method ............................................................................................ 189
    12.3.1 Materials ............................................................................... 189
    12.3.2 Procedure ............................................................................. 192
    12.3.3 Participants .......................................................................... 193
A Publications 255

B Video Demonstration 257

C Survey and Follow-Up Interview Materials 258
  C.1 Survey Questionnaire ........................................ 258
    C.1.1 Introduction ........................................... 258
    C.1.2 Questions ............................................. 258
    C.1.3 Information About Yourself ........................... 267
  C.2 Follow-Up Interviews Questionnaire ....................... 271
    C.2.1 Warm-up .............................................. 271
    C.2.2 Reading .............................................. 271
    C.2.3 Learning for Work .................................... 271
    C.2.4 Facing Complicated Text ............................... 272
    C.2.5 Reactions to Tool .................................... 272

D Social Accessibility Interview Materials 274
  D.1 Questionnaire ............................................ 274
    D.1.1 Warm-up .............................................. 274
    D.1.2 Work ................................................. 274
    D.1.3 Reading / Assistive Technologies ..................... 275
    D.1.4 ATS Tools ............................................ 275

E Articles for Methodological Studies 278

F Articles for Chapters 11 & 12 281
List of Figures

1.1 A flowchart of the various parts of this dissertation, summarizing the research questions addressed and the methods employed to address those questions. 8

4.1 Screenshots from the video demonstration shown to participants in the survey and follow-up interviews demonstrating (a) syntactic simplification and (b) lexical simplification. Simplified text is highlighted in yellow, while in (b), words identified as complex are highlighted in gray. 27

4.2 Participants’ responses to the questions: On a day that you read, how much time do you spend reading text (a) on a screen (e.g. computers, laptops, phones, tablets, etc.) and (b) that is not on a screen (e.g. books, magazines, newspapers, print-outs). There was a significant difference between the two (p <0.01). 32

4.3 Counts for responses to the different reading purposes included as options for the question “For what purposes do you read? (select that apply).” 33
4.4 Participants’ responses to the questions: “How often do you read to learn about technical topics for your work (e.g. information about new technologies, new software, programming information, etc.)?” and “How often do you watch videos to learn about technical topics for your work (e.g. information about new technologies, new software, programming information, etc.)?” The full text in the options was: “rarely (less than once a month),” “monthly (one to three times a month),” “weekly (once a week),” “often (two to four times a week),” and “daily (five or more times a week).” There was a significant difference between the two ($p < 0.05$).

4.5 Counts for responses to the different workarounds included as options for the question “How do you try to understand text that is complicated? (select all that apply)”

4.6 Survey respondents’ agreement to a Likert-scale question “I would be interested in a tool that helps me to understand text by making it simpler” (a) overall and (b) broken down by communication preferences. Communication preference levels: Spoken = “Spoken language only”, Mostly Spoken = “Mostly spoken language, with a little sign language”, Half = “About half spoken language, half sign language”, Mostly Sign = “Mostly sign language, with a little spoken language”, and Sign = “Sign language only.”

4.7 Participants’ Likert-scale responses to the question “Please choose your interest in a tool like the one in the video for each activity below.”

7.1 Low-linguistic complexity comprehension questions success for H1, with a max. value of 100% ($* = p < 0.05$).

7.2 Score predictions for the comprehension questions for H1, with a max. value of 100% ($* = p < 0.05$).

7.3 Understandability judgements for H1 using a Likert-type agreement scale ($* = p < 0.05, *** = p < 0.001$).
7.4 Readability judgements for H1 using a Likert-type agreement scale ( ** = p < 0.01, *** = p < 0.001). .................................................. 90

7.5 Reading speed for H2, calculated in words per minute (wpm) ( *** = p < 0.001). .......................................................... 92

7.6 High-linguistic complexity comprehension questions success for H2, with a max. value of 100% ( *** = p < 0.001). ......................... 92

7.7 Low-linguistic complexity comprehension questions success for H2, with a max. value of 100% ( *** = p < 0.001). ......................... 93

7.8 Score predictions for the comprehension questions for H2, with a max. value of 100% ( *** = p < 0.001). ........................................ 93

7.9 Understandability judgements for H2 using a Likert-type agreement scale ( *** = p < 0.001). .................................................. 94

7.10 Readability judgements for H2 using a Likert-type agreement scale ( ** = p < 0.01). ............................................................... 94

8.1 Reading speeds for H1, measured in words per minute ( *** = p < 0.001). ........... 113

8.2 Understandability judgements for H1 using a Likert-type agreement scale ( * = p < 0.05). ...................................................... 114

8.3 Readability judgements for H1 using a Likert-type agreement scale ( * = p < 0.05). ............................................................... 115

8.4 Grammaticality judgements for H1 using a Likert-type agreement scale ( ** = p < 0.01, *** = p < 0.001). ................................. 116

8.5 Low-linguistic complexity comprehension questions scores for H2, with a max. value of 100% ( ** = p < 0.01). ......................... 117

8.6 High-linguistic complexity comprehension questions scores for H2, with a max. value of 100% ( *** = p < 0.001). ......................... 117

8.7 Understandability judgements for H2 using a Likert-type agreement scale ( ** = p < 0.01). .......................................................... 118
10.1 An article used for the video demonstrations in our preliminary study, with a zoomed-in view of what some text looked like under different design variations for user initiative and change visibility, which were displayed to users in this study. ................................................................. 137

10.2 Participants’ response to a 1-to-6 scalar question on how likely they were to use a system based on each level of the user initiative and change visibility design dimensions, from 1 (very unlikely) to 6 (very likely). This small study was insufficiently powered for statistical difference testing, but it was meant to guide prototype selection for the main study. Responses are displayed using a diverging stacked bar graph, as recommended in [104] ........................................ 143

10.3 One of the articles used as stimuli in our experimental study, along with zoomed-in views of how text appeared under the four different conditions, e.g. clicking on a word in the decoration condition or hovering over a word in the pop-up condition. .................................................. 145

10.4 Participants’ responses to questions about all four conditions in the experimental study, including subjective Likert-scale responses for (a) the text was easy to read and (b) I was able to understand this text well, with significant pairwise differences marked with asterisks (* p <0.05). In (c), analysis on objective comprehension questions did not reveal any significant differences between the four conditions. .................................................. 150

10.5 Participants’ agreement to a Likert-scale question, presented at the end of the study, as to whether they would be likely to use each of the three simplification conditions, with asterisks marking significant pairwise differences (**) p<0.01, (***) p<0.01). .................................................. 151
11.1 Screenshots of the prototype browser extension. Image (a) shows the prototype when lexical simplification is selected. Image (b) shows the option for “words” in the options for Quantity greyed out as it is not compatible with both and syntactic simplification.

11.2 Screenshots of portions of a ScienceDaily article showing 3 combinations of the various parameter settings. Figure (a) shows lexical simplification being employed at a word-level, using all highlights and providing replacements permanently in place. Figure (b) shows lexical + syntactic simplifications provided as a temporary pop-up for sentences, without any highlights. Figure (c) shows syntactic simplification alone, applied at a sentence-level, provided manually on the side with highlights only for suggested sentences.

12.1 Screenshots of the browser extension with a toggle for enabling (as shown in screenshot a) or disabling (as shown in screenshot b) the extension. An additional setting shows the options for the simplification type.

12.2 Screenshots of a paragraph from a ScienceDaily article when using the 3 simplification types for one word or sentence. Gray indicates suggestions, while yellow indicates replacements. The lexical + syntactic option is labeled as “Both.”

12.3 Participants reading speed under the conditions of using the tool and not using the tool ( * = p < 0.05)

12.4 Divergent stacked bar charts [104] summarizing participants’ judgements of the usefulness of the tool. The top 3 bars summarize participants’ judgements for each simplification type, while the bottom bar summarizes participants’ judgements for the tool as a whole, which were provided separately. (** indicates p < 0.01, * indicates p < 0.05).
List of Tables

2.1 Example from [129] of syntactic and lexical simplification (for lexical, the replaced word is shown in bold). .......................................................... 12

7.1 One of the comprehension questions for an article, in its two versions. . . . . 80

7.2 A summary of the results for each metric across each hypotheses. Hypotheses were not supported if the metric did not discriminate between any complexity levels, and partially supported if it discriminated between some but not all pairs of complexity levels. .......................................................... 85

8.1 Excerpt from each condition, illustrating different levels of fluency. . . . . 106

8.2 A summary of the results for each metric, for each of the hypotheses. Hypotheses were not supported if the metric did not discriminate between any complexity levels, and partially supported if it discriminated between some but not all pairs of complexity levels. .......................................................... 111
Chapter 1

Introduction

In this dissertation, we investigate Automatic Text Simplification (ATS) technologies to support the reading tasks of adults who are Deaf or Hard of Hearing (DHH). Fifteen percent of U.S. adults are DHH [16], and prior literacy research has established that there is a wide range of reading skills among DHH adults: While some are strong readers, studies have also reported median reading skills of DHH high school (secondary school) graduates in the U.S. being at the “fourth-grade level,” which is a standard for U.S. students in elementary school typically at age 9 [120]. Other studies have found that over 17% of deaf adults can be considered as readers with low literacy [42]. Furthermore, research has found that there are lower educational outcomes among DHH individuals, e.g. in 2017, 34% of hearing individuals in the U.S. had completed a bachelor’s degree, compared to only 18.8% of deaf individuals [46]. Reading literacy is also important in an employment context, e.g. for learning new skills mid-career [109]. Research has revealed a 22% lower rate of employment among DHH adults, and DHH adults’ salaries are only 64%, in comparison to their hearing peers [124]. This impact may be greater in contexts such as the computing field in which, prior work has established that professionals often need to learn about new technologies in their own [109, 140]. These findings motivate our focus on DHH individuals.

ATS consists of a variety of computing techniques for rewriting or modifying a text to
improve its readability and understandability, while preserving the original meaning [111]. This is typically done by restructuring phrases or sentences in a text, which is known as syntactic simplification; by rewriting individual words or phrases, which is called lexical simplification; or hybrid combinations of both [111]. While ATS can have computing-based applications such as providing simplifications for a machine-translation pipeline, a growing body of research has evaluated the use of ATS to provide reading assistance to various user groups, including low-literacy readers [128], readers with aphasia [34] or dyslexia [99, 101], or second-language learners [9]. As prior research had identified great diversity in literacy skill among DHH adults in the U.S., e.g. [84, 93, 120], a few researchers have investigated the use of ATS as reading assistance for DHH readers. For instance, studies have found benefits from providing DHH adults with syntactic simplification in medical texts [73]. However, many open questions remain in this context of using ATS to provide reading assistance to DHH adults, including their needs and interests in these technologies, how to best design and evaluate these technologies with this group, and how they may benefit from other approaches to ATS such as lexical simplification.

1.1 Focus of This Dissertation

While there has been a lot of progress in the field of natural language processing (NLP) on the advancement of the computing techniques powering ATS, much less research has focused on the human computer interaction (HCI) aspects of these technologies. While there has been a considerable amount of computing accessibility research with users with disabilities on the use of ATS to provide reading assistance tools for different user groups, as outlined above, most of that work has focused on evaluating the underlying technology itself, i.e. the quality of the text that had been processed by ATS systems (e.g. [105]), identifying linguistic properties that affect text readability for different user groups [92, 100], measuring the benefits of providing ATS-based reading assistance tools [73, 82], and, in a few cases, investigating some aspects of the user interface of these tools [101].
CHAPTER 1. INTRODUCTION

However, with the exception of one study focused on the needs for reading assistance tools for reading online reviews among adults with autism [137], little prior work has focused on investigating the initial needs, interests and requirements of the respective target user groups of their research. Further, as an emerging field, evaluations of these technologies also remain an important part of the research conducted. However, while prior work has called for research identifying the best manual evaluation methods for evaluating ATS with target users [117], to the best of our knowledge no prior work has provided methodological guidance on how to best evaluate these technologies among target users, let alone DHH adults. Finally, while there are numerous design aspects involved in the design of ATS-based reading assistance tools, little work in the reported literature has investigated the preferences of target user groups for the possible settings of those parameters, nor how they may affect their preference and adoption of these tools.

Thus, the focus of this dissertation is on the HCI aspects of developing ATS-based technologies, specifically for DHH adults, as motivated above. More specifically, we aim to bridge the gap between the needs of DHH adults for these technologies, and the researchers in NLP who are developing the underlying technologies. The contributions of this dissertation include the identification of the needs, requirements and interests of this user group for ATS-based reading assistance tools. Our contributions also include methodological guidance on what are the best metrics for reliably evaluating various aspects of the output of ATS among this user group, as well as identifying their preferences and perspectives for the various design parameters involved. Through evaluations of a fully functional prototype based on those previous contributions, we inform NLP researchers about the ways DHH adults may interact and benefit from ATS technologies. Finally, the approaches we employ in this dissertation may be used by other researchers aiming to investigate ATS-based reading assistance tools with other user groups from an HCI perspective.
1.2 Research Questions Investigated in This Dissertation

In order to investigate the use of ATS to provide reading assistance to DHH adults and provide the contributions outlined above, we rigorously investigate several research questions, which have been grouped in 3 parts related to: interest and requirement gathering, methodological research on methods to evaluate the output of ATS among DHH adults, and to the design and evaluation of ATS-based reading assistance tools. More specifically, we investigate the following research questions:

1.2.1 Part I: Needs, Interests and Requirements for ATS Among DHH Computing Professionals

Part I of this dissertation focuses on the initial requirement and interest gathering of DHH adults with experience in computing for ATS-based reading assistance tools. Two big research questions are investigated in Part I:

RQ1.1. What are the current reading experiences of DHH individuals with work experience in the computing field, and are they interested in ATS-based tools to assist those reading experiences? This overarching research question includes their current reading practices, such as how much reading are they doing, how are they reading (i.e. on a screen or paper-based), and what are they reading for. We also investigate how often they engage in reading for learning about new topics at work, what their views are about their experience with complicated text, and how do they overcome it. And finally, whether they are interested in ATS-based reading assistance tools, and if so, for which reading activities would they be interested in it for.

RQ1.2. What are the perspectives of DHH adults with work experience in the computing field on the social accessibility of ATS-based reading assistance tools? This question includes their perspectives on what
they value at work, how they want to be perceived by others, and how they perceive the personal and social acceptability of using ATS-based reading assistance tools in a workplace environment. This question also includes potential the investigation of potential design solutions for mitigating social accessibility issues that may arise.

1.2.2 Part II: Methods for Evaluating ATS among DHH Adults

Part II then focuses on comparing methods for evaluation of ATS-based reading assistance tools among DHH adults. This methodological research is needed to determine the best methods for evaluating these technologies among this user group. Thus, Part II begins by investigating the following research questions:

RQ2.1: When English texts of different complexity levels are evaluated by DHH individuals with (a) lower English literacy skill and (b) higher English literacy skill, which metrics reveal statistically significant differences among texts that vary in a specific characteristic of the output of ATS? We refer to this characteristic as the discriminative ability of the metric, and this is a desirable characteristic. If there is a significant difference, then the metric is effective for use in evaluating that particular characteristic being evaluated. We investigate this research question for two characteristics of the output of ATS:

- **RQ2.1.1.** Which metrics exhibit a discriminative ability in evaluating the complexity of simplified texts among DHH adults at the different literacy levels? This question focuses on the complexity (or simplicity) of the simplified text, a key aspect of the output of ATS systems.

- **RQ2.1.2.** Which metrics exhibit a discriminative ability in evaluating the fluency of simplified texts among DHH adults at dif-
different literacy levels? This question, in turn, focuses on the fluency of simplified texts, which may be affected by errors introduced in the simplification process.

RQ2.2. In an overall analysis of response scores for all texts varying in a particular characteristic of the output of ATS, when we compare response scores between DHH individuals in the higher-literacy and lower-literacy groups, do we observe a significant difference? We refer to this as the literacy bias of the metric. While this is not necessarily a problem with the metric that would prevent its use in evaluating text complexity, researchers who use this metric must ensure that they report the literacy skill level of their participants in a study, in order for results across studies to be comparable. Similar to RQ2.1, we explore this question for the same two characteristics of the output of ATS:

- **RQ2.2.1.** Which metrics exhibit a literacy bias when evaluating the complexity of simplified texts among DHH adults at the different literacy levels? This question focuses on the complexity (or simplicity) of the simplified text.

- **RQ2.2.2.** Which metrics exhibit a literacy bias when evaluating the fluency of simplified texts among DHH adults at different literacy levels? This question, in turn, focuses on the fluency of simplified texts.

1.2.3 Part III: Design and Evaluation of ATS-based Reading Assistance Tools

After the initial requirement gathering in Part I and methodological research summarized in Part II, Part III of this dissertation focuses on investigating the preferences of DHH adults
for the parameters involved in the design of ATS-based technologies. More specifically, Part III investigates the following overarching research questions:

**RQ3.1:** Does providing simplification for individual complex words affect DHH users’ perception of their ease in reading a text or perception of how well they understood it? While prior research had identified that simplifying the syntactic structure of texts in medical texts provided benefits to DHH readers [73], no prior work had identified whether there are benefits obtained by DHH adults when simplifying individual complex words.

**RQ3.2:** When employing ATS-based reading assistance tools that simplify individual complex words, are DHH users’ subjective preferences about these tools affected by whether they are provided with greater autonomy (i.e. in requesting simplifications on-demand or seeing what words have been replaced)? As an initial effort into investigating the various design parameters of ATS tools, this question focuses on whether there are benefits from employing settings for two parameters that provide the user with more autonomy when using ATS-based reading assistance tools.

**RQ3.3.** When employing ATS-based reading assistance tools that simplify individual complex words or the syntactic structure of sentences, what are the preferences of DHH users for the various parameters involved in the design of these tools? This question further explores design parameters including ones explored in RQ3.2 for simplifying words, but also those related to how much text is replaced at once, as well as the location and permanence of the simplifications.

**RQ3.4.** When providing an ATS system that provides both lexical and syntactic simplification features, as well as hybrid approaches that combine lexical and syntactic simplification to DHH users, how
do users interact with and benefit from such system, and how do their preferences vary for the various simplification approaches? This question focuses on the first comparative exploration in which users actually engage with a simplification system that provides various approaches to simplification that reflect the state of the art of these systems.

Figure 1.1: A flowchart of the various parts of this dissertation, summarizing the research questions addressed and the methods employed to address those questions.

1.3 Overview of This Dissertation

In order to provide the background necessary for a reader for subsequent chapters, Chapter 2 begins by summarizing the essential details about ATS technologies and their use for reading assistance, prior work highlighting the diversity in literacy skills among DHH adults, as well as comparisons between ATS and providing translations, which could be another alternative for DHH adults who are ASL users. The rest of the dissertation is organized in the 3 parts outlined above, which are also summarized in Figure 1.1.

Part I of this dissertation begins by exploring the needs, interests and requirements of DHH adults in the computing field for ATS-based reading assistance tools. Part I thus begins with Chapter 3 by outlining prior work motivating the need for domain-specific user research in the context of ATS (which motivates our focus in Part I on DHH adults in the computing field), investigating the reading habits of DHH adults, as well as research into
the social accessibility of assistive technologies, which is important to consider early on in the design of these technologies [116]. Then, Chapter 4 summarizes a survey study, which also included follow-up interviews to gain a deeper understanding of the survey results. The survey and interviews focused on learning about the reading experiences of DHH adults with experience in computing, including what and how they typically read at work. In that study, we also explored participants’ interests in ATS-based reading assistance tools for various purposes identified in the survey. Finally, Chapter 5 then summarizes an interview study focused more specifically on the social accessibility of ATS-based reading assistance tools, identifying participants’ values at work, how the issues of social accessibility emerging from the use of ATS-based reading assistance tools may come into conflict with those values, and which design aspects may help mitigate those issues.

Then, Part II focuses on the methods for evaluating the underlying ATS technologies among DHH adults. Part II thus begins with Chapter 9 by outlining prior work from a lens of how prior work has evaluated ATS technologies, identifying potential methodologies to use, as well as a lack of methodological guidance on the effectiveness of those methodologies. Thus, Chapters 7 and 8 summarize methodological studies focused on identifying appropriate metrics for use among DHH adults at different literacy levels when evaluating two key characteristics of the output of ATS: the complexity (Chapter 7) and fluency (Chapter 8).

Finally, Part III of this dissertation focuses more specifically on the design of ATS-based reading assistance tools for DHH adults. In Chapter 6, prior work on ATS-based reading assistance tools is summarized with a focus on the user interface of those tools, along with HCI work on user autonomy. Then, Chapter 10 summarizes an experimental study focused on investigating whether there are particular benefits obtained from providing DHH adults with a specific type of ATS, namely, lexical simplification. That study also included initial explorations of participants’ preferences for design parameters that related to the level of autonomy the tools provide to the users, and how varying those levels affect users’ preferences for these systems. Then, Chapter 11 first summarizes the various design parameters involved in the design of such tools, and the possible settings of those parameters that have
been used in prior work. Chapter 11 also summarizes our study exploring this design space with DHH adults. More specifically, we investigate participants’ preferences for the various design parameters, and how participants’ preferences for those design parameters may vary depending on the type of ATS employed. Finally, drawing from the design explorations in Chapter 11, Chapter 12 summarizes our evaluation of realistic ATS systems with DHH adults to explore how they interact with and potentially benefit from the various types of ATS. Lastly, Chapter 12 also explores whether participants’ preferences towards the various approaches of ATS vary upon engaging in reading tasks with a tool that provides those types of ATS on demand.
Chapter 2

Background

As summarized in the Section 1, ATS consists of computing techniques that aim to rewrite text to improve its readability or understandability, and thus it involves the identification of complex text, as well as the generation and selection of appropriate alternatives [111, 117].

There are three main approaches to modifying text for readability or understandability [111, 117]: syntactic simplification, in which the structure of sentences or phrases are modified to reduce grammar complexity, e.g. [33]; lexical simplification, which involves identifying and replacing complex words with simpler synonyms, e.g. [80]; and hybrid approaches that combine both technologies, e.g., [141]. Table 2 shows an example of a sentence simplified using both approaches, from [129].

2.1 Using Automatic Text Simplification as Reading Assistance

While text simplification can have technical applications within various NLP pipelines, e.g., as a precursor for machine translation, a growing body of research has examined its use for providing reading assistance tools to specific groups of people who may benefit from reading support. Prior work on ATS as an assistive technology has focused on its use for
The senators alleged that public opinion would not receive the disbursement willingly.
Then, the senators rejected the proposal from the Information Technology Division of the Senate to change all computers in the plenary room.

The senators rejected the proposal from the Information Technology Division of the Senate to change all computers in the plenary room, alleging that public opinion would not receive the expense willingly.

| Original: | The senators rejected the proposal from the Information Technology Division of the Senate to change all computers in the plenary room, alleging that public opinion would not receive the disbursement willingly. |
| Syntactic | The senators alleged that public opinion would not receive the disbursement willingly. Then, the senators rejected the proposal from the Information Technology Division of the Senate to change all computers in the plenary room. |
| Lexical | The senators rejected the proposal from the Information Technology Division of the Senate to change all computers in the plenary room, alleging that public opinion would not receive the expense willingly. |

Table 2.1: Example from [129] of syntactic and lexical simplification (for lexical, the replaced word is shown in bold).

There have been many user studies in this area, focusing on different aspects of the systems. For instance, prior ATS studies that have focused on specific target user groups have investigated whether users benefit from the various approaches to text simplification [73, 82], or comparing how the use of different systems impact those benefits [100]. Few studies measured any difference in objective measures (e.g. comprehension questions) [73], but some measured a difference in users’ subjective perception of whether they benefited from the technology [99, 105].

A number of user groups including people with dyslexia, e.g. [99]; with aphasia, e.g. [34]; second-language learners, e.g. [9]; children, e.g. [30]; low-literacy readers, e.g. [129]; and more closely related to our research, DHH adults and children, e.g. [58, 73]. The three approaches identified above may benefit user-groups to a different degree, depending on their literacy profile [105]. Prior work sheds light on the reasons for this, identifying that the sources of reading difficulty for different user groups may vary [92], and thus the findings from one user group may not generalize to another as different user groups may benefit from the various approaches to ATS.
Other work has focused on identifying the linguistic needs of different user groups (i.e. which linguistic properties affect the readability of a text), finding that the linguistic properties that affect readability for different user groups may not always be the same [92, 100]. One prior study also investigated the user needs and interest of adults with autism for reading in ATS tools for reading online reviews [137]. There has also been research that incorporates evaluations of the quality of the text produced by ATS systems, traditionally measured by “expert” readers – usually native speakers – rather than the target users (e.g. [48, 105, 133]).

2.2 Diversity in Literacy Skill Among DHH Readers

Prior standardized testing has measured a median of fourth-grade reading levels among DHH high school graduates in the United States [47], with other studies on reading comprehension reporting sixth-grade reading levels among DHH university students [1, 93]. Furthermore, prior research has described over 30% of deaf high school graduates in the United States as “functionally illiterate” [84], with other studies suggesting that over 17% of deaf adults have “low literacy” [42]. It is important to note that these are subsets of the samples in these study, and they are not reflective of the entire population. Instead, what these suggest is that while many DHH readers have age-appropriate reading skills, there are also significant subsets of DHH adults who encounter challenges when reading and thus, there is great diversity in literacy skill among this user group. ATS has the potential to benefit DHH readers given that prior literacy research has identified both syntactic and lexical aspects of text as key sources of difficulty for lower literacy DHH readers. In fact, prior research on ATS has found benefits from syntactic approaches to ATS [73].

Prior research has also examined the particular literacy profile of DHH readers, which sheds light on aspects of the reading task that may be the greatest challenge. For instance, some prior work on DHH readers has identified syntactic structure as a potential source of reading difficulty [29]. However, recent research has also suggested that syntax is not
the only source of difficulty, as vocabulary knowledge also plays a role in the diversity of DHH readers’ literacy [28, 78]. Furthermore, research into the reading strategies employed by DHH readers has also identified unfamiliar vocabulary as a key challenge [10]. Due to the literacy challenges that many DHH readers face, ATS tools provide a possible solution. Furthermore, the specific research on DHH reader’s literacy summarized above motivates research into a variety of ATS technologies, operating at various linguistic levels, to support DHH readers. Notably, prior studies with lexical simplification among users with dyslexia [99] did not reveal significant effects on comprehension scores, but participants reported perceiving the text significantly easier to read and understand under certain interface treatments. However, while there have been benefits identified from using syntactic approaches to provide reading assistance to DHH readers, it remains unknown whether automatic lexical simplification would be also beneficial for DHH readers or whether they would be interested in – and perceive a benefit from – a system that provides this form of simplification alone or in combination with syntactic simplification (which is also investigated specifically in this dissertation).

2.3 Automatic Text Simplification versus Translation

While for second language learners, translation of text into another language might be beneficial, translating English text on websites into American Sign Language (ASL) would not benefit all DHH individuals in the U.S., since there is great diversity in the levels of ASL proficiency [44]. In addition, even for DHH users who are ASL signers, providing ASL translations of English text on-demand is currently infeasible: Requesting human sign-language interpreters to translate texts online into ASL would be very resource-intensive, and the state of the art for automatic ASL machine translation is not yet sufficient for this task, as discussed in [18]. Furthermore, research has found that providing text-simplification not only has language-learning benefits, but also when non-native readers were provided with both translation and simplification technology, they preferred simplification [39].
Part I:

Requirement and Interest Gathering
Prologue to Part I

The overall goal of this dissertation is to investigate the use of ATS-based reading assistance tools to support the reading tasks of DHH adults. To that end, this work seeks to inform researchers in ATS about the best ways to design and evaluate ATS-based reading tools. Thus, a necessary first step is to motivate this research by investigating the requirements of DHH adults with their direct participation. Part I of this dissertation thus begins with direct participation of a subgroup of DHH adults: those experience in the computing field, a field that often requires self-directed learning. This first part focuses on learning about their needs, interests and requirements for ATS-based reading assistance tools. The results from Part I also serve as a foundation for Parts II and III, which focus on the design and evaluation of such tools. Because of the preliminary nature of the studies presented in this part, participants in these studies only interact with video demonstrations of the tools (as opposed to reading simplified texts or interacting with functional prototypes as in Parts II and III) in order to gather their impressions of those demonstrations.

Part I begins by outlining prior work on the need for domain for domain specific research (which motivates our focus on DHH adults in the computing field), the reading habits of DHH adults, and considerations of social accessibility in the design of assistive technologies in Chapter 3. Then, Chapter 4 presents a survey and interview study focused specifically on investigating the needs and interests of DHH adults in ATS-based reading assistance tools. Finally, Chapter 5 presents an interview study focused on the social accessibility of ATS-based reading assistance tools in a workplace context.
More specifically, in Part I of this dissertation, the following overarching research questions are explored:

RQ1.1. What are the current reading experiences of DHH individuals with work experience in the computing field, and are they interested in ATS-based tools to assist those reading experiences? This research question, explored in Chapter 4, was broken down into the following sub-questions:

RQ1.1.1: What are the reading practices of DHH individuals with work experience in the computing industry? More specifically, how much reading are they doing, how are they reading (i.e. on a screen or paper-based), and what are they reading for?

RQ1.1.2: How much do DHH individuals in the computing industry engage in reading for learning about new topics at work?

RQ1.1.3: What are the views of DHH individuals in the computing industry about their experience with complicated text, and how do they overcome it?

RQ1.1.4: Are DHH individuals in the computing industry interested in ATS-based reading assistance tools? And if so, for which reading activities would they be interested in it for?

RQ1.2: What are the perceptions of DHH individuals with experience in the computing field on the social accessibility of ATS-based reading assistance tools? How can these tools be better designed to address potential issues of social accessibility? This research question is examined in chapter 5.
Chapter 3

Background and Related Work

This chapter provides context for Part I of this dissertation by first motivating the need for domain-specific user research in the context of ATS-based reading assistance tools. Then, this chapter describes research looking into assistive technologies for DHH adults in a workplace context, as well as DHH adults’ general reading habits. Finally, this chapter summarizes research on the social accessibility of assistive technologies, including related work on the social accessibility of technologies for DHH users.

3.1 Need for Domain-Specific User Research

ATS research has recently emerged as a sub-field of NLP, and a key challenge in the field has been access to training data (e.g. simplification corpora) [111]. One approach that has been proposed to address this challenge is the collection of data sets of judgements from non-native English speakers on the complexity of individual words from a general lexicon, which can be used to train simplification models [80]. Maddela and Xu [80] suggest that it would be useful to gather additional data sets with judgements from other specific user groups or vocabulary on specific domains (e.g. medical, computing or legal domains). However, to inform the construction of such data sets with judgements from particular user
groups, or within particular domains, it is important to conduct user research into those user groups’ reading habits, reasons for reading, and other details of a particular group of readers in specific domains.

Considering this part’s focus on the needs and interests of DHH users working in the computing field, and the lack of prior work on reading behaviors of DHH individuals in the workplace or their interests for assistive technologies to assist with those reading tasks, we also reviewed research into other assistive technologies for DHH users in the workplace. Most prior work in this area, however, has centered around the use of Automatic Speech Recognition (ASR) and captioning. Prior research includes investigations of the use of ASR to facilitate communication with hearing colleagues [37] or customers [81], or how the presence of ASR technologies may affect hearing people’s behavior, which might in turn affect the technology’s performance [108].

### 3.2 DHH Readers’ Reading Habits

In order to investigate the needs and requirements for ATS-based reading-assistance tools among DHH users in the computing field, it is important to consider what their current reading habits are (i.e. how much reading they do now and what purposes they read for), as well as what tools or workarounds they currently use to support themselves when reading complicated text. When reviewing related work for the latter, we found prior literacy research that has investigated the reading strategies DHH readers used to understand text (e.g. [10, 20, 53, 107]). This prior research has primarily focused on investigating the internal strategies employed by DHH readers, which include inferential and metacognitive strategies – i.e. strategies that rely on the reader’s own awareness and control of their understanding – such as identifying alternative meanings for individual words, replacing unfamiliar words or phrases with more familiar words or paraphrases, or translating words into American Sign Language (ASL) [10]. However, to the best of our knowledge, no prior work has investigated what specific external tools DHH readers employ when encountering
complicated text.

While the prior work above has examined how DHH individuals engage in reading, relatively little work has investigated their general reading habits. The only study we found focusing on the reading habits of DHH individuals, in which researchers compared the reading habits of DHH and hearing university students, found that DHH participants reported reading more often (ranging from 6 to 56 hours a week) than hearing participants (ranging from 1 to 43 hours a week) [85]. Furthermore, the most frequent reading activities reported in that study by DHH participants’ involved e-mail and other Internet media [85].

There has been wide diversity in the way reading habits of the general American population (not focusing on DHH individuals) have been measured and reported in prior work. For instance, there are studies that report general frequency of reading, such as a survey in which 80% of participants reported reading occasionally for pleasure and 50% of full-time workers reported reading every day for work or academic purposes [94]. Others, in turn, report the amount of time spent reading on a day, including studies in which individuals ages 15 to 54 reported reading on average 10 minutes per day [89] or which estimate that the general population reads an average of 15.6 minutes per day [130]). This diversity is also present in what the focus of their investigations are. For instance, while some prior work compared people’s reading frequency between different purposes for reading, including comparison between work or academic purposes and leisure (e.g. [54, 68]), others have focused on leisure alone (e.g. [47, 89]), or books (e.g. [142]). Other studies have focused on the relationship between participants’ demographic factors and their reading habits, including factors such as race (e.g. [106]), reading proficiency (e.g. [113]) or levels of education and occupation (e.g. [68, 118]).

However, these findings – for DHH readers or Americans in general – may not generalize to DHH adults, let alone professionals in a specific field, for prior research on the general population has found that not only the “setting” (i.e. work vs. leisure), but also an individual’s occupation affected their reading habits [68]. Thus, the authors of that study suggest, when investigating the reading habits of a particular user group, it is important to
CHAPTER 3. BACKGROUND AND RELATED WORK

consider the contexts and settings of interest in which reading may occur, as well as readers’ characteristics and reading habits. To the best of our knowledge, however, no prior work has investigated the contexts and settings of interests for reading, nor the reading practices of DHH individuals in the computing industry. Thus, in this work, as we investigate these individuals’ needs and interests in ATS-based reading assistance tools, we also examine their reading habits and contexts of interest to them, and in which of these contexts users would be interested in having ATS-based reading assistance.

3.3 Social Accessibility

Finally, as we investigate preliminary interests in potential assistive technologies for the workplace, it is important to consider related work on the social accessibility of assistive technologies as encouraged by recent work in the area [116]. Social accessibility, a term defined in [116], is used to distinguish the social aspects of assistive technologies from their functional aspects. In this sense, the term social accessibility encompasses the understanding of perceptions of assistive technologies in social contexts as they relate to users’ social identities and how their abilities are portrayed. Many research studies have focused on one aspect of social accessibility, social acceptability, which in the context of social accessibility relates to perceptions of the appropriateness of using assistive technologies in social environments, and it involves tensions between users’ aspirations or fears, and social norms or onlooker’s perceptions [65].

Prior work has established that social influences have significant effects in personal adoption of technologies when their use of is mandatory (as opposed to voluntary) for users without disabilities [122]. However, the most closely related work to our present work lies in studies of social acceptability in the context of users with sensory disabilities, where the choice of use of a particular assistive technology may not necessarily be mandatory, but intended to provide access and may involve social stigmas [116].

Research with users with different sensory disabilities has found that well-functioning
assistive technologies enable greater feelings of independence and self-efficacy [76, 116]. However, research has also identified bystanders’ misperceptions of assistive technologies, which include beliefs that assistive technologies eliminate someone’s disability, or that users with disability cannot do anything without their assistive technologies [115]. These misperceptions, especially when a device malfunctions, can also lead to ambiguities around the users’ abilities, such as the bystander concluding that users are incompetent or not capable of doing tasks that they are actually able to do [115]. Those ambiguities, along with potential breakdowns or malfunctions of the assistive technologies, can cause negative feelings, such as self-consciousness [115].

In the context of DHH users, prior work on the social acceptability of assistive technologies has focused on devices such as wearables for accessing spoken or auditory information [41], finding that users’ perceptions of the social acceptability of wearable sound-awareness devices was most affected by the context of use [41]. Other research into similar devices has found that whether technologies are framed as assistive also influences onlookers’ perceptions of their social acceptability [96], while other work with users with disabilities in general identified that designing assistive devices resembling mainstream devices may also help in mitigating social acceptability issues [115].

The prior work with DHH users outlined above, however, focused mostly on the use of assistive devices (e.g. [41, 96]), as opposed to software tools that may be used within already-mainstream devices. Thus, while their results regarding the sources for social accessibility issues as well as potential solutions to mitigate those may be informative, they may not generalize directly to assistive software, such as ATS-based reading assistance tools. Also, considering that many members of Deaf culture do not consider deafness a disability, but a cultural difference instead, the social accessibility of assistive technologies relating to language may involve nuanced tensions not present in other contexts, and other perceptions (or misperceptions) in addition to those identified in [115] may be at play. However, to the best of our knowledge, issues of social accessibility of ATS-based reading assistance tools, such as what kinds of feelings they may generate among users or what aspects can affect
their social acceptability, have not been explored among DHH adults (or any user group), let alone in a workplace context. Considering encouragement from prior work to consider these issues early in the design of new assistive technologies [74, 116], we also explore the perceptions of DHH adults with experience in the computing industry towards the acceptability of ATS-based reading assistance tools from a social accessibility perspective.
Chapter 4

Reading Experiences and Interest in Reading-Assistance Tools Among DHH Computing Professionals

4.1 Introduction

In addition to the reading skills necessary in personal and social contexts, written language literacy is also important for professional success for many fields. For instance, lower reading skills may pose challenges to individuals working in the computing and information technology field, as prior research has established that such workers are often required to learn about new topics on their own to keep their technical skills up to date [109]. Prior work has found that computing professionals rely to a large extent on reading text-based resources for this learning and thus low literacy may pose a challenge for computing professionals who are DHH [140]. Furthermore, people who are DHH are underrepresented in the computing field, with a recent Stack Overflow survey finding that only 0.8% of users
CHAPTER 4. READING EXPERIENCES AND INTERESTS IN ATS

identified as DHH [90]. This underrepresentation thereby motivates research on potential barriers to professional success.

This chapter presents a survey of 32 people who are DHH with experience working in computing and information technology fields, to investigate their needs and interests in ATS-based reading-assistance tools. The questionnaire included a brief non-interactive video demonstration of ATS-based reading assistance tools so that participants could discuss their interest in using such technologies. After the survey study, follow-up interviews were conducted with five respondents to gain a deeper understanding into the survey results.

4.2 Method

To investigate domain-based user needs for reading assistance tools based on automatic text simplification, we conducted a mixed-method study including pilot interviews, an online survey, and follow-up interviews with DHH individuals who have had experience in the computing and information technology fields. The pilot interviews (N=12) informed the design of the online survey (N=32), and the follow-up interviews with a subset of survey respondents (N=5) provided a deeper understanding of the patterns that had emerged from in survey results. In this section, we present the methods for each phase of this study.

4.2.1 Pilot Interviews

To inform the design and terminology used in our survey study, we first conducted pilot interviews with 12 DHH participants. In these interviews, we explained the concept of ATS-based reading assistance tools and asked them questions about situations in which they could envision using (or not using) such tools, as well as what they currently do when they encounter text they could not understand. This data allowed us to pilot-test the language we would later use for video demonstrations of the tool, as well as gather lists of reading purposes and workarounds to overcome complicated text, which we could use when preparing answer-choice options for similar items in the questionnaire for our survey
(section 4.2.2). A total of 12 DHH participants were recruited through e-mail and social media. Participants self-identified as male (N = 7) and female (N = 5), with mean age of 24 (SD = 1.5). There were 8 participants who identified as culturally Deaf \(^1\), 3 as Hard-of-hearing and 1 as deaf. Participants met in person with a research assistant and the interviews were conducted in English or ASL at the participants' preference. Participants were compensated with $40 for their participation. The analysis of these pilot interviews was primarily formative: Specifically, any interview questions that had required clarification during pilot interviews were edited for clarity when authoring related items on the survey questionnaire (section 4.2.2), and the open-ended responses from pilot study participants informed the list of answer choices for some questionnaire items. Full details of the survey questionnaire appear in section 4.2.2.

### 4.2.2 Survey

**Participants.**

Our participant-selection criteria included identifying as Deaf or Hard-of-Hearing, as well as having had work experience (including internships) in the computing or information technology within the past 5 years. Participants were recruited through social media posts, e-mail advertisements, and word of mouth, through the career center and alumni networks at our institution, as well as colleagues at tech companies and computing accessibility groups. Participants were offered the opportunity to enter into a raffle to win a $100 gift card.

We received a total of 32 responses (an additional 17 started, but did not complete it, yielding a dropout rate of 34%). Participants' mean age was 28.3 (SD = 7.9), ranging from 20 to 54. Participants self-identified as male (N = 18), female (N = 13) and agender (N = 1). The highest degrees obtained by participants included high school (N = 4), associates (N = 9), bachelor's (N = 13) and masters (N = 6), with 17 out of 32 participants indicating

---

\(^1\)As described in [91], Deaf with a capital “D” refers to people who identify as members of Deaf culture. We thus use this nomenclature throughout this dissertation.
Figure 4.1: Screenshots from the video demonstration shown to participants in the survey and follow-up interviews demonstrating (a) syntactic simplification and (b) lexical simplification. Simplified text is highlighted in yellow, while in (b), words identified as complex are highlighted in gray.

they are still students. Participants had an average of 5.5 years of work experience (SD = 7.37), ranging from less than a year to 32 years. There were 8 participants who identified as deaf, 12 as culturally Deaf [91], 10 as hard-of-hearing, and 2 as other (one indicated being “deaf with cochlear implants in both ears,” and one as being “deaf in only one ear, but hard-of-hearing on the other”). In terms of communication preferences, there were 6 participants who preferred spoken language only; 7 who indicated mostly spoken language, with a little sign language; 6 who preferred about half of each; 6 who preferred mostly sign language, with a little spoken language; and 6 who preferred only sign language. Lastly, most participants reported feeling very comfortable with reading English (N =18), while others indicated feeling comfortable (N = 8), neutral (N = 5), and not comfortable (N = 1).

Materials.

The survey, hosted using Qualtrics, consisted of 30 questions, and it required approximately 15 minutes to complete. Most of the questions were Likert-type items on a 5-point scale or
multiple-select questions (in which a participant could select more than one option among a set of choices presented). The questions were grouped by the following topics: 1) participants’ reading habits and their practices for learning on their own at work; 2) their experiences encountering difficult text and the workarounds used to overcome it; 3) their interest in ATS-based reading assistance tools in general and for specific activities that may involve reading; and 4) their thoughts on issues of autonomy and social acceptability in the context of the design of such tools. Before answering questions about ATS tools (group “3” of questions, as listed above), participants were shown a brief video demonstrating a reading-simplification tool, which showed a demo of both lexical and syntactic simplification. As ATS tools are not yet widely available to consumers, we created this video demonstration following the style of completed preliminary work (presented in Chapter 10) into the visual design of such systems. Figure 4.1 shows screenshots of this video demonstration, and the original video is shared in Appendix B – along with a complete copy of our survey questionnaire in Appendix C.

**Data Analysis.**

We calculated descriptive statistics (e.g. median, quartiles) for our ordinal scale data, as well as frequencies for data obtained from multiple-select questions. Furthermore, since our scalar-response data was not normally distributed, we conducted non-parametric statistical tests for difference testing, including Friedman and Kruskal-Wallis tests. Similarly, for correlation analysis, e.g. between user’s communication preference and their interest in the ATS tool in different contexts, Spearman correlation analysis was performed for this non-parametric data.

**4.2.3 Follow-up Interviews**

The final question in our survey asked participants if they would be willing to participate in a follow-up interview study, which we conducted with a subset of our survey respondents
who expressed a willingness to participate. Before presenting the survey results, we present here our methods for this interview portion of our study. Afterward, we will provide a combined results section of both the survey and the interview results, in an interleaved manner.

Materials.

These semi-structured interviews consisted of 30 questions, and they were grouped in similar categories as those of the survey, except in this case the categories were more open-ended in nature. The first category was about reading, which included questions about which activities participants read for, what they enjoy and do not enjoy reading about, as well as what they enjoy or do not enjoy about the activity of reading itself. We then asked participants about how they learn in the context of work, and how reading compares to other activities such as watching videos. Then, we asked participants about facing complicated text, what they believe affects someone’s reading skills and text difficulty, as well as how they personally try to understand complicated text. Questions about difficulty in reading were first posed in the third-person, due to the sometimes-sensitive nature of asking about literacy difficulty, under the assumption that respondents may be more willing to share their personal experiences after grounding it amid that of others. Finally, we showed participants the same video shown in the survey again and asked them questions such as the positive and negative impacts they could imagine the tool could have, as well as how they would feel if they either saw or were seen by co-workers using such tools. The full list of questions is shared in Appendix C.

Procedure.

A total of 7 participants who had responded “yes” to whether they would be willing to participate in interviews were randomly selected and contacted via e-mail, and 5 of these individuals responded to this request for an interview appointment. These 5 interviews
were conducted using video-conferencing because of social-distancing restrictions due to COVID-19. Participants were provided with informed consent forms ahead of the interview via e-mail. All of the interviews were conducted via video-call using Zoom and lasted 35 minutes on average, and were recorded for later reference with the participants’ consent. Four of the interviews were conducted in ASL by a researcher in the team who is hard of hearing and fluent in ASL, and one was conducted in English using the chat feature on Zoom. At the end of the interview, participants were compensated with $40 for their participation.

Participants.

Participants’ mean age was 28 (SD = 4.5), ranging from 27 to 37. Participants self-identified as female (N = 3) and male (N = 2). The highest degrees obtained by participants included associates (N = 1), bachelor’s (N = 3) and masters (N = 1), with 3 participants indicating they were still students. Participants had 3.8 years of experience on average (SD = 1.48), ranging from 2 to 6. There were three participants who identified as deaf, while the other two identified as Deaf. In terms of communication preferences, one participant indicated preferring mostly spoken language, two participants who preferred about half of each, but mostly sign language; and two preferred only sign language, one of which also specified preferring “written language.” Lastly, one participant reported feeling very comfortable with reading English, another one indicated feeling comfortable, two neutral, and one not comfortable.

Data and Analysis.

Of our five interviews, four had been conducted in ASL, and these were interpreted and transcribed by a researcher who identifies as hard of hearing and is fluent in ASL. Those four interviews amounted to a total of 119 minutes of video recording. Together with the fifth interview, conducted through chat, there were a total of 5,100 words of transcripts.
These transcripts were analyzed by one researcher in the team using an inductive approach by conducting a round of selective coding followed by axial coding to identify categories in their responses.

4.3 Results

In this section, we present the results of the survey and the follow-up interviews. Based on the results from both studies, our findings are grouped into five categories: 1) our participants’ reading frequencies and activities; 2) learning in the context of work; 3) participants’ perceptions of complicated text and workarounds to overcome it; 4) participants’ interest in ATS-based reading assistance tools; and 5) design considerations for such tools.

4.3.1 Reading

Reading Frequency.

Following the methodology of [68], we asked participants in the survey to report their frequency of reading on a 5-point scale of: “rarely (less than once a month),” “monthly (one to three times a month),” “weekly (once a week),” “often (two to four times a week),” and “daily (five or more times a week).” The majority of survey respondents reported reading at least once a week, including 16 who reported reading daily and 7 who reported reading two to three times a week. Another 3 participants reported reading one to three times a month, while five reported reading less than once a month. Similarly, four of the interview participants indicated reading often, except for P4, who indicated “not reading as often as I would like.” When comparing the reading frequency responses on this question to response data on similar types of questions collected from among the general U.S. population in prior work (as discussed in Section 3.2), our participants’ frequency of reading was relatively similar.
Figure 4.2: Participants’ responses to the questions: On a day that you read, how much time do you spend reading text (a) on a screen (e.g. computers, laptops, phones, tablets, etc.) and (b) that is not on a screen (e.g. books, magazines, newspapers, print-outs). There was a significant difference between the two (p < 0.01).

**Reading on an Electronic Screen.**

Survey participants next estimated the number of minutes spent reading, on a day that they read. However, in order to investigate how much of their reading happens on a screen (where ATS-based reading assistance tools are typically deployed), we asked for this estimate in two separate questions: how many minutes a day spent reading on a screen (e.g. computers, laptops, phones, tablets, etc.) and not on a screen (e.g. books, magazines, newspapers, print-outs). As illustrated in Figure 4.2, 25 survey respondents reported reading over 30 minutes a day on a screen, 15 of which reported reading over 60 minutes a day on a screen. In comparison, most survey respondents (N = 21) reported reading text not on a screen...
less than 15 minutes a day, with 7 indicating spending no time at all. A Wilcoxon signed rank test indicated a significant difference between the reported time reading on a screen and not on a screen ($Z = -3.93$, $r = 0.695$, $p < 0.000$).

The data from the interviews showed a similar trend, with three out of five participants explicitly saying that they tend to read more “online,” or in the words of P5, “I read mostly on the computer.” Some of the sources for content online cited by participants included blogs on platforms such as “Medium,” forums on platforms like “Reddit,” as well as online newspapers and social media.

**Purposes for Reading.**

We then asked participants in the survey about the purposes for which they read, by selecting as many items as they wish from among the following list, which was obtained from our pilot interviews (section 4.2.1): work (e.g. technical text), academic (e.g. research papers, scientific articles, class, exams, textbooks), medical (e.g. health insurance, diagnosis), legal (e.g. terms of service, contracts), personal communication (e.g. e-mail, text messages,
social networks), visual media (e.g. movies, tv shows), personal reading (e.g. books), recreation (e.g. restaurant menus), and news (e.g. newspapers, magazines). Finally, a write-in “other” option was provided so they could mention other purposes for reading. Work was the purpose that survey respondents reported most often (N = 29), with academic (N = 26), personal communication (N = 25), and visual media (N = 24) coming close behind. The least selected options were medical and legal, with 8 respondents each. Figure 4.3 summarizes the frequencies for all options.

Interview participants drew distinctions between “personal reading” as compared to reading for work or academic purposes. Two interviewees reported mainly reading for work (P1 and P4), with P12 commenting: “With regarding to reading (online), I use for work stuffs and learning things to improve my career (programming for example).” Another two interviewees mentioned reading for academic purposes, with P2 conditioning her frequency of reading on whether she is at school, saying, “If I’m in school I tend to read weekly almost every day.” All interview participants, in turn, mentioned different forms of personal reading out of interest or curiosity, such as reading “storybooks to enjoy” (P1), or “technology, bible, something that interests my curiosity like martial arts” (P4). Two other participants mentioned being motivated to read to be aware of what is going on in the world, e.g. with P3 commenting “since this situation (COVID-19) we have to read emails, social media every day, all day.” Finally, two participants also specified that, for personal reading, they prefer reading content such as stories “with simpler words” (P2) or “that are simple like thriller novels and not like Shakespeare” (P1).

4.3.2 Learning for work

In order to further understand the importance of reading in the context of work, the survey asked participants to rate how often they read to learn about technical topics at work, as well as how often they watched videos for the same purpose. The rationale for including this question about watching videos is the increasing prevalence of online viewership [13], as well as using this set of two questions to help further quantify the amount of reading
Figure 4.4: Participants’ responses to the questions: “How often do you read to learn about technical topics for your work (e.g. information about new technologies, new software, programming information, etc.)?” and “How often do you watch videos to learn about technical topics for your work (e.g. information about new technologies, new software, programming information, etc.)?” The full text in the options was: “rarely (less than once a month),” “monthly (one to three times a month),” “weekly (once a week),” “often (two to four times a week),” and “daily (five or more times a week).” There was a significant difference between the two (p < 0.05).
these users engage in (by comparing it to frequency responses they provide for some other activity, namely watching videos). Both of these questions were on the same 5-point scale used for the general reading frequency above, going from rarely to daily. As shown in Figure 4.4, the majority of survey respondents (N = 23) reported reading at least once a week for learning about new topics at work, 9 of which reported doing so two to three times a week, and another 9, daily. The majority of participants (N = 18) also reported watching videos to learn about new topics at work at least once a week, with 9 reporting doing so two to three times a week, and 3, daily. A Wilcoxon signed rank test indicated a significant difference between the reported time reading and watching videos to learn about technical topics at work (Z = -2.089, r = 0.369, p = 0.037).

We asked about similar topics in our interviews. There were idiosyncratic differences in terms of whether interview participants preferred reading vs. watching videos overall, with two participants explicitly saying they prefer videos in general, one saying they typically prefer reading, and one rather simply stating it “depends on my mood” (P5). However, when talking about specific situations in which they would prefer one or the other, most participants indicated preferring videos for learning new things, topics that are unfamiliar, that are practical, or in the words of P4 “when not a lot of thinking is required.” Participants indicated preferring reading over videos when they are already familiar with or passionate about topics, or when they are reading concepts that are more technical in nature or that are “hard to memorize” (P1).

4.3.3 Complicated text

The survey asked participants to rate “how often do you encounter text that is complicated” on a 5-point scale from “Never” to “Very often”. There was a wide range of responses to this question: The median response was “neither seldom nor often” (14 participants). The lower quartile response was “seldom” (8 participants), while the upper quartile was “often” (9 participants).
Interview participants indicated text being “hard to read” is something that they do not enjoy about reading. For example, in the words of P2, “I don’t like to read theory related readings because they’re too hard to read.”

Two participants (P1 and P3) used Shakespeare as their examples of what complicated text looks like, with P1 specifically mentioning Shakespeare text as something she does not like to read. When discussing what makes a text complicated, all interview participants mentioned vocabulary or terminology that they are unfamiliar with as one of the main sources of difficulty. As P4 put it: “Because sometimes you are reading and understanding, but all of a sudden there’s a word that makes you lose your train of thought. You have to stop there, analyze what it means, look it up and then look at the reading to figure out where you left off to continue reading.” Furthermore, three participants associated negative feelings with complicated text, two of which mentioned feeling frustration (P4 and P5), and P2, a loss of confidence: “I felt pressured to use that same level of English [as my classmate], so it takes me more time to read and write. I really didn’t like that pressure. It also caused me to lose confidence in class. There is a lot of discussion in class and I would feel like my classmates were smarter than me.” Finally, some participants seemed to quantify complicated text in terms of the time it adds to reading. P5, for example, stated that “I consume so much time trying to figure out what a word means. Then, once I figure it out, I continue to read.”

Overcoming complicated text.

Our survey participants were also asked to indicate resources they typically use to try to understand text that is complicated, selecting as many items they wish from the following list: a dictionary, looking for a translation to American Sign Language, asking coworkers for help, asking a supervisor for help, and looking for other websites talking about the same topic. Finally, a write-in “other” option and “this doesn’t apply to me” were provided. The composition of answer choice options on this list was informed by several sources, including our pilot interviews (discussed in section 4.2.1) as well as considering whether there exist
external tools analogous to internal metacognitive strategies employed by DHH readers who encounter complex text [10]. As shown in Figure 4.5, the most frequently selected response was looking up words in a dictionary (N = 25), followed by looking for other websites talking about the same topic (N = 21). The least frequently selected options were looking for a translation to ASL (N = 4) and asking a supervisor for help (N= 3).

In our interview study, “looking up” words was mentioned by all interview participants as a way to overcome complicated text. Notably, at least three participants specified that these “look-ups” may consist a Google web search for the word, rather than using a specific dictionary resource. As P4 said: “If I have dictionary, I use it or just do a quick research on Google.” Furthermore, four participants mentioned asking others for help, including friends, interpreters, co-workers, supervisors and professors. P3 indicated that her asking someone else for clarification depends “on where I am and my surroundings.” Notably, P1 mentioned trying to understand some text first before asking others, especially more senior colleagues, because “I hate to ask ’bigger’ people since it brings embarrassment to me.”
Figure 4.6: Survey respondents’ agreement to a Likert-scale question “I would be interested in a tool that helps me to understand text by making it simpler” (a) overall and (b) broken down by communication preferences. Communication preference levels: Spoken = “Spoken language only”, Mostly Spoken = “Mostly spoken language, with a little sign language”, Half = “About half spoken language, half sign language”, Mostly Sign = “Mostly sign language, with a little spoken language”, and Sign = “Sign language only.”

4.3.4 Interest in tool

We showed participants a video demonstration of a prototype ATS-based reading assistance tools, using the two most common approaches in the ATS literature: lexical and syntactic simplification. The video was approximately two minutes long and included demonstrations of using the tool for both of these approaches; a copy of this video is included as an electronic supplementary file. We then asked participants to indicate their agreement to the statement “I would be interested in a tool that helps me to understand text by making it simpler” on a 5-point Likert scale from “strongly disagree” to “strongly agree.” As illustrated in Figure 4.6(a), the overall response was positive, with a median answer of “strongly agree,” with 25 participants responding at least “somewhat agree,” out of which 19 responded “strongly
In our interview study, the same video demonstration was shown to all 5 participants. All indicated that the main benefit they envision an ATS tool would provide would be saving them time. As P1 put it, “It would speed up my reading pace that is all-important.” Three participants also mentioned not having to ask others for help or clarification as a benefit of the tool. P2, for example, stated “It would help us to read easily without asking others to help.” Other benefits mentioned by participants included learning new words (P2 and P4), as well as reducing frustration (P5) and increasing confidence (P1). Notably, three participants explicitly mentioned that the benefits would not be limited to DHH readers, but also, as P1 put it, “hard readers (Deaf and other disabilities and non-English speaking).”
CHAPTER 4. READING EXPERIENCES AND INTERESTS IN ATS

**Tool Interest and Communication Preference.**

Considering that prior work with DHH users in other domains (e.g. sound-awareness technologies [41] or sign language animations [61]) has identified that users’ communication preferences (i.e. sign-language vs. spoken) may influence their interest in or opinions on certain technologies, we wanted to investigate whether that was the case here as well. Survey participants’ response to the question about communication preference was coded using a scale 1 (spoken); 2 (mostly spoken, some sign); 3 (half and half); 4 (mostly sign, some spoken), and 5 (sign). A Spearman correlation test revealed a significant correlation between users’ reported communication preference and their interest in the tool (rho: 0.5, p = 0.0034), with participants who preferred sign language reporting higher interest in the tool, as illustrated in Figure 6(b). A Kruskal-Wallis test, however, did not reveal any significant difference between the groups.

In our interview study, when participants talked about issues that affect people’s reading levels and their perception of difficult text, all participants mentioned English not being their first language: “I struggled and it was hard for me to pick it up, and for Deaf people ASL is their first language, and English is not their first language. ASL is completely different” (P5). Furthermore, participants talked about people’s upbringing or exposure to English – either by listening or reading – as another source of difficulty. P3, who identified as a strong reader, said “I am from a 5th generation Deaf family, but my grandma was strict with English and forced me to learn English all of my life. I am grateful for her, so really the main point is the people in your life and how involved they are and how they empower you.” Thus, our question related to “communication preference” may have correlated with interest in the tool because it may have shed some light as for which language participants feel more confident with.
Tool Interest and Purposes for Reading.

Survey participants were also asked whether they would be interested in a tool like the one in the video for each of the purposes for reading (the same list we had provided previously in the Reading Habits section), using a 5-point Likert-type question from “Not Interested” to “Extremely Interested”. A Friedman test did not indicate a significant difference between the responses when comparing the various purposes for reading. Figure 4.7 indicates that the responses that received the most “extremely” or “very” interested responses from our respondents were: academic, medical, legal and work.

In our interview study, while all participants mentioned they would personally use the tool, four mentioned they would not use it for texts that are written informally or using “basic English.” As P1 commented, “Any content that are already in basic English or written in informal talking style [the tool] isn’t needed.” These comments suggest that participants’ interest in using the tool when reading for various purposes may be related to their estimate of how likely they would be to encounter complex English text when reading for that purpose.

4.3.5 Design considerations

Autonomy.

In Part III of this dissertation, we thoroughly explore autonomy as a design aspect that may affect acceptability of these tools, where autonomy is framed as whether the tool gives user control over which texts were simplified (vs. doing so automatically) and whether the system makes it clear whether the user is looking at an original or a simplified version of some text. However, in our survey study, participants were preliminarily asked to imagine using such a tool (after viewing the video demonstration) and to respond to two questions related to the autonomy the tool would provide: “I would be upset if the tool replaced text before I got to see it” and “I would be upset if the tool replaced text without asking me.” The majority of survey respondents reported they would be upset if the tool replaced text
before they got to see it, and also that they would be upset if the tool replaced text without asking them, with median agreement responses of “strongly agree” for both question items, and more than 75% of respondents responding at least “somewhat agree.”

In our interview study, autonomy was not a topic that arose directly, as this would be explored in Part III of this dissertation. However, one participant mentioned autonomy as an expectation: “When the website loads, I don’t expect to immediately translate. It is unto the people’s preference to choose to switch if they like” (P1). However, two participants brought up a related issue: P3 highlighted as a downside of the tool that there would be “additional clicking to do.” P1, in turn, suggested that “Since the system works on sections rather than the whole paragraphs, it requires switching every paragraph which is complicated. I would like if the whole website translates the content to ‘simple language.’” These comments relate to a specific aspect of autonomy in the design of user interfaces for these tools: how much text is replaced at once upon the users’ request.

Social Acceptability.

Considering issues of social acceptability in assistive technology design [116], we also asked participants to rate whether they would be embarrassed if a colleague saw them using a tool like the one in the video, which may shed light into whether the visibility of such a tool should be further investigated. The majority of survey respondents responded either “somewhat disagree” or “neither agree nor disagree,” with a median response of 2.5 (between “somewhat disagree” and “neither agree nor disagree”).

When looking at the interview data, however, three interview participants indicated that the social acceptability of the tool may be dependent upon the environment, specifically on co-workers understanding of the users’ situation. As P2 puts it: “I would feel ashamed. If we all have the same problem with English then I would be fine but if people whose first language is English, they probably wouldn’t understand why or look down upon or judge me so I wouldn’t want to use it in that environment.” P5, in turn, expected some judgement
from co-workers: “Maybe they would think he’s really on a tool? Or be impressed because I’m using a tool as an effort.” On the other hand, P4 thought that it was “not about how I feel if it helps me then why not,” or P1 who thought personal values were more important than what co-workers think: “I don’t mind either too [if a co-worker uses it]. I care that the information should be readable and accessible.”

Accuracy.

A topic we did not specifically ask about in a question in the survey, but which several participants mentioned during interviews was that the accuracy of the tool was a key concern of these users. More specifically, the possibility of the tool “causing people to misunderstand the word if it’s not the right meaning or replacement word” (P2), or as P3 put it: “Maybe the tool could misinterpret the word. The wording in the sentence could mean one thing while the tool may interpret it to something else. Whatever the coding or something could misunderstand the true intent of the sentence.”

However, when asked about how the perceptions of the system would change if the system was not 100% accurate, interview participants reported they would think the tool is still useful if the tool meets a certain threshold of accuracy, with two participants estimating that threshold to be “90%.”

4.4 Discussion

Our results suggest that DHH individuals with work experience in the computing field read often, mostly on a screen. While participants also reported reading for personal purposes, our results suggest that a lot of their reading relates to computing-related topics, as they are reading for work or academic purposes. By means of comparison with another activity our participants engage in (watching videos to learn about computing topics), our results suggest that they do read a lot to learn about computing topics. While videos may be preferred for learning about completely unfamiliar topics at a high-level, our results suggest
that participants still read more often than watching videos when investigating topics with which they already have some deeper familiarity. Considering that NLP researchers have suggested there may be benefits from training systems based on judgements from specific user-group on specific domain [80], these findings thereby motivate further research into ATS-based reading assistance tools for DHH computing professionals as a potential user group.

Our results indicate that while participants do not report facing complicated text very often, when they do face it, it affects their enjoyment of reading. Participants quantified complicated text in terms of the time it takes to overcome it, and difficulty with complex or unfamiliar words was a key source of difficulty that participants reported. Thus, it is not a surprise that the most frequent workaround to facing complicated text was “looking up” words, which closely parallels the solutions that lexical approaches to ATS would provide. Further, the second most reported workaround was looking for an alternative text, which in turn parallels the solutions syntactic approaches to ATS would provide. Notably, however, our results suggest that while there may be some openness to asking others for help or clarification, our participants prefer to attempt to overcome it on their own before asking others and thus ATS tools may be helpful to avoid asking others. Furthermore, while our participants expressed interest in ATS tool, our results highlight that there were no commercially-available ATS tools in use by this user group as has been suggested by [111] for the general population.

We also found that there is a lot of interest overall in having a tool to assist with complicated text by making it easier to read. We note that it may be worth further exploring the relationship between this interest and users’ spoken-vs-sign communication preferences or what they consider as their first language. The main benefits participants envisioned from a tool like this were related to saving time and not having to ask others for help. Interest varied depending on the reading purpose, with participants indicating interest for such tools when reading for work, academic, medical, or legal purposes. Notably, in our results on frequency of reading for each of these purposes (section 4.3.1), work and academic
had been frequent purposes for reading, with medical and legal as more rare purposes for reading, suggesting that indeed ATS tools applied to texts in the computing domain could be explored by NLP researchers for this particular user group.

Our results highlight three aspects of the design of these technologies that may be worth further exploring. First, how much text is transformed upon a single user’s request (i.e. only one word, one sentence, one paragraph, the full text) emerged as a concern, which is further explored as a design parameter in Part III of this dissertation. The less text that is replaced per request, the more effort it requires from participants. This is independent, however, from whether the simplifications themselves happen at the lexical or syntactic level (i.e. whether only words are replaced, or sentences are rewritten too), which the present study did not explore. Thus, this relationship between what is replaced and how much is replaced at once is important to explore. Second, our results suggest it is worth further exploring how to mitigate the social acceptability of these tools, since that may vary depending on the environment users are using it in. Lastly, our results indicate that the relationship between the accuracy of the tool – specifically the meaning preservation of the transformations – and its usability should be further investigated.

4.5 Limitations and Future Work

There were several limitations of our study: Because we conducted an online survey, we had to rely on time estimates from participants as a way to learn about their reading habits, which prior research has shown provides only a glimpse into people’s reading habits – since reading is a complex social phenomenon that looks differently in different contexts [68]. While we tried to mitigate this by comparing reading specifically in the workplace against another activity (i.e. watching video), it is difficult to obtain accurate estimates of every reading activity our participants may engage in.

A second limitation was the sample size of our study: Because we were looking at a specific user group, i.e. DHH adults with work experience in the computing field, our sample
size was not large enough to support fully investigating whether there may be a relationship between users’ interest in reading assistant, and with various demographic factors, e.g. participants’ communications preference. In future work, research with a larger population of DHH in the U.S. could look deeper into this issue by building regression models with the demographic factors. Similarly, research with the larger DHH population in the U.S. could examine how their interests and needs may differ from those in the computing industry.

Our study design was based on survey and interviews, which included a brief video demonstration of reading assistance tools. We had asked participants to imagine using the such tools, to gain insights into their views on various design issues. Thus, this hypothetical nature of the questions is a limitation that is addressed in Part III of this dissertation in which participants actually engage with a tool. However, there are critical questions that emerge from our results regarding issues of social acceptability, autonomy and accuracy, which are explored in subsequent chapters of this dissertation through interviews and usability studies with interactive prototypes. While our video had shown both lexical and syntactic approaches to text-simplification, we did not compare these approaches in this chapter, as we were concerned that our short video may not provide sufficient context to differentiate these approaches. Future work can also include usability studies that focus on the comparison of both of these approaches. Furthermore, the color coding used in the video demonstration for complex and simplified text (gray and yellow, respectively) was based on suggestions from DHH participants during pilot studies with a prototype; however, future work should also ensure that the color coding used for prototypes deployed in the wild is accessible for users with color blindness.

While our work has examined this specific user group, in a specific domain, future studies could investigate the needs and interests of other user groups and in other domains, who may also benefit from ATS-based reading assistance tools. Finally, because we were focusing on DHH computing professionals, our participants’ self-reported English levels and level of education may not be fully representative of the entire DHH population in the U.S. Thus, future work may focus on the interests of DHH adults with lower levels of literacy or
education, to understand how their views may differ.

4.6 Conclusion

Through a survey and follow-up interviews with a subset of survey respondents, our study investigated the needs and interests of DHH individuals in the computing field for ATS-based reading assistance tools. Our results suggest that DHH individuals read often, frequently on electronic devices and to learn about new topics for their work, and indicated strong interest in ATS-based reading assistance tools. Our results also include a prioritized list of the most frequent workarounds our participants currently use for overcoming complicated text, with looking up words and finding other texts with the same content, which are analogous to typical ATS approaches, being the most frequent ones reported by participants. We also provide a prioritized list of reading purposes for which participants reported interest in using ATS-based reading assistance tools, as well as their frequent purposes for reading. These findings thereby motivate further technical work on such tools for this user group, which may require gathering user-and-domain-specific datasets for this setting, as needed by NLP researchers. Finally, our results provide insights into certain design considerations for ATS-based reading tools, namely expanding the user autonomy they provide, and highlighting participants’ concerns about ATS accuracy and the social acceptability of these technologies. These findings, in turn, motivate further design work into such tools for this particular user group.
Chapter 5

Social Accessibility of ATS-based Reading Assistance Tools

5.1 Introduction

Based on our results from our study summarized in Chapter 4, we wanted to further investigate one of the design considerations outlined above, namely the social accessibility of the tools, given the inconclusive results from our survey and interview results. Thus, we conducted a new interview study with 7 participants in which we focused only on issues related to social accessibility specifically in the work context, which include participants’ personal values in a professional context (e.g. how they want to be seen by their colleagues) and how those may relate to using ATS-based reading assistance tools. Further, we investigate how they perceive the use of these tools, as well as how they would expect their colleagues to perceive it, and how those expectations may affect their view of the social acceptability of using ATS-based reading assistance tools in the workplace.
5.2 Method

5.2.1 Materials

The questionnaire for these semi-structured interviews consisted of 20 open-ended questions, which were grouped into four sections. In the first section, we asked participants about their work experiences (e.g. what type of work they do and how long they have done it for), as well as their reading experiences (e.g. how much they read and what they like to read), as a way to get them grounded on these experiences for our later questions in which we would ask them to imagine using reading assistance tools in a workplace environment. We then asked participants questions about what they value at work and how they want to be perceived by their colleagues in a professional environment. Then, we asked participants more specific questions about reading and reading skills, which included questions about how they would describe the diversity in literacy among DHH readers, as well as whether they personally use tools to support their reading tasks. Finally, after showing participants the same video demonstration used in the survey and the previous sets of interviews, we asked participants more directly about their perceptions of using ATS-based reading assistance tools both from a personal and a social perspective. On the personal perspective, we first asked them questions about what they thought about the tool. Because we had observed some people express some reservations about the label “simplification” in our pilot interviews for the study described in Chapter 4, we also asked them about their thoughts on those tools being called “simplification” tools and, if they were not hesitant themselves, why they thought some people may be hesitant to use a tool with such a name, as well as what alternative names they thought may help mitigate that hesitancy. On the social perspective, our questions included what they would expect others to think if they were seen using ATS-based reading assistance tools as well as how they themselves would feel if others saw them. Because in the results from our study summarized in Chapter 4, some participants expressed they might feel embarrassment, in these interviews we asked participants why they thought someone would feel embarrassed if seen using these types of tools, if they would not feel
so themselves. Finally, we also asked participants for ways that could help mitigate that embarrassment. The full interview questionnaire for this set of interviews is also included in Appendix D.

5.2.2 Procedure

As our recruitment criteria remained the same as in 4, we reached out to 13 survey respondents who had indicated willingness to participate in follow-up interviews but had not participated in our previous set of interviews, and two of them participated in this new set of interviews. We recruited another five participants through posting advertisements in social media groups of DHH users, for a total of seven participants. Participants were provided with informed consent forms ahead of the interview via e-mail, and met with a researcher via video-call using Zoom. The interviews lasted 35 minutes on average, and were recorded for later reference and transcription, with the participants’ consent. Two of the interviews were conducted in ASL by a researcher on the team who is hard of hearing and fluent in ASL, and the rest were conducted in English using professional captioners. Similar to our previous set of interviews, at the end of the interview, participants privately filled out an anonymous demographics form which included questions about their communication preferences and comfort in reading English. Participants in this study were also compensated with $40 for their participation.

5.2.3 Participants

The average age of our seven participants was 34 (SD = 12.7), ranging from 21 to 55. Participants self-identified as female (N = 3), male (N = 2), agender (N = 1) and non-binary (N =1). The highest degrees obtained by participants included high school (N = 2), bachelor’s (N = 3) and master’s (N = 2), with 3 participants indicating they were still students. Participants had an average of 10.6 years of experience in the computing field (SD = 10.6), ranging from less than one year to 30. There were 2 participants who
identified as deaf, 4 who identified as hard-of-hearing, one who identified as both deaf and hard-of-hearing, and one who identified as culturally Deaf [91]. In terms of communication preferences, 3 participants preferred spoken language only, 1 participant preferred spoken language with a little sign language, 2 participants who preferred about half of each, and 1 preferred mostly sign language, with a little spoken language. Lastly, 5 participants reported feeling very comfortable with reading English, and another 2 indicated feeling comfortable.

5.2.4 Data Analysis

Our seven interviews yielded 245 minutes of recordings. The interviews conducted in ASL were transcribed by the research assistant in our team who conducted the interviews, and for those that were conducted using professional captioners, the captions provided were saved after the interviews. Combined, these transcriptions amounted to a total of 19,293 words.

This data was then analyzed by three researchers in our team using a thematic analysis approach as described in [19] by first doing a round of open coding independently and identifying categories of codes, followed by a few rounds of discussion to refine the codes and identify themes from the data.

5.3 Results

After conducting a thematic analysis on our data, we identified four key themes related to our research question. First, we found that participants described wanting to be seen by others in their workplace as competent and reliable, and these perceptions are important to them, especially so in a workplace or academic setting. Second, participants described a potential perceived relationship between literacy skill and cognitive ability or intelligence. Thus, being seen using an ATS tool, which may imply literacy issues, could come in conflict with their desired image of competence given that others may relate lower literacy to incompetence. Third, participants shared how text simplification technologies may not
only be for DHH readers, but described shared experiences with other user groups, such as people whose first language is not English. And finally, we found that framing these technologies as something more mainstream than an assistive technology, as well as making the user-interface of these technologies more discreet or easily hidden, may help mitigate some of the issues identified.

5.3.1 Values and Expectations

When we asked about how they want to be seen in the workplace, most participants commented on wanting to be seen as “competent” (P1), knowledgeable (P5: “someone who knows how to get the job done”), or skilled (P4: “I think everybody wants to be seen as a skilled reliable worker”). Participants also want to be seen as reliable and responsible (P5: “I want them to think that I am a reliable person”), or someone who can be trusted to get the job done (P7: “someone who is dedicated and trusted with a job duty. To see it to completion”).

When we talked about what a good working environment would include, most participants talked about clear communication and a good sense of community in which people are accommodating to their needs, but where the support is mutual. For instance, P4 listed communication as having the topmost importance in the workplace. P1, in turn, described a good working environment as one in which “people are caring for each other and having a sense of community” and “where respect is both ways.”

5.3.2 Reading skill and intelligence

While at first none of the participants indicated they would feel offended by the recommendation of using ATS tools or embarrassed if seen by others themselves, when we asked them why they thought a third person could potentially feel those ways, most of their responses related to how literacy skill may imply a cognitive or intellectual disability, or be interpreted as the user being less capable, which could result on people looking down on
them. Thus, if someone recommends an ATS-based reading assistance tool, they may be interpreted as implying that they believe the person is less capable, or as P1 puts it “If someone recommended to a person and that person may feel they are not being viewed as smart because they are implying that they have a low IQ.” If a colleague sees someone using ATS-based reading assistance tools, in turn, the colleague may interpret that as a lack of intelligence too and look down on them. P2 puts it as “it’s almost like an indication that I can’t read the original wording and that I need a text simplifier. Maybe they would look down on me that I’m not able to read,” while P6 said “They might be like this guy’s not that smart.”

These perceptions are important because they may come in conflict with participants’ desires to be seen as competent, knowledgeable and skilled. This may be especially amplified in a computing workplace, where open offices are now more common, as P3 puts it “I think having a level of privacy is important, like open workspaces are a trend but it is important to respect the work space”. P2 also talked about how using it in a workplace “seems public,” because “if I have my monitor up here and then I am reading something on the screen, people around me can see. So it’s the public aspect of it.”

The feelings described may also be further amplified by cultural expectations on reading and intelligence, as P4 describes “Intelligence and reading comprehension and everything is a big thing in our culture.” However, while participants expected others to think this way, they also were clear that literacy skill and intelligence are not related. For example, when talking about why someone may feel offended by the recommendation to use an ATS-based reading assistance tool, P5 said “because it would make them feel like their intelligence is being questioned […] but it’s nothing to do with how smart they are, it is just focusing on the English literacy.” In addition to clarifications indicating that reading skill is not necessarily related to intelligence, P5 also argued that the reasons why a DHH person may struggle with reading certain texts or why ATS tools could be needed are not because of being DHH nor a lack of intelligence, but because of societal or structural issues instead. More specifically, P5 explained how “the issue starts with the education system,” arguing
that:

“The problem is because there are some deaf and hard-of-hearing people who can read and write very well. While others may not because some deaf hard-of-hearing teachers prioritize lip reading, producing speech and not focusing on English writing and reading itself. Some people may not teach them proper English and Reading Writing and just focus on learning how to speak.” - P5

The relationship between reading skill and intelligence was also present when participants talked about their thoughts of labeling the tool as a “simplification” tool, with 3 participants comparing it to “dumbing down.” While P2 thought that “simplification is a better word than dumbing down obviously,” P5 felt that they sounded alike, and P1 worried that it could have that connotation “when it is not really happening.” P3 also commented on how the technology may feel like “not giving information but also spoon feeding and that could be offensive.”

We also observed in our data that participants’ level of comfort with others seeing them using ATS-based reading assistance tools may be affected by who the onlooker is. For instance, some participants expected family and friends to be less judgemental, like P5 who said “I would feel they would be less judgmental about it than people I work with.” However, in the workplace, it may be more of an issue if the person who sees you using it is your boss or a supervisee. For instance, P2 said “I would have to justify to my boss why I’m using this tool.” P6, on the other hand, would expect feeling embarrassed if seen by a supervisee “because I am the boss.”

Finally, it is also important to note that many participants commented on how either they actually did not, or other people should not, care about what others think when using these kinds of technologies – especially when those technologies provide utility to the user. For example, using hearing aids as an example for a device people may be embarrassed to use, P4 commented “I give very little craps about what people think [...] if I could use
hearing aids I would probably wear them [...] I still wouldn’t care what anybody says or thinks.” P7, in turn, focused more on how other people may be busy and how “nobody really cares what the other is doing.” But some of these same participants also suggested it was not always this way. For example, as the reason for why P3 does not worry about what co-workers think, P3 mentioned “I have been through a lot.” Similarly, P4 expects younger people to be more self-conscious, which suggests it may have taken time to get to the level of comfort where one does not care about what others think. However, even when someone may be comfortable enough not to care about what others think, there may be negative feelings involved depending on particular connotations about the technologies being used and, as mentioned above, who is watching. For instance, P4, who commented on a level of comfort being seen using assistive technologies, still had some reservations about using them in the context being seen by a supervisee if the tool has a negative connotation: “If someone that I’m in charge of sees me using a particular tool that might have a negative perception or connotation or something with it, that would really bother me.”

5.3.3 Not just for DHH readers

While many participants thought these tools would be useful, many also talked about how these tools may not just be useful for DHH readers, but other user groups may benefit from using them as well. P4, for example, was even “perplexed” by our study’s focus on DHH readers. However, most of the other participants’ comments seemed to relate to people who use English as a second language. For example, P5 talked about how “my mom whose first language is Chinese and second language is English would really benefit from this tool because she could improve and work on her English skill.” P5 also talked about how this may be beneficial for international students: “a lot of international students in graduate schools have to read a lot and it can be very tiring.”

While not all of our participants were sign language users, participants seemed to describe a shared experience between those for whom English is their second language and ASL users, like P5, who said “Some deaf hard-of-hearing people or people of color or immi-
grants their parents may not know English either and so they may know another language or maybe they just only know how to communicate in ASL so they don’t have the opportunity to practice English so that could impact their reading.” P5’s acknowledgement of how many DHH people may use ASL as their primary language and English only as a second language may help explain why participants’ talked about these shared experiences. However, P2 also felt that these tools could even benefit hearing people and children, and P5 also included people with intellectual disabilities.

P4, in turn, believes that complicated text is actually the author’s fault. Thus, P4 believed that using ATS tools for DHH readers was a “misguided approach,” but instead these should be tools for authors to make their writing more accessible. In P4’s words:

“I think your approach needs to be more than possibly similarly to grammarly or spell check for something on this could be a tool for the author or editor of an article to make their documents more accessible from the start, rather than require the end user, to use the tool.” - P4

5.3.4 Mitigating Social Accessibility Issues

In addition to participants’ suggestions that these technologies may not just be useful for DHH readers, some participants also used this same idea as a way to mitigate potentially negative perceptions of using these tools: framing or marketing these technologies in a way that does not feel like an assistive technology, but rather as something that is useful for the general population. For instance, P4 said it is “just something that anybody can use so if you ever make a tool like this it definitely needs to be marketed as for anyone.” Participants’ recommendations for alternative ways of naming these tools also seemed to reflect this, with P5 indicating that this is more of a “plain language” tool. Others, instead of focusing on complexity, thought of the tool more as a translator (P1), to a more “layperson’s English,” as a way to “un-jargon” text (P3), or to “paraphrase” (P5 and P7) or “summarize” (P5) text. Others focused on the efficiency aspect of the tool, including P6, who suggested
CHAPTER 5. SOCIAL ACCESSIBILITY

naming it an “efficient studying tool” because “it sounds more productive and intelligent this way;” P3, who suggested calling it “ez-pz;” or P7, who suggested calling it “easy read” or “quick grasp.”

As a way to help mitigate the potential issues with using ATS-based reading assistance tools in a public space, many participants suggested designing the tools in a way that blends in with other existing tools. For example, P1 suggested that “making it look like the native text would make it less embarrassing” or if the software gives an option to “simplify the entire webpage when it is loaded so the user won’t have to manually click to simplify in front of others.” P3 recommended making it available “seamlessly like I was not using the tool, maybe [changing] the highlighter color that you use,” but making it configurable on demand. Other participants, like P3 and P6 recommended providing ways to switch back to the original text or quickly hiding the tool. For instance, P6 said “if the person really feels embarrassing, there should be a way to hide the tool fast (like a shortcut on a keyboard)” and P4 also commented “a keyboard shortcut should be fine.”

5.4 Discussion

When considering issues of social accessibility, our study revealed how a potentially prevalent misperception of a relationship between reading skill and intelligence was at the heart of the potential social acceptability issues with this technology. While this perceived relationship may not a unique concern to DHH readers, potential issues arising from its perception had not been previously identified in the context of ATS-based reading assistance tools. Our results suggest that while our participants may be comfortable being seen as DHH, and even if it takes time for many individuals to reach that level of comfort, few were okay with being seen as incompetent or as lacking intelligence. Misperceptions regarding the abilities of users of assistive technologies echo the findings of prior work social accessibility work [115], which identified potential misperceptions of incompetence that can arise if a user appears unable to use an assistive technology. Thus, those misperceptions of incom-
petence relate to breakdowns in the use of assistive technologies, i.e., situations in which the assistive technologies failed to function properly. In contrast, our results suggest that, in the context of ATS-based reading assistance tools, DHH users believed they could be seen as incompetent simply by using ATS-based reading assistance technologies, even if the tools function properly, because of the perceived relationship between reading skill and intelligence.

While prior work had identified potential ways of mitigating social accessibility issues in assistive devices or products by making them resemble more “mainstream” devices (e.g. [115]) or framing them as something other than “assistive technologies” (e.g. [96]), our work sheds light on potential solutions for social accessibility issues in the context of software tools that may be used in a workplace environment. While participants’ suggestions for characterizing these tools as “productivity tools” parallel suggestions for designing devices that resemble “mainstream” devices, the focus on productivity here provides more insight into the reasons DHH users identify as motivating the use of ATS-based reading assistance tools stemming a desire to read efficiently rather than simply identifying as DHH. Furthermore, our results also suggest that it may indeed be important to design ATS-based reading assistance tools in ways that blend in with other existing software tools, that allow them to be hidden quickly, or to be loaded for longer chunks of text at once.

These possibilities for potential design solutions need to be considered in the context of prior work in ATS which had involved designs that provide simplifications for limited amounts of text at a time in ways that are also visible (e.g. [14, 99]) which, according to our results, may present socially accessible issues. However, as the autonomy of these tools is also important to consider in the design of assistive technologies (as explored in Chapter 10), there may be tensions between potential desires for visual indications to provide autonomy and the ability to have these tools blend in or be hidden easily for social accessibility, which should be further explored.

Considering participants’ suggestions for mitigating social accessibility issues, and how these may conflict with participants’ preference for autonomy, further research is needed
into the relationship between user autonomy and visual indications for replacements, while at the same time making these technologies blend in with other existing tools. Our results also shed light on other important considerations for the design of these technologies. For instance, participants’ suggestions for how to name ATS tools and the way they envisioned them as not being something that should be targeted as an assistance technology for DHH readers, but potentially more as productivity tools for the general population, suggests that perhaps researchers in this area should focus on how to make universally designed tools that may benefit all. However, this does not mean that all research in the area should be done universally, as prior work has found that the linguistic sources of complexity may vary for different user groups [92]. Instead, this suggests that research focusing on the individuality of particular groups should be done within the context of striving towards universal design, identifying the shared experiences the different user groups may have as well. Finally, it is important to note participants’ emphasis on how ATS may be beneficial to other user groups beyond DHH users, as highlighted in Chapter 4 and section 5.3.3. What is notable about this result is the specific group most participants in our second study emphasized, namely, non-native speakers or second-language learners; this sheds light on how DHH users view the reason why these technologies may be beneficial to them as relating to a diversity in English literacy skill as opposed to someone who is DHH, which may also shed light on the proper ways to frame these technologies to support social accessibility.

5.5 Limitations and Future Work

As an interview study, this study shared a lot of the limitations with the study presented in Chapter 4, including those regarding the sample size and the participants’ demographics. However, a specific avenue for future work based on the results from this study include employing larger sample sizes that enable the investigation of a potential relationship between age and concerns about social accessibility issues, considering participants’ comments regarding age and personal acceptance of assistive technology use.
Furthermore, there are critical design considerations that emerge from our results regarding issues of social accessibility, as well as potential directions for mitigating issues that may emerge in those respects. Thus, subsequent chapters in Part III of this dissertation explore how our findings from this study may translate to the actual design of technologies, including how to manage the relationship between user autonomy and privacy in using these tools when needed. Future work should also explore how to balance moving towards universal design while also attending to potentially different user needs.

5.6 Conclusion

This chapter presented an interview study with seven DHH adults with experience in computing and IT fields. This study focused on their perspectives on what they value at work, how those values relate to personal adoption or issues of social accessibility of ATS-based reading assistance tools, and what design aspects may help mitigate potential social accessibility issues in a workplace context. Our findings suggest that participants want to be seen as competent and reliable, but public use of ATS-based reading assistance tools may come into conflict with that desired image because of a perceived relationship between reading skill and intelligence. Our findings also include insights into design solutions to help mitigate the potential social accessibility issues identified.
Epilogue to Part I

Part I of this dissertation focused on identifying the needs, interests and requirements of DHH adults with experience in computing fields for the use of ATS-based reading assistance tools. To do so, Part I presented two studies: an initial study which included an online survey followed by a set of interviews with a subset of the survey respondents to get a deeper understanding of their responses. That study focused on learning about the reading experiences of DHH adults with experience in computing fields, as well as their interests in ATS-based reading assistance tools. Then, a second interview study focused on the perspectives of DHH adults with experience in computing fields on the social accessibility of ATS-based reading assistance tools. More specifically, Part I of this dissertation explored the following research questions:

RQ1.1. What are the current reading experiences of DHH individuals with work experience in the computing field, and are they interested in ATS-based tools to assist those reading experiences? This research question was explored in Chapter 4, and was broken down into the following sub-questions:

RQ1.1.1. What are the reading practices of DHH individuals with work experience in the computing industry? More specifically, how much reading are they doing, how are they reading (i.e. on a screen or paper-based), and what are they reading for? We found that these users read relatively often, mostly on a screen, and especially for work and academic
purposes, which motivates further work into reading assistance tools to support these reading tasks.

**RQ1.1.2. How much do DHH individuals in the computing industry engage in reading for learning about new topics at work?** We found that participants read significantly more often than other activities that support learning, such as watching videos.

**RQ1.1.3. What are the views of DHH individuals in the computing industry about their experience with complicated text, and how do they overcome it?** We identified the most frequently used workarounds participants use for overcoming complicated text, which included the use of dictionaries or finding alternative versions of the text. These workarounds parallel the lexical and syntactic approaches to ATS, which suggests that these approaches to ATS may potentially be beneficial to these readers.

**RQ1.1.4. Are DHH individuals in the computing industry interested in ATS-based reading assistance tools? And if so, for which reading activities would they be interested in it for?** We found that participants expressed strong interest in ATS-based reading assistance tools, and work and academic purposes where among the top purposes of interest, along with medical and legal texts, which motivates future work into domain-specific ATS-based reading assistance for this group.

**RQ1.2. What are the perspectives of DHH adults with work experience in the computing field on the social accessibility of ATS-based reading assistance tools?** This research question was explored in Chapter 5, where we found that while participants want to be seen as competent and knowledgeable in a work environment, the use
of ATS-based reading assistance tools may come into conflict with that because of a perceived relationship between reading skill and intelligence. We also identified user-interface design aspects that may help mitigate these social accessibility issues, which highlights the importance of the design of the user interface of these tools further explored in Part III.
Part II:

Methods for Evaluating ATS Tools
Among DHH Adults
Prologue to Part II

Evaluation is an important aspect of the design of assistive technologies, especially when those technologies are supported by automatic tools that may produce imperfect output. Thus, the focus of Part II of this dissertation consists of methodological research to identify the best ways to measure two key characteristics of the output of ATS systems: the complexity and fluency of the texts. For these methodological studies, participants do not engage with ATS tools directly; instead, participants read texts that have been carefully engineered to meet certain criteria using simplified versions obtained through both manual or automatic simplifications. The results of these methodological studies, inform NLP researchers about the possibility of evaluating the complexity and fluency of automatically simplified texts among DHH adults, and which methods to trust for doing so.

Part II of this dissertation thus begins with a summary of general background on the evaluation of ATS, followed by methodological research on the evaluation of linguistic technologies with DHH adults in Chapter 6. Then, Chapter 7 summarizes a first methodological study exploring different methods for evaluating the complexity of simplified texts. Finally, Chapter 8 summarizes a second methodological study comparing methods for evaluating the fluency of automatically simplified texts among DHH adults at different literacy levels.

More specifically, in Part II of this dissertation, the following research questions are explored:

**RQ2.1:** When English texts of different complexity levels are evaluated by DHH individuals with (a) lower English literacy skill and (b) higher English literacy
skill, which metrics reveal statistically significant differences among texts that vary in a specific characteristic of the output of ATS? In other words, when evaluating texts known to be at different complexity level with these participants, which metrics reveal statistically significant differences among the complexity levels? If there is a significant difference, then the metric is effective for use in evaluating text complexity. We refer to this characteristic as the discriminative ability of the metric, and this is a desirable characteristic. This question is explored in Chapter 7. We investigated this research question for two characteristics of the output of ATS:

RQ2.1.1. Which metrics exhibit a discriminative ability in evaluating the complexity of simplified texts among DHH adults at the different literacy levels? This question is explored in Chapter 7.

RQ2.1.2. Which metrics exhibit a discriminative ability in evaluating the fluency of simplified texts among DHH adults at different literacy levels? This question is explored in Chapter 8.

RQ2.2. In an overall analysis of response scores for all texts varying in a particular characteristic of the output of ATS, when we compare response scores for each metric between DHH individuals in the higher-literacy and lower-literacy groups, do we observe a significant difference? We refer to this as the literacy bias of the metric. While this is not necessarily a problem with the metric that would prevent its use in evaluating text complexity, researchers who use this metric must ensure that they report the literacy skill level of their participants in a study, in order for results across studies to be comparable. Similar to RQ2.1, we explore this question for the same two characteristics of the output of ATS:

RQ2.2.1. Which metrics exhibit a literacy bias when evaluating the complexity of simplified texts among DHH adults at the different literacy levels? This question is explored in Chapter 7.
RQ2.2.2. Which metrics exhibit a literacy bias when evaluating the fluency of simplified texts among DHH adults at different literacy levels? This question is explored in Chapter 8.
Chapter 6

Background and Related Work

As a field that is rapidly progressing within the natural language processing (NLP) research community, much ATS research has focused on its evaluation, e.g. [73, 99, 105]. Evaluations of ATS typically focus on evaluating new machine learning techniques or its benefits as a reading assistance tool for a particular user group. When evaluating the former, ATS systems are typically evaluated in relation to the complexity (or simplicity) of the resulting text, its faithfulness in preserving the meaning of the original text, and the fluency (grammatical correctness) of the resulting text [83, 117].

In this chapter, we discuss prior work on the evaluation of ATS, focusing on why each evaluation was conducted, whom it was conducted with, and which methods were used. This analysis of prior work reveals the lack of methodological guidance and lack of consensus as to how text complexity and fluency should be evaluated with target users. Finally, we consider prior work among DHH adults focusing on the evaluation of other linguistic technologies among these users. While some of this work was not specifically related to ATS, our analysis of this prior research provides guidance for how to conduct methodological research on linguistic technology evaluation with DHH users.
6.1 Evaluating Complexity of Simplified Texts

As mentioned above, one of the key aspects of ATS that is typically evaluated is the complexity of the resulting text. There has been wide diversity in how these evaluations have been conducted, including both manual, and automatic evaluation methods. Most research that relies on manual evaluation typically provides participants with texts at different levels of complexity; these stimuli texts often consist of an original version and simplifications obtained as the output of ATS, e.g. [83, 99, 105].

6.1.1 Metrics Employed in Evaluations of Complexity

There is a lot of diversity in how complexity is measured, including the use of various objective or subjective metrics. Objective metrics rely on behavioral metrics or scores from a task (e.g., [73, 99]). Some studies measure reading speed, under the premise that better comprehension would lead to higher reading speed, e.g. [77, 99, 103]. Other studies include comprehension questions, where the expectation is that better comprehension will lead to better scores, e.g. [73, 99, 105]. Other behavioral metrics have included eye-tracking, based on the premise that text complexity affects fixations duration [60, 99, 117].

In the case of subjective metrics, which rely on direct judgments from participants, researchers often distinguish between two dimensions of text complexity: understandability and readability. Understandability, which focuses more on the user, refers to how easy a text is to understand for a particular reader. Readability, in turn focuses more on the text, referring to how easy the text is to read intrinsically. Thus, subjective metrics often consist of judgements of understandability such as “This text was easy to understand” or judgements of readability such as “This text was easy to read” (e.g., [101, 102, 105, 139]). While the two can seem similar, researchers have argued that they can be measured independently by using examples such as a well-written scientific paper which may be highly readable, but difficult to understand without the proper training [111]. Finally, other work has focused on asking participants for an estimate of one of the objective metrics, under
the assumption that better estimates (e.g. a participant thinking they did better on a quiz or read faster) may indicate better comprehension [77].

Automatic evaluation of ATS, in turn, uses software to analyze texts based on automatic metrics. These metrics can be based on machine translation metrics (e.g. BLEU) or on comparisons against human references of complexity (e.g. SARI) [135]. While the latter directly relies on human judgements, all automatic metrics ultimately rely on human judgements for validation. Thus, manual evaluation of ATS is still crucial even for the development of automatic metrics for automatic evaluation.

6.1.2 User Groups Involved in Manual Evaluations of Complexity

Manual evaluations of the complexity of ATS output are typically done with either high literacy readers, often referred to in the literature as “expert” or “native” readers, or with target users of the technologies. However, the focus of the evaluations is typically different when employed with those different groups, and prior work has observed inconsistencies between the results obtained from evaluations with expert or native readers and evaluations with target users, as well as specific challenges when conducting evaluation with target users.

Evaluations with expert or native readers typically focus on evaluating ATS in terms of whether the resulting simplified text is actually simpler itself. Thus, they typically only employ subjective judgements of complexity or readability e.g., [105, 141], and do not include any of the performance-based objective metrics outlined above. Prior work has also explored the recruitment of these expert or native readers through crowdsourcing, e.g. [75, 83].

In contrast with evaluations with expert or native readers, evaluations with target users typically focus on whether those users obtain benefits from those technologies and thus employ both objective and subjective metrics, to determine whether there are any measurable differences in their performance as measured by the objective metrics, or on users’ perceived levels of understandability or readability, e.g., [99, 105].
However, while some studies have found significant differences from objective metrics such as comprehension questions, e.g. [73], and reading speed, e.g. [139], there have also been challenges: For example, in one study, expert readers judged automatically simplified texts as simpler, but evaluations with target readers with down syndrome showed no significant differences in their scores on comprehension questions when comparing between the original and simplified texts [105]. Other studies revealed significant differences in subjective judgements from target readers, but not on comprehension questions [99]. Inconsistencies with objective metrics have also been observed in the context of evaluating text complexity specifically among DHH readers. For instance, in prior work on measuring text complexity in the health domain with both DHH and hearing readers, significant differences in comprehension questions scores were only revealed for the DHH group, who had significantly lower functional health literacy than the hearing group [73]. Finally, prior work has also suggested that comprehension questions can be confounded by a number of different factors, such as the participants’ understanding of the questions themselves [56].

Studies that employ subjective responses with target readers tend to reveal significant differences between text complexity levels more readily, e.g. [77, 99] However, when using subjective metrics in accessibility, there are concerns of positive bias as well as accessibility barriers preventing users from appropriately judging their own performance [121].

6.2 Evaluating Fluency of Automatically Simplified Texts

As with any automatic system, errors may be introduced in the process of automatically simplifying texts. Errors from ATS are often grammatical or semantic [110], and thus output texts must be evaluated in terms of fluency (grammatical correctness) and faithfulness (preservation of meaning of the original text). Thus, as mentioned above, researchers often wish to measure these other key aspects of ATS output in addition to the complexity of texts, e.g., [105, 141].

Contrary to evaluations of complexity, fluency and faithfulness are not typically eval-
uated by the target reader population, under the assumption that asking a lower-literacy reader to evaluate these aspects of the text would be more difficult. Instead, researchers traditionally conduct evaluations of these two aspects exclusively with expert or native readers. To conduct these evaluations, researchers generally collect subjective judgements of fluency and faithfulness from expert or native readers, e.g. using scalar instruments, often with readers comparing simplified texts to the original, e.g., [105].

Expert readers may indeed be well-suited for evaluating how faithfully the ATS output has preserved the information conveyed by the original text, e.g., through side-by-side comparison of original and simplified texts. However, while there are benefits from asking an expert reader to closely examine an original and simplified text to determine whether the meaning has been faithfully preserved, the exclusion of lower-literacy readers from evaluations of text fluency may be more difficult to justify. In fact, prior work has revealed how specific target groups differ in their linguistic needs and preferences when reading texts, e.g., [92], and other experimental studies have revealed that a reader’s level of literacy influences their subjective judgements about the fluency of a text, e.g. [40].

While there is undoubtedly benefit from gathering the opinion of expert readers about how grammatically correct the output of an ATS system is, that prior research suggests that it would be more valid if the fluency of ATS output could be evaluated with the target reader group. Identifying how target readers can also contribute to evaluations of text fluency may yield new insights, since their perception as to which linguistic aspects of ATS text output may contribute to its overall fluency may differ from those of expert readers.

6.3 Evaluating Other Linguistic Technologies Among DHH Adults

There has been prior methodological research on evaluating other automatic linguistic technologies that may produce imperfect outputs among DHH adults at different literacy levels. Berke et. al investigated how to evaluate imperfect automatic video-captioning technolo-
gies supported by ASR with DHH adults at various literacy levels [12]. In that study, they found subjective judgements to be more effective than comprehension questions overall, but also that the literacy levels of participants affect the effectiveness of the metrics [12]. More specifically, that study revealed it was easier to measure differences in caption quality with higher-literacy participants using subjective metrics, while for lower-literacy participants, subjective metrics requiring meta-cognitive insight into text quality were not effective. Only some objective metrics were effective at evaluating the accuracy of automatic captioning technologies among lower-literacy participants [12].

Prior methodological research has also been conducted on the evaluation of errors in American Sign Language (ASL) animations, e.g. [55, 57]. Studies with DHH participants revealed benefits from collecting responses to both subjective judgements and objective comprehension questions about the content of the message [57]. In contrast with [12], researchers suggested that subjective questions should not be a replacement for comprehension questions in the context of ASL animations with DHH participants [55].

6.4 Limitations and Future Work Needed

All of the prior work on evaluating complexity with target readers relies on being able to effectively measure differences in complexity using the metrics described above (i.e. that if two texts are indeed at different complexity levels, the metrics would reveal significant differences). Despite researchers making use of various metrics for measuring text complexity in studies with target users of ATS technologies, and in spite of calls from prior work for identifying the best manual methods [111, 117] for ATS evaluation, there has been a relative lack of prior methodological work establishing the validity of these instruments for measuring complexity with such users.

Furthermore, the context provided above highlights how it is important to identify whether it is possible to measure the fluency of automatically simplified texts among target users, and if so, what are the best ways to do so. To the best of our knowledge, however,
no prior methodological work has identified guidance on whether fluency can be evaluated among a target population such as DHH adults and if so, which metrics would be more effective, and whether those metrics would be affected by the literacy levels of participants.

As illustrated by the contrasting results from prior methodological research on the effectiveness of different types of metrics for evaluating different linguistic technologies, outlined above, the results for evaluating specific aspects of particular linguistic technologies may not generalize to other aspects of other linguistic technologies. Thus, additional methodological work is needed to understand the effectiveness of different metrics for evaluating different aspects of the output of ATS systems, which is the focus of Chapter 7 and Chapter 8. Considering influences of participant’s literacy levels on the effectiveness of particular metrics identified in prior work [12], the methodological studies presented in Part II closely follow the methodology of [12], by using literacy level as a factor and, in our case, text complexity (Chapter 7) and fluency (Chapter 8) levels as another factor. While there is overlap between this work and that of [12], the studies presented in this dissertation also differ in terms of the specific metrics used since we focus on those identified above, from prior work on evaluating ATS.
Chapter 7

Identifying Appropriate Metrics for Evaluating Complexity of Simplified Texts Among DHH Adults

As a first step to reliably evaluate ATS systems with DHH adults, methodological research is necessary to determine what types of metrics are effective in measuring the complexity level of texts when evaluating them among DHH readers. Thus, this chapter presents an investigation of which metrics are effective for evaluating the complexity of the output from ATS with DHH readers, and how respondents’ literacy levels may influence that effectiveness. Thus, as a methodological study, this study relies on carefully selected texts with human-made simplifications that are known to be at different complexity levels, as judged by a DHH literacy expert.
7.1 Hypotheses Evaluated in This Study

The following hypotheses are evaluated for each metric included in the study:

1. H1: When English texts of different complexity levels are evaluated by DHH individuals with (a) lower English literacy skill and (b) higher English literacy skill, the response scores for this metric will reveal statistically significant differences among the complexity levels. We refer to this characteristic as the discriminative ability of the metric, and this is a desirable characteristic. If there is a significant difference, then the metric is effective for use in evaluating text complexity.

2. H2: In an overall analysis of response scores for all texts, when we compare response scores between DHH individuals in the higher-literacy and lower-literacy groups, we will observe a significant difference. We refer to this as the literacy bias of the metric. While this is not necessarily a problem with the metric that would prevent its use in evaluating text complexity, researchers who use this metric must ensure that they report the literacy skill level of their participants in a study, in order for results across studies to be comparable.

7.2 Method

7.2.1 Reading Stimuli

This study required texts at different levels of complexity as a source for ground truth. One approach to obtain these could have been using an ATS system for acquiring the different levels of complexity (by simplifying an already-complex text). However, because ATS systems are still rather experimental, that approach would likely introduce errors [110], which could confound our results. This study focuses on evaluating one aspect of the quality of ATS output text: complexity, not on its level of error. Thus, to ensure having texts at different levels of complexity – and with good quality – a Wizard-of-Oz approach, using
human-made simplifications, was employed. The texts for this study were selected from Newsela\(^1\), a website that provides news articles at different levels of linguistic complexity as a resource for school teachers, which had been used by prior work in NLP as a source of simplifications for training ATS systems [134].

**Text selection**

We first selected 10 articles from the Science section of Newsela that were relatively similar in word-length (between 550 and 750), had Flesch-Kincaid grade levels close to 12th grade and also had the same number of simplified versions available of similar complexity\(^2\). For each article, we selected three versions: the original version, the one with medium-level complexity as well as the simplest version available. An expert in Deaf literacy then selected a subset of 6 articles to use in this study based on 1) whether the simplifications provided by Newsela would actually be simpler for DHH readers; 2) the background knowledge one could expect an average Deaf reader to have about the topic (e.g. one of the articles that were part of the initial set of 10 related to the use of sounds); and 3) prioritizing the ones who had the least number of grammatically complex comprehension questions (described in next subsection). Specifically, articles were omitted if the accompanying questions had more than one dependent clause or phrase because that makes them more difficult for DHH readers to understand [36].

The final set of 6 articles used in this study had an average length of 669 words (SD = 85.15) and Flesch-Kincaid grade level of 12.41 (SD = 0.86) in their original versions (henceforth referred to as our high-complexity condition), which means a high-school graduate in the U.S. should be able to understand them. In their medium-complexity version, their

\(^1\)https://newsela.com

\(^2\)We rely on and report Flesch-Kincaid grade levels [67] as an initial measurement of readability, as this readability formula has been widely used for measure text simplification tasks [111]. However, considering that the formula only looks at surface-level characteristics such as word- and sentence-length, additional human validation of various characteristics of the texts is needed for the work presented in Part II to ensure the reliability of our ground truth
average length was of 714 words (SD = 76.78), with an average Flesch-Kincaid grade level of 8.9 (SD = 0.76), meaning a student in their first year of high-school in the U.S. should be able to understand them. Finally, in their low-complexity version, their average length was 412 words (SD = 128.51), and the average Flesch-Kincaid grade level was 4.3 (SD = 0.31), which should be understandable for a 4th grade student in the U.S. Links to all of these articles are included in Appendix E.

7.2.2 Metrics

We selected a set of metrics identified from prior work, including both objective and subjective metrics to evaluate in this study. These metrics included reading speed, comprehension questions (at different levels of linguistic complexity), participants’ predictions of how well they did on the comprehension questions, as well as subjective judgements of the understandability and readability of the texts.

Reading Speed

We included reading speed as a potential objective metric of readability, since it had been used in prior work, as discussed in the Background and Related Work section, under the assumption that more readable text leads to higher reading speed. The reading speed was measured in words per minute (wpm).

Comprehension Questions

Our comprehension questions consisted of main-factual questions written as multiple-choice, which varied in their linguistic complexity. This is because, first, when using comprehension questions, most prior work has written them as multiple-choice e.g. [35, 56, 73, 99]. Second, considering that there are also different types of questions in terms of what they ask of the users [35], we decided to use main-factual questions (i.e. questions that ask about relatively important aspects of a text) because prior work has identified them as easier than other types
CHAPTER 7. METRICS FOR EVALUATING COMPLEXITY

<table>
<thead>
<tr>
<th>High Linguistic Complexity</th>
<th>Which of the following statements was true about the animal’s physical composition?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a) It had no mechanism for chewing</td>
</tr>
<tr>
<td></td>
<td>b) Its head was relatively small in relationship to its body</td>
</tr>
<tr>
<td></td>
<td>c) It was able to blow fire out of its mouth</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Low Linguistic Complexity</th>
<th>What is true about the animal?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a) It had no way to chew.</td>
</tr>
<tr>
<td></td>
<td>b) Its head was small.</td>
</tr>
<tr>
<td></td>
<td>c) It could blow fire.</td>
</tr>
</tbody>
</table>

Table 7.1: One of the comprehension questions for an article, in its two versions.

of questions [35], which allowed us to control for question-type difficulty. Finally, prior work has suggested that the varying linguistic complexity of questions (i.e. some questions can be harder to read than others) can affect whether a user actually understands the questions themselves [56]. So, we decided to test whether varying the degree of linguistic complexity of the comprehension questions themselves would have an effect on their effectiveness, so we produced low and high-linguistic complexity comprehension questions using the procedure outlined below.

Now, we could have authored different comprehension questions for each article version (i.e. their high, medium and low complexity conditions) each article. However, we wanted to determine whether asking the same questions would effectively distinguish the varying complexity levels of the different article versions. This approach thus required that we authored the questions in a way that would not favor a particular article version by asking about a fact that could have been trimmed during the simplification process and thus not be present in one of the simplified versions. To address this, we identified facts that would be present in all three versions of each article. Then, we created 10 pairs of multiple-choice main-factual comprehension questions for each article. As illustrated in Table 7.2.2, each pair contained the same question (i.e. they were asking about the same fact and the multiple-choice options were the same fact), but both the question items and the options
were written at a high and a low linguistic complexity level. When looking at the low-
linguistic-complexity question in each pair across all of the texts, their average Flesch-
Kincaid grade level was 3.75 (SD = 0.67) which means we could expect a U.S. 3rd- to
4th-grader to understand them. In turn, the average Flesch-Kincaid grade level for the
high-linguistic-complexity questions was 9 (SD = 0.42), which should be understandable
for a student in their first year of high school in the U.S. Then, we selected a subset of 6
pairs of questions by discarding those with higher grammatical complexity in their high-
linguistic-complexity versions, or questions that asked about facts that were not directly
tied to the main idea of the article.

The final set of questions consisted of 6 pairs of questions for each article. Now, while
we could have shown both versions in each pair to participants, the version that would
come second would have been likely to seem repetitive to participants and to elicit the same
answer as the version of the same question that appeared first. So, instead we chose to
show each participant 3 questions in their low-linguistic complexity version and 3 in their
high-linguistic complexity version for each text, and rotated their selection for each article
across participants. The 6 questions were then arranged randomly as a single quiz. Then,
we scored participants’ scores on the low-linguistic-complexity questions separately from
the high-linguistic complexity questions to determine if either set was more effective than
the other at identifying the complexity of the articles.

Score Prediction

We also included a question that asked participants to give a subjective estimate of how
well they would do on a task. Specifically, we asked participants to estimate on a 0-100
scale the grade they expected to get on the quiz they had just completed. This question is
similar to one used in prior research among DHH individuals, in which they were asked to
predict their success at other academic tasks [125].
Understandability and Readability

We also included two subjective questions that have been widely used in the literature on the evaluation of text simplification. The first question, which focuses on a text’s understandability and is thus more focused on the reader, reads as “I was able to understand this text well” and uses a 5-point Likert-type scale of agreement going from “Strongly disagree” to “Strongly agree.” The second question, which focuses on the text’s readability and is thus more focused on the text, reads as “This text was easy to read,” and also uses the same 5-point Likert-type scale.

7.2.3 English Reading Literacy

In order to group participants by their literacy level, we needed a reliable metric of their English reading literacy skill. Thus, we administered the sentence comprehension sub-test of the Wide Range Achievement Test in its 4th edition (WRAT4) [132], which has been used in prior work as a measurement of literacy level with DHH readers because it is brief, can be administered without audio stimuli and has been previously validated with DHH people [64, 95].

7.2.4 Data Collection

Procedure

This IRB-approved study was conducted remotely due to social-distancing restrictions during the COVID-19 pandemic. Thus, participants were provided with a consent form via e-mail ahead of the study. Then, participants met via Zoom with a researcher on the team who is hard-of-hearing and fluent in ASL for a 70-minute appointment. Participants were directed to a website created using jsPsych [31], which contained all of the stimuli for the study. The first screen on the website was an introduction, which contained detailed instructions of the study and indicated to participants that they would be reading texts that had all been simplified using different simplification tools. We told participants that all
texts had been simplified to avoid a placebo effect. No other indications of whether the
texts had any transformations were provided to participants.

Each participant proceeded through all 6 articles, reading 2 at each complexity level
(high, medium and low). The order of articles and text complexity conditions was rotated
using a Graeco-Latin square design. After reading each article, participants answered the
quiz described above, which contained 6 comprehension questions in total, 3 with low lin-
guistic complexity and 3 with high linguistic complexity. Which 3 questions were selected
at each linguistic complexity level was rotated across participants, and the order in which
they were displayed in the quiz was randomized. After the quiz, participants were asked to
predict their grade, followed by the subjective metrics of understandability and readabil-
ity. After the third article, participants were encouraged to take a quick break to prevent
fatigue.

After reading all 6 articles, participants filled out the sentence comprehension sub-test of
the WRAT, followed by a demographics questionnaire. Participants were then compensated
with $40 for their participation.

Participants

Participants were recruited through social-media and email advertising based on the criteria
of identifying as Deaf or Hard-of-Hearing (DHH) and being over 18 years old. We recruited
a total of 59 participants for our study. Participants’ self-identified genders included female
(N = 31), male (N = 25) and non-binary (N = 1). Participants’ average age was 27.33
(SD = 10.17, range = 18 - 63). A total of 19 participants identified as hard-of-hearing,
with 28 identifying as culturally Deaf [91], 9 as deaf and one as Deaf/Blind. Participants’
average WRAT score was 87.82 (SD = 15.58, range = 63 - 126), which is lower than the
national average in the U.S. of 100 [132]. To investigate hypothesis H1, which focused on
participants with lower (H1a) and higher literacy skills (H1b), we split participants into
two groups based on their median WRAT score, which was 86. Our two groups, labeled as
WRAT-L and WRAT-H, respectively, were as follows:

- **WRAT-L**: participants with WRAT scores of 86 or lower (mean = 75, SD = 6.95, range = 63 - 85).
- **WRAT-H**: participants with WRAT scores higher than 86 (mean = 100.6, SD = 10.34, range = 87 - 126).

Three participants did not have time to complete the WRAT sentence comprehension form, and thus their responses were not included in the analysis. Upon careful analysis of the reading speed of participants, following the methodology of [77], we excluded the data from two participants whose reading speed was higher than the median reading speed plus 3 times the Interquartile Range (IQR), which means they may have just been skimming through the text \(^3\). This left us with 54 participants and, after splitting them into two groups based on their median WRAT score as detailed above, each group consisted of 27 participants.

### 7.2.5 Data Analysis

After conducting Shapiro-Wilk tests of normality for all of the results, none of them followed a normal distribution, and thus we conducted non-parametric tests for our difference testing. For each of the literacy groups, we conducted Kruskal-Wallis tests to determine if there were statistical differences between the text complexity conditions (H1).

If there were significant differences, then we conducted post-hoc pairwise comparisons using Mann-Whitney U-tests with Bonferroni corrections. To compare between the two literacy groups (H2), given that this was a non-parametric between-groups comparison we conducted Mann-Whitney U-tests as well.

\(^3\)One participant only met this criteria for one of the articles, so we only excluded the data for that specific article for that participant.
### Table 7.2: A summary of the results for each metric across each hypotheses. Hypotheses were not supported if the metric did not discriminate between any complexity levels, and partially supported if it discriminated between some but not all pairs of complexity levels.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Discriminative Ability among Lower Literacy DHH Respondents (H1a)</th>
<th>Discriminative Ability among Higher Literacy DHH Respondents (H1b)</th>
<th>Literacy Bias (H2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading speed</td>
<td><strong>H1a was not supported.</strong> This metric was not discriminative between any text complexity levels.</td>
<td><strong>H1b was not supported.</strong> This metric was not discriminative between any text complexity levels.</td>
<td><strong>H2 was supported.</strong> Higher literacy readers had significantly higher reading speed than lower literacy readers.</td>
</tr>
<tr>
<td>High-linguistic-complexity comprehension questions</td>
<td><strong>H1a was not supported.</strong> This metric was not discriminative between any text complexity levels.</td>
<td><strong>H1b was not supported.</strong> This metric was not discriminative between any text complexity levels.</td>
<td><strong>H2 was supported.</strong> Higher literacy readers had significantly higher scores than lower literacy readers.</td>
</tr>
<tr>
<td>Low-linguistic-complexity comprehension questions</td>
<td><strong>H1a was partially supported.</strong> Worked well to distinguish between lowest and highest text complexity only.</td>
<td><strong>H1b was not supported.</strong> This metric was not discriminative between any text complexity levels.</td>
<td><strong>H2 was supported.</strong> Higher literacy readers had significantly higher scores than lower literacy readers.</td>
</tr>
<tr>
<td>Score prediction</td>
<td><strong>H1a was partially supported.</strong> Worked well to distinguish between lowest and highest text complexity only.</td>
<td><strong>H1b was not supported.</strong> This metric was not discriminative between any text complexity levels.</td>
<td><strong>H2 was supported.</strong> Higher literacy readers predicted significantly higher scores than lower literacy readers.</td>
</tr>
<tr>
<td>Understandability</td>
<td><strong>H1a was partially supported.</strong> Worked well to distinguish low-complexity texts from both the medium and high-complexity texts.</td>
<td><strong>H1b was not supported.</strong> This metric was not discriminative between any text complexity levels.</td>
<td><strong>H2 was supported.</strong> Higher literacy readers had significantly higher judgements than lower literacy readers.</td>
</tr>
<tr>
<td>Readability (Best Metric)</td>
<td><strong>H1a was partially supported.</strong> Worked well to distinguish low-complexity texts from both the medium and high-complexity texts.</td>
<td><strong>H1b was partially supported.</strong> Worked well to distinguish between lowest and highest text complexity only.</td>
<td><strong>H2 was supported.</strong> Higher literacy readers had significantly higher judgements than lower literacy readers.</td>
</tr>
</tbody>
</table>
7.3 Results

As summarized in Table 7.2, we observed significant differences between text complexity levels when using low-linguistic-complexity comprehension questions as well as all of the subjective metrics (i.e. the score prediction, and the understandability and readability judgements) with the WRAT-L group. With the WRAT-H group, in turn, we only observed significant differences between text complexity levels when using readability judgements. There were significant differences between the two groups (WRAT-L and WRAT-H) with all of the metrics.

In the rest of this section, we present the detailed results of the statistical analysis first for H1, and its sub-hypotheses, followed by the results for H2 and all of its sub-hypotheses. Complementary to these detailed results, figures 7.1 through 7.10 illustrate significant results using whisker plots for continuous data (i.e. the reading speed, comprehension question scores for both the high-linguistic and low-linguistic complexity questions, and their score predictions), and stacked bar charts for the Likert-type data (i.e. understandability and readability judgements) which are the recommended way of plotting Likert-type scales [104]. Figures 7.1 through 8.3 show the results for each text complexity, grouped by literacy group (for hypotheses H1a and H1b), while figures 7.5 through 7.10 show the results for hypotheses H2, which looked at the results for each group as a whole.

7.3.1 Discriminative Ability (H1)

Reading Speed

For each group, the statistical analysis revealed no statistically significant differences for either group (p-value of 0.37 for WRAT-L and 0.12 for WRAT-H).
Figure 7.1: Low-linguistic complexity comprehension questions success for H1, with a max. value of 100% (* = p < 0.05).

High-linguistic-complexity Comprehension Questions

The results from the statistical tests were not significant for either group (p-value of 0.78 for WRAT-L and 0.76 for WRAT-H).

Low-linguistic-complexity Comprehension Questions

The difference between the different text complexity levels was statistically significant for the WRAT-L group ($\chi^2 = 8.1383, p = .017$). Pairwise comparisons revealed statistical differences between the low and high complexity conditions ($Z = -2.55, r = 0.34, p = 0.011$). For the WRAT-H group, however, the results were not statistically significant ($p = 0.75$). These results are illustrated in Figure 7.1.

Score prediction

As illustrated in Figure 7.2, there were significant differences between the text complexity conditions for the WRAT-L group ($\chi^2 = 8.1135, p = .017$), with pairwise tests revealing differences between the low and high text complexity conditions ($Z = -2.315, r = 0.315, p$
Figure 7.2: Score predictions for the comprehension questions for H1, with a max. value of 100% ( * = p < 0.05).

= 0.021). On the other hand, the results for the WRAT-H were not significant (p = 0.08)

Understandability

The results revealed significant differences between the text complexity conditions for the WRAT-L group (χ² = 13.0794, p = 0.001), with pairwise tests showing significant differences between the low and the medium complexity conditions (Z = -2.15, r = 0.293, p = 0.031), as well as between the low and the high complexity conditions (Z = 0.293, r = 0.482, p < 0.001). However, the results for the WRAT-H group revealed no significant differences (p = .35). Figure 8.2 illustrates these results.

Readability

Finally, as illustrated in Figure 8.3, the tests revealed significant differences between the text complexity levels for both literacy groups. For the WRAT-L group (χ² = 24.2346, p < 0.0001), pairwise comparisons revealed significant differences between the low and the medium complexity conditions (Z = -3.161, r = 0.43, p = 0.0016), as well as between the low and the high complexity conditions (Z = -4.81, r = 0.655, p < 0.0001).
Figure 7.3: Understandability judgements for H1 using a Likert-type agreement scale (\( * = p < 0.05 \), \( *** = p < 0.001 \)).
Figure 7.4: Readability judgements for H1 using a Likert-type agreement scale (\( ** = p < 0.01 \), \( *** = p < 0.001 \)).
CHAPTER 7. METRICS FOR EVALUATING COMPLEXITY

For the WRAT-H group ($\chi^2 = 7.0867, p = .028$), pairwise comparisons only revealed significant differences between the low and high text complexity conditions ($Z = -2.29, r = 0.312, p = 0.022$).

### 7.3.2 Literacy Bias (H2)

H2 was supported for all of the metrics, with the WRAT-H group obtaining higher scores or providing higher predictions of their scores, as well as providing higher ratings for understandability and readability. For each metric, the results of the Mann-Whitney U-tests were:

- **Reading speed**: $Z = -3.7662, r = 0.206, p < .001$ (Figure 7.5).
- **High-linguistic-complexity comprehension questions**: $Z = -7.199, r = 0.566, p < .00001$ (Figure 8.6).
- **Low-linguistic-complexity**: $Z = -5.6, r = 0.44, p < .00001$ (Figure 8.5).
- **Score prediction**: $Z = -4.499, r = 0.353, p < .00001$ (Figure 7.8).
- **Understandability**: $Z = -3.758, r = 0.295, p = .001$ (Figure 8.7).
- **Readability**: $Z = -2.757, r = 0.217, p < .01$ (Figure 7.10).

### 7.4 Discussion

In this section, we discuss the results and their implications first for the hypothesis related to the discriminative ability of the metrics (H1), followed by the literacy bias of the metrics (H2).

#### 7.4.1 Discriminative Ability (H1)

Overall, the metric that was most effective with both groups (WRAT-L and WRAT-H) was the subjective readability question “This text was easy to read.” This was also one of only
Figure 7.5: Reading speed for H2, calculated in words per minute (wpm) (*** = p < 0.001).

Figure 7.6: High-linguistic complexity comprehension questions success for H2, with a max. value of 100% (*** = p < 0.001).
CHAPTER 7. METRICS FOR EVALUATING COMPLEXITY

Figure 7.7: Low-linguistic complexity comprehension questions success for H2, with a max. value of 100% (*** = p < 0.001).

Figure 7.8: Score predictions for the comprehension questions for H2, with a max. value of 100% (*** = p < 0.001).
Figure 7.9: Understandability judgements for H2 using a Likert-type agreement scale (*** = p < 0.001).

Figure 7.10: Readability judgements for H2 using a Likert-type agreement scale (** = p < 0.01).
two metrics that was able to distinguish between the low and medium complexity conditions with WRAT-L, as well as the only one that was effective with the WRAT-H group.

When focusing on the objective metrics included (i.e. reading speed and comprehension questions), only comprehension questions were effective under three different conditions. First, comprehension questions were effective only when focusing on readers with lower literacy levels; we only observed significant differences across the complexity conditions when focusing on the WRAT-L group. The wording of the questions themselves has to be written in a way that is easy enough for those readers to understand them; we only observed significant differences when focusing on the low-linguistic-complexity comprehension questions. Finally, the texts also need to be far apart in complexity; we only observed significant differences between the low and high complexity conditions. Thus, we conclude that while some of these conditions may be of particular interest for researchers who want to employ comprehension questions, the results of comprehension questions as an objective metric of text complexity should be interpreted carefully unless those questions have been either carefully validated or piloted prior to using them as a metric with the particular literacy group of interest. However, as discussed in the next paragraph, subjective questions were highly effective for evaluating text complexity; so, while comprehension questions may have value in ensuring that participants pay attention to readings because of the pressure of being tested afterwards, subjective questions may be sufficient to evaluate the complexity of texts with DHH readers when those texts are known to be of good quality (as they were in our study given that they were human-made simplifications).

The subjective metrics (i.e. score prediction, understandability and readability judgements) were all effective with the WRAT-L group. While prior work has highlighted concerns around the meta-cognitive literacy knowledge skills of DHH readers (i.e. their ability to make judgements about their literacy skills), especially those with lower literacy levels [87, 119, 136], our results suggest that when focusing on the complexity of texts, readers with lower literacy are able to effectively judge this difference in relative complexity. However, only subjective judgements of readability were effective with the WRAT-H group, which
highlights a trend in our results: it is more difficult to effectively differentiate complexity levels with higher literacy readers.

Our results are in line with prior work on the evaluation of the quality of captioning tools with DHH readers in terms of subjective metrics being more effective [12]. This is in contrast with prior work on the evaluation of ASL animations with DHH adults, which suggested subjective metrics should not be a replacement for comprehension questions [55]. However, our results are in stark contrast with the results of the captioning evaluation study [12] in terms of which group the metrics were more effective with; while their study revealed it was harder to evaluate imperfect captioning tools with lower literacy readers, our results suggest that it is harder to evaluate text complexity with higher literacy readers. We speculate that in our case, the articles with high complexity (which were around the 12th grade level) may not have been challenging enough for the higher literacy readers. Two of our subjective metrics focused on the readers: predicting how they had scored on the quiz and judging how easy the text was for them to understand. However, when focusing on the text, namely how easy the text was to read, higher literacy readers were able to judge it properly.

These results also help illuminate inconsistencies described in the background and related work section on the evaluation of text complexity with DHH readers. In a study [73] that had revealed significant differences using comprehension questions among DHH readers, these significant differences were only revealed for the DHH group, who had significantly lower functional health literacy than the hearing control group. This suggests that one of the conditions identified in our results for comprehension questions may have been met: having a group with lower literacy. The text complexity levels in that study were also reportedly around 4 grade-levels apart, which may have met another condition identified in our results: text complexity levels being far apart. However, the researchers in that study did not report the linguistic complexity of their comprehension questions, so it is hard to determine if the condition of having questions that are easy to understand was met.

Our results suggest that subjective metrics may not only be discriminative of text com-
plexity levels when evaluating them with DHH readers, but they are also more effective than our objective metrics. Thus, we recommend always including subjective metrics in text-complexity studies with DHH readers. However, if researchers still want to include comprehension questions in their studies, we recommend being careful about the interpretation of a lack of significant differences in comprehension questions, unless they have been carefully crafted with low linguistic complexity and validated after controlling for the literacy level of participants. It is important to note that these recommendations are in the context of non-erroneous texts. However, given that, as with any automatic system, errors are inevitable in the output of ATS systems [110], further research is necessary to determine how the introduction of both semantic and grammatical errors would impact our results since some metrics may be able to capture the different types of errors better than others. Further, while we acknowledge we cannot guarantee that with more statistical power statistically significant differences would not emerge for some of the objective metrics in our study, we note that most prior work on evaluation of ATS systems with DHH readers has used smaller sample sizes without controlling for literacy levels. Thus, if we did not observe significant differences with our sample size and controlling for literacy levels, it is unlikely they would emerge in a study with fewer participants.

7.4.2 Literacy Bias (H2)

Unlike prior work on the evaluation of captioning tools, in which some metrics did not exhibit a literacy bias, our results suggest that all of the metrics we used had a literacy bias. More specifically, participants in the WRAT-H group scored and predicted higher scores, and also judged texts to be easier than participants in the WRAT-L group. This does not mean that these are bad metrics. Instead, this suggests that researchers should always carefully report and/or control for the literacy levels of the participants in their study when evaluating the complexity of texts with DHH readers. Readers of published research studies should then consider this literacy bias when comparing across studies, since texts of equal complexity may elicit different scores depending on the literacy level of
the participants they are evaluated with.

It is difficult to generalize this result to other user groups (i.e. whether these metrics would have a literacy bias with other user groups) given that prior work has established that the sources of difficulty with texts may be different for different user groups [92]. Nevertheless, these results are still useful to researchers working with other user groups to know that these metrics may be biased for participants’ reading skill. Similar analysis using measurements of the relevant skill a specific user group struggles with could reveal similar trends.

7.5 Limitations and Future Work

There were several limitations to our study. While it was our original intention to include measures obtained through eye-tracking (e.g. fixation duration) in our study, given that they have been used in prior work on evaluating ATS with target readers, we were prevented from using eye-tracking because we had to conduct our study remotely to abide by social-distancing restrictions due to COVID-19. Future work can thus investigate the effectiveness of eye-tracking for distinguishing text complexity levels among DHH readers.

In this study, we decided to focus only on one aspect of the potential output of an ATS system: the level of complexity of the output. But as mentioned in our discussion, the introduction of errors in the output is inevitable [110]. So, future work should focus on how to measure the overall quality of the output of simplification systems with DHH readers, and how the introduction of semantic and grammatical errors could affect the effectiveness of the metrics for evaluating text complexity suggested by our results, which we explore in Chapter 8.

As mentioned in our discussion, the high-complexity conditions for the articles in our study may not have been challenging enough for participants with higher literacy levels. Future work could determine whether some of the metrics analyzed in our study would be effective for evaluating texts with higher complexity among participants with higher
literacy. Further, while we grouped participants into two groups using their WRAT scores, we could have recruited a greater number of participants or used a different literacy test to determine whether further subdivision could reveal other patterns.

Another limitation of our study was that our sample size did not allow us to guarantee that with a larger sample size, and thus more statistical power, significant differences would not emerge for metrics for which we did not observe significant differences in this study.

In this study, our participants identified as DHH, but not necessarily as users of ASL. A future study could recruit a more narrow demographic of participants, with a specific focus on ASL signers, in which ASL-based comprehension questions could also be examined. While we only included texts in the science genre as stimuli, future work could explore whether our findings would still hold with other text genres. Finally, future work could focus on exploring the efficacy of the metrics explored in our study with other user groups and broader literacy levels.
Chapter 8

Identifying Appropriate Metrics for Evaluating Fluency of Simplified Texts Among DHH Adults

This study investigates which metrics (including the ones evaluated in Chapter 7 and others identified from prior work focused on fluency and system performance) are effective among DHH adults for evaluating the fluency of ATS output texts. Given the diversity in literacy among DHH readers, we also investigate whether participants’ responses to metrics varies depending upon their reading-literacy level. As this is a methodological study, we have carefully engineered stimuli texts to be at specific known levels of fluency.

8.1 Hypotheses Evaluated In This Study

For each metric, we evaluate the following hypotheses:

1. **H1**: When evaluating English texts at different fluency levels, due to errors introduced
by ATS, participants’ responses for this metric will reveal statistically significant differences among texts of different fluency levels. This characteristic, which is desirable, has been referred to as the **discriminative ability of the metric** in prior work [12] and Chapter 7. If there is a significant difference, then the metric is effective for use in evaluating the fluency of the texts. For each metric, we investigate this hypothesis among two sub-groups of DHH readers: **(H1a)** those with lower English literacy skill and **(H1b)** those with higher English literacy skill.

2. **H2**: When analyzing response scores for all texts, comparing the response scores of DHH individuals in a higher-literacy and lower-literacy group, a significant difference will be observed. This has been referred to in prior work (and Chapter 7) as the **literacy bias of the metric** [12] and it is not necessarily a problem with the metric nor would it prevent its use for evaluating the fluency of texts. Instead, when using this metric in a study, researchers need to consider and report the literacy skill level of their participants so that results across studies can be comparable.

### 8.2 Method

#### 8.2.1 Reading Stimuli

In a typical study to evaluate the output of an ATS system, participants may be asked to respond to a set of trusted question-instruments to assess a set of texts with unknown levels of fluency. However, in this work, we are conducting a methodological study, and thus the design of the study differs (informally, the study design may feel somewhat “backwards”): In our study, we need a set of texts that are known to be at specific levels of fluency, and we will ask participants to respond to question-instruments in order to determine whether those question-instruments are able to discriminate between texts of different fluency and whether there are any literacy biases in the response to those items. Therefore, it is essential that our text stimuli are carefully selected such that they have specific levels of fluency.
Given that our goal is to inform researchers evaluating ATS systems, we wanted the texts included in our study to exhibit realistic levels of quality from automatic systems. However, it would have been difficult to simply process texts with ATS while carefully controlling for both fluency and complexity levels. Thus, we employed a semi-automated process using the output of two state-of-the-art ATS systems as well as manual simplifications. Then, we mixed sentences from each of these sources, such that the resulting text consisted of an interleaved mixture of sentences from: the original (complex) text, the output of one of the ATS systems, and a simplified text that had been produced by a human author. Details of our procedure for generating text stimuli are described below.

8.2.2 Stimuli Generation Procedure

Articles.

We selected 6 articles from Newsela\(^1\), an educational website that provides human simplified versions of news articles. Our 6 articles were from the Science section of Newsela, which had been identified as appropriate to use with DHH readers, and had been used in our previous methodological study for evaluating the complexity of simplified texts among DHH readers. Before simplification, these articles had an average Flesch-Kincaid grade level of 12.4 (SD = 0.86). We will refer to these original versions of these articles as “Original.” Links to these articles are provided in Appendix E.

Simplifications.

We processed each of the original articles through two state-of-the-art ATS systems.

1. A state-of-the-art hybrid approach to ATS that incorporates both syntactic rules and data-driven neural models, and allows control over parameters, such as: the number of sentence-splits and deletions, the average sentence lengths in the output, and the percentage of words from the original text that still remain in the paraphrased output.

\(^1\)https://newsela.com
For this study, that soft constraint over the number of words copied from the original input when paraphrasing was set at 70%. The output from this system had an average Flesch-Kincaid grade level of 8.8 (SD = 0.77). We will refer to the versions of the articles obtained through this ATS system as “Hybrid.”

2. A Transformer-based seq2seq model trained using the datasets presented in [59], which had established a new state of the art. This recent system benefited from an improved sentence alignment step in order to create new sentence-aligned datasets, and the researchers had compared different simplification models trained using these new datasets [59]. The output from this system had an average Flesch-Kincaid grade level of 6.7 (SD = 0.9). We will refer to the versions of the articles obtained through this ATS system as “Transformer.”

As mentioned above, while it was one of our goals to evaluate texts that exhibited realistic levels of fluency from ATS systems, we also needed to carefully control the fluency and complexity levels of the texts. So, to obtain sentences with high-fluency and low-complexity (for use in creating our text stimuli), we selected one of the human-authored simplifications for each article that had been included in the Newsela dataset. On average, these human-made simplifications were at a Flesch-Kincaid grade level of 8.9 (SD = 0.76), which were the closest to the average literacy grade-levels observed in prior work with DHH adults (e.g. in [12]) as well as the participants in Chapter 7. We will refer to these versions of the articles as “Newsela.” Links to these articles are also provided in Appendix E, where they are labeled as ”medium.”

Thus, at this stage, for each of our six articles, we had four different versions of the text: (1) an Original complex version, (2) output from the Hybrid ATS system, (3) output from the Transformer ATS system, and (4) a human-authored Newsela simplification of the article. To produce text stimuli for our study, our goal was to assemble texts consisting of concatenated subsets of sentences from each of these four sources. In order to guide this assemblage of stimuli texts, we needed to know the fluency and complexity of each sentence
in each source text. With this information, a mixture of sentences could be selected from each, to achieve a final text of a specific fluency and complexity level.

Annotations.

We first segmented the original article at a sentence level. Then, we aligned each sentence from the original article to the sentence(s) produced by the ATS or human simplifications. In many cases, the ATS systems and human authors had split sentences, and thus multiple simplified sentences aligned to a single original sentence in such cases. Two high-literacy native English speakers rated each original sentence and its possible replacements on 5-point Likert scales for grammaticality and complexity. The judgements of grammaticality were collected using an item from prior work [105] that asked the annotators to indicate their agreement with the item: “This sentence is grammatically correct.”

Judgements of complexity were collected on a scale in which disagreement indicated harder to read texts, using the item “This sentence is easy to read,” which had been used in prior work (e.g. [101, 102, 139]) and explored in Chapter 7 for evaluating complexity of texts. For clarity, we reverse-scored annotators’ responses for this particular item, using 5 for Strongly Disagree and 1 for Strongly Agree, such that we may more intuitively refer to the resulting score as complexity.

The average judgements from our annotators for all sentences in the Original versions of texts were 1.6 for complexity and 4.5 for fluency. The average judgements among all sentences in Hybrid texts were 2.8 for complexity and 3 for fluency. For the average judgements among all sentences in Transformer texts were 2.5 for complexity and 3.2 for fluency. Finally, for sentences in Newsela texts, the average judgements were 1.1 for complexity and 4.8 for fluency.

Finally, we asked a DHH literacy expert to judge a subset of 24 sentences selected

---

2This researcher is a university professor of Deaf and Hard-of-Hearing (DHH) literacy and education who has published over 10 journal articles in DHH literacy venues, and worked as a teacher of DHH students for several years before pursuing a PhD.
at random from the six articles so we could compute inter-rater reliability on that subset, to determine the level of agreement in the expert’s ratings with those of our two natively fluent annotators. As these were ordinal data, we computed Krippendorf’s alpha, obtaining a moderate alpha value of 0.554.

**Generation.**

For each sentence, since there were four options of possible sources to use, the set of all possible texts that could have been generated would be combinatorially numerous. For instance, given that for each sentence, we had four options (the original, the two obtained from ATS, and one human simplification), an article with 26 sentences would generate $4^{26}$ (281,474,976,710,656) possible combinations. Rather than attempting to generate all possible options, instead we created a Python script that executed a top-N greedy algorithm, as follows:

1. It first considered the annotators’ judgements for each original sentence and its three possible replacements.

2. At each sentence, it selected the best among the four options, such that the fluency and complexity of the overall article remained closest to the desired level. For instance, when attempting to create a text stimulus with low fluency but high complexity, it would favor local choices to achieve this result.

3. The algorithm output the top-N articles closest to the desired levels of fluency and complexity (after identifying an article, it would backtrack and select the next best choice until N articles were obtained).

In this manner, we identified articles that were all at average complexity of 2.0, but average levels of fluency of: 3.5 for our low fluency condition, 4 for our medium fluency condition, and 4.5 for our high fluency condition. Table 8.1 shows an excerpt from each condition, illustrating the different levels of fluency.
CHAPTER 8. METRICS FOR EVALUATING FLUENCY

<table>
<thead>
<tr>
<th>Fluency Level</th>
<th>Excerpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>“Much of this has been swallowed up by agriculture, there is still much land,” said farmers who don’t like agriculture. “We are committed to continuing to look for this small, elusive lizard, elusive and lizards.”</td>
</tr>
<tr>
<td>Medium</td>
<td>“Much of this has been swallowed up by agriculture, there is still much land,” said farmers who don’t like agriculture. “We are committed to continuing to look for this small, elusive and cryptic lizard.”</td>
</tr>
<tr>
<td>High</td>
<td>He added that much of this land has been taken up by agriculture, but there is still a lot more land to survey. “We are committed to continuing to look for this small, elusive and cryptic lizard.”</td>
</tr>
</tbody>
</table>

Table 8.1: Excerpt from each condition, illustrating different levels of fluency.

Validation.

As a final step to ensure that our resulting text stimuli met our conditions, we conducted an expert review, with the DHH literacy expert in our team. We asked the expert to consider the three versions of each article, each labeled with generic labels (A, B, C). We then asked the expert to rank them by fluency, from the perspective of how an average DHH reader would interpret the texts. Our team member ranked all of the articles in the same order as we had ranked them in the generation step; this outcome gave us further confidence that these articles were indeed at the three different levels of fluency.

8.2.3 Metrics

Considering that the goal of this study was to identify which metrics are able to distinguish different levels of fluency among texts that were carefully engineered to be at different levels
of fluency, we employed a diverse set of possible metrics, which had been identified in prior work on evaluating ATS and related fields. These included objective metrics such as reading speed and comprehension questions, as well as subjective judgements, which included score predictions and judgements of: readability, understandability, grammaticality, and system performance.

Reading Speed

This metric has been used in prior work measuring the readability of texts in the context of comparing the impact of user-interface elements, e.g. [77, 103], as well as evaluating ATS systems, e.g. [99], with the premise that more readable texts would lead to higher reading speed. Our previous methodological study on how to evaluate the complexity of texts among DHH adults did not identify reading speed as an effective metric, but it remains unclear whether it can distinguish texts at different fluency levels. We measured reading speed by measuring the time taken to read each text in minutes, then dividing the number of words in the text over the minutes taken.

Comprehension Questions

The comprehension questions employed in our study were obtained from our previous methodological study on how to evaluate the complexity of ATS output among DHH adults presented in Chapter 7. In that study, we had written two versions of each multiple choice question, at two different levels of linguistic complexity. That is, each question asked about the same fact, but the wording of the question item and answer choices varied in their linguistic complexity. Following the approach of that study, we created a quiz for each participant that, for each fact, randomly selected a low-linguistic or high-linguistic comprehension question, such that the quiz contained 3 of each level. This approach thereby controls for the difficulty level of responding to a question about any specific fact, while enabling the study to examine the efficacy of comprehension questions at different levels of
linguistic-complexity in their wording.

**Score Prediction**

This subjective-response item asked participants to predict how well they thought they had done on the comprehension questions, by indicating a numerical value from 0% to 100%. This item had originally been used in a prior study measuring how reading comprehension can predict academic achievement [125]. In our previous methodological study on how to evaluate the complexity of texts among DHH adults, this score prediction item had not been able to distinguish among complexity levels of texts. However, it remains unclear whether subjective judgements for this item might distinguish between different levels of fluency of automatically simplified texts among DHH readers.

**Likert Subjective Judgements**

Four subjective items used a 5-point scale: “Strongly Disagree” to “Strongly Agree.”

- **Understandability.** Judgements of understandability have been used in our previous methodological study on evaluating text complexity with DHH readers. The question item for an understandability judgement reads: “I was able to understand this text well.” While this item for understandability focuses more on the reader’s experience, the prior item for readability focuses somewhat more on the text itself.

- **Grammaticality.** Judgements of grammaticality have been used in prior work on evaluating the fluency of ATS output with high-literacy readers, e.g., [105, 141]. However, they have not been validated with DHH readers. We employ an item from prior work [105]: “This sentence is grammatically correct.”

- **Readability.** Judgements of readability have also been widely used in prior work, e.g., [99, 105], including our previous methodological study on how to evaluate text complexity with DHH adults. The question item for a readability judgement reads
“This text was easy to read.”

- **System Performance.** Prior methodological work on how to evaluate the quality of automatic video-captioning tools supported by ASR among DHH adults included a subjective measurement of whether the participant believed the ASR had done a “good job” [12]. Thus, we included an adapted version: “The automatic text simplification system did a good job simplifying this news story.”

### 8.2.4 Data Collection Procedure

After filling out an informed consent form for our IRB-approved study, participants met remotely over Zoom with a research assistant fluent in ASL. Participants were shown articles sequentially on a website created using the jsPsych library [31]. Articles were counterbalanced using a Graeco-Latin-Square schedule, which rotated the order of the articles (on different topics) as well as the order of the conditions (low, medium, and high fluency). Each participant read the six articles, with two of each condition.

After each article, participants responded to 6 comprehension questions as a single quiz, each containing the 3 low-linguistic complexity and 3 high-linguistic complexity questions as described above, with their order randomized and which 3 questions were at each level being counterbalanced across participants. Then, participants were asked to predict their scores in the comprehension quiz from 0 to 100. Participants then responded to the other subjective Likert-scale items (understandability, readability, grammaticality, and system performance).

Participants were given the option to take a break after the third article to avoid fatigue. After participants read all articles and responded to all questions, they filled out the sentence comprehension sub-test of the Wide-Range Achievement Test (WRAT), which has been validated as an indicator of literacy levels for DHH readers [64, 95]. Finally, participants completed a demographic questionnaire and were compensated with $40 USD for their participation.
8.2.5 Participants

Participants were recruited through advertisements posted on social media. We had a total of 29 DHH participants. Participants’ average age was 25.6 (range 18 to 35, SD = 5.4). A total of 10 participants identified as culturally Deaf [91], while 6 participants identified as deaf, 12 as hard-of-hearing, and one as Deaf-blind (This participant indicated that they were able to adjust the font size of the text on their web browser to read it comfortably during the study.) Participants self-identified as female (N = 21), male (N = 7), and one participant preferred not to say.

Participants’ average WRAT scores were 94.6 (range 67 to 128, SD = 16.8). In order to compare the effectiveness of the metrics for different literacy groups among our participant pool, as well as to identify the potential literacy bias of our metrics, we followed the approach of our previous methodological study presented in Chapter 7 and split our participants into two groups based on their median WRAT score (93):

- **WRAT-H**: 13 participants who had WRAT score higher than 93.
- **WRAT-L**: 16 participants who had a WRAT score of 93 or lower.

8.3 Results

Table 8.2 summarizes the results for each metric in terms of: **discriminative ability** (how effective each metric was in measuring differences among the three fluency levels of text) and **literacy bias** (whether metric scores were overall higher/lower depending upon the literacy level group of participants). These items correspond to the two Hypotheses presented in section 3.

In regard to discriminative ability, both reading speed and judgements of grammaticality were effective at measuring differences between some levels of text fluency with WRAT-H and WRAT-L readers. However, judgements of understandability and readability were only
### CHAPTER 8. METRICS FOR EVALUATING FLUENCY

<table>
<thead>
<tr>
<th>Metric</th>
<th>Discriminative Ability among Lower Literacy DHH Respondents (H1a)</th>
<th>Discriminative Ability among Higher Literacy DHH Respondents (H1b)</th>
<th>Literacy Bias (H2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading speed (Best Metric)</td>
<td>H1a was partially supported. Worked well to distinguish the lowest fluency texts from both the medium and highest fluency texts.</td>
<td>H1b was partially supported. Worked well to distinguish the lowest fluency texts from both the medium and highest fluency texts.</td>
<td>H2 was not supported. There were no measurable differences between lower and higher literacy readers.</td>
</tr>
<tr>
<td>High-linguistic-complexity</td>
<td>H1a was not supported. This metric was not discriminative between any text fluency levels.</td>
<td>H1b was not supported. This metric was not discriminative between any text fluency levels.</td>
<td>H2 was supported. Higher literacy readers had significantly higher scores than lower literacy readers.</td>
</tr>
<tr>
<td>comprehension questions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low-linguistic-complexity</td>
<td>H1a was not supported. This metric was not discriminative between any text fluency levels.</td>
<td>H1b was not supported. This metric was not discriminative between any text fluency levels.</td>
<td>H2 was supported. Higher literacy readers had significantly higher scores than lower literacy readers.</td>
</tr>
<tr>
<td>comprehension questions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score prediction</td>
<td>H1a was not supported. This metric was not discriminative between any text fluency levels.</td>
<td>H1b was not supported. This metric was not discriminative between any text fluency levels.</td>
<td>H2 was not supported. There were no measurable differences between lower and higher literacy readers.</td>
</tr>
<tr>
<td>Understandability</td>
<td>H1a was not supported. This metric was not discriminative between any text fluency levels.</td>
<td>H1b was partially supported. Worked well to distinguish between the lowest and highest text fluency only.</td>
<td>H2 was supported. Higher literacy readers had significantly higher judgements than lower literacy readers.</td>
</tr>
<tr>
<td>“I was able to understand this text well”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Readability</td>
<td>H1a was not supported. This metric was not discriminative between any text fluency levels.</td>
<td>H1b was partially supported. Worked well to distinguish the high-fluency texts from both the medium and low-fluency texts.</td>
<td>H2 was not supported. There were no measurable differences between lower and higher literacy readers.</td>
</tr>
<tr>
<td>“This text was easy to read.”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammaticality</td>
<td>H1a was partially supported. Worked well to distinguish between the lowest and highest text fluency only.</td>
<td>H1b was partially supported. Worked well to distinguish between the lowest and highest text fluency only.</td>
<td>H2 was not supported. There were no measurable differences between lower and higher literacy readers.</td>
</tr>
<tr>
<td>“This text was grammatically correct.”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System performance</td>
<td>H1a was not supported. This metric was not discriminative between any text fluency levels.</td>
<td>H1b was not supported. This metric was not discriminative between any text fluency levels.</td>
<td>H2 was not supported. There were no measurable differences between lower and higher literacy readers.</td>
</tr>
<tr>
<td>“The tool did a good job simplifying the news story.”</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.2: A summary of the results for each metric, for each of the hypotheses. Hypotheses were not supported if the metric did not discriminate between any complexity levels, and partially supported if it discriminated between some but not all pairs of complexity levels.
effective when used with the WRAT-H group. No other metrics were effective at measuring
differences among the fluency levels of texts.

Our analysis revealed literacy bias for understandability judgements and comprehension
questions (at both levels of linguistic complexity of questions). WHAT-H readers gave
higher responses for the understandability item, and they also achieved higher scores on
comprehension questions. To determine whether other metrics had statistically equivalent
response scores, when comparing WRAT-H and WRAT-L readers, we conducted a Two
One-Sided Test (TOST) for all remaining metrics. The TOST revealed that WRAT-L
and WRAT-H readers had statistically equivalent responses for: reading speed (within a
margin of 74 wpm), score predictions (within a margin of 15 percentage points), as well
as judgements of readability, grammaticality, and whether the tool had done a good job
(within a margin of 0.75 for all three).

For many readers, Table 2 may be a sufficient summary of the results. The remainder
of this section provides detailed results for each metric. First, we present the results for
Hypothesis 1 (the discriminative ability of the metrics). Significant results are accompanied
by Figures 8.1 through 8.4, which include whisker-plots for continuous data and stacked
divergent bar charts\(^3\) for Likert-type data, separated by literacy group. Then, we present
the results for Hypothesis 2 (the literacy bias of the metrics), which are accompanied by
Figures 8.5 through 8.7, which include the same type of plots outlined above, but in this
case compare the groups WRAT-L and WRAT-H.

\(^3\)Stacked divergent bar charts are recommended for the display of Likert-type data in [104], and these
graphs indicate the percentage of each response, with the stacked bars centered on the “neutral” response
item.
8.3.1 H1: Discriminative Ability

Reading Speed (wpm).

As illustrated in figure 8.1, the results from the Kruskal-Wallis test revealed significant differences in reading speed between the conditions for both groups (WRAT-L: $\chi^2 = 18.707$, df = 2, p-value < 0.001; WRAT-H: $\chi^2 = 27.41$, df = 2, p-value < 0.001). Pairwise Mann-Whitney U tests with Bonferroni corrections revealed statistically significant differences for both the WRAT-L and the WRAT-H group between the low and medium fluency conditions (WRAT-L: Z = -3.62, r = 0.64, p < 0.001; WRAT-H: Z = -5.01, r = 0.98, p < 0.001), and between the low and high fluency conditions (WRAT-L: Z = -3.47, r = 0.61, p-value < 0.001; WRAT-H: Z = -3.93, r = 0.77, p-value < 0.001).

Low-Complexity Comprehension Questions.

There were no statistically significant differences revealed by the analysis between the conditions for either group (p-value = 0.895 for WRAT-L, and p-value = 0.858 for WRAT-H).
Figure 8.2: Understandability judgements for H1 using a Likert-type agreement scale ( * = p < 0.05).

High-Complexity Comprehension Questions.

Similar to the low-complexity comprehension questions, no statistically significant differences were revealed for the high-complexity comprehension questions (p-value = 0.425 for WRAT-L, and p-value = 0.176 for WRAT-H)

Score Prediction.

No statistically significant differences were revealed for the score predictions with either group (p-value = 0.326 for WRAT-L, and p-value = 0.422 for WRAT-H).
Understandability.

The judgements of understandability only revealed significant differences between the fluency conditions for the WRAT-H group ($\chi^2 = 7.751$, df = 2, p-value = 0.02). Pairwise comparisons revealed differences between the low and high fluency conditions (Z = -2.4, r = 0.47, p-value = 0.016). Figure 8.2 summarizes these results.

Readability.

As illustrated in Figure 8.3, judgements of readability also revealed statistically significant differences between the fluency conditions for the WRAT-H group ($\chi^2 = 7.932$, df = 2, p-value = 0.019), with pairwise comparisons revealing differences between the low and high fluency conditions (Z = -2, r = 0.39, p-value = 0.045) and the medium and high conditions (Z = -2.04, r = 0.4, p-value = 0.041).
Grammaticality.

Judgements of whether the text were grammatically correct revealed significant differences for both groups (WRAT-L: $\chi^2 = 11.482$, df = 2, p-value = 0.003; WRAT-H: $\chi^2 = 13.355$, df = 2, p-value = 0.001). For both groups, pairwise comparisons revealed differences between the low and high fluency conditions (WRAT-L: $Z = -2.89$, $r = 0.51$, p-value = 0.004; WRAT-H: $Z = -3.54$, $r = 0.69$, p-value < 0.001). Figure 8.4 summarizes these results.

System Performance.

No statistically significant differences were revealed for either group for the judgements of whether the tool had done a good job (p-value = 0.191 for WRAT-L, and p-value = 0.058 for WRAT-H).
CHAPTER 8. METRICS FOR EVALUATING FLUENCY

Figure 8.5: Low-linguistic complexity comprehension questions scores for H2, with a max. value of 100% (* = p < 0.01).

Figure 8.6: High-linguistic complexity comprehension questions scores for H2, with a max. value of 100% (** = p < 0.001).
8.3.2 H2: Literacy bias

H2 was only supported for some of the metrics. In the cases in which it was supported, participants in the WRAT-H scored higher overall (in the case of comprehension questions) or provided higher judgements (of understandability). The metrics for which H2 was not supported all passed TOST equivalence tests, at the alpha=0.05 level.

- **Reading Speed (words per minute)**: \( Z = -0.506, p\text{-value} = 0.61 \). TOST equivalence testing revealed no significant difference (with a margin of 74 words per minute).

- **Low-Complexity Comprehension Questions**: \( Z = -2.994, r = 0.306, p\text{-value} = 0.003 \) (Figure 8.5)

- **High-Complexity Comprehension Questions**: \( Z = -3.964, r = 0.405, p\text{-value} < 0.001 \) (Figure 8.6)

- **Score Prediction**: \( Z = -1.59, p\text{-value} = 0.11 \). TOST equivalence testing revealed no significant difference (with a margin of 15 on a 0 to 100 scale).

- **Understandability**: \( Z = -2.627, r = 0.268, p\text{-value} = 0.009 \) (Figure 8.7)
• **Readability:** $Z = -0.21$, p-value = 0.8. TOST equivalence testing revealed no significant difference (margin 0.75 on a 1 to 5 scale).

• **Grammaticality:** $Z = -1.296$, p-value = 0.19. TOST equivalence testing revealed no significant difference (with a margin of 0.75 on a 1 to 5 scale).

• **System Performance:** $Z = -0.381$, p-value = 0.7. TOST equivalence testing revealed no significant difference (with margin of 0.75 on a 1 to 5 scale).

8.4 **Discussion**

In the following subsections, we first discuss the results of the discriminative ability of the metrics (H1), followed by the discussion of the results of the literacy bias of the metrics (H2).

8.4.1 **H1: Discriminative Ability**

The best metric overall was reading speed; it was able to measure differences between more text fluency levels and worked with both WRAT-H and WRAT-L participants. To the best of our knowledge, reading speed had only been used in prior work as a way to measure reading comprehension, and in our previous methodological study on evaluating the complexity of simplified texts with DHH readers, this metric had not been an effective way to measure text complexity. This lack of use in prior work for evaluating fluency may come from the fact that most evaluations of fluency have relied upon an expert reader making a side-by-side comparisons between the original text and a simplified text. Measuring reading speed is less suitable when a participant is making such side-by-side comparisons. Our findings suggest that displaying the output text to a DHH participant and measuring their reading speed is also an effective way of measuring the fluency of that text.

The other objective metrics employed in our study were the comprehension questions written at different complexity levels. However, given that these were not able to distinguish
between any of the fluency conditions with either one of the groups, we do not recommend their use for evaluating the fluency of simplified texts. This result is in line with our previous methodological study on evaluating the complexity of simplified texts, in which comprehension questions had only worked under a number of conditions (namely, that the texts were far enough in complexity, that the group had lower literacy, and that the questions were written in low complexity). As highlighted in that prior study, comprehension questions may have value in keeping participants engaged with the reading task, even if they are not useful as metrics for evaluation of the text itself. In this case, we found no evidence of them being effective in distinguishing between the fluency levels of texts included in our study.

Among the subjective metrics, judgements of grammaticality were the most effective; this metric was able to reveal the difference between the low and the high fluency conditions of text. This metric may be used in cases where reading speed may not be available, e.g., if it is not feasible for a researcher to capture reading time due to their study setup. Among the remaining subjective metrics, we found that judgements of readability and understandability were only effective with the WRAT-H participants. Thus, in our study, we found that a greater number of our metrics were effective among higher-literacy participants, as compared to the number of metrics that had been effective among lower-literacy participants. These results are in line with prior methodological work on evaluating the quality of automatic video-captioning tools with DHH readers [12]; in that prior study, more of the caption-quality metrics that researchers had investigated were effective among their higher-literacy readers. We speculate that this may be because these judgements (of understandability and readability) require higher metacognitive awareness when disfluencies are introduced in the text—and are therefore more suitable among higher-literacy readers.

In summary, future researchers who wish to make use of the methodological findings of our study should utilize reading speed to measure fluency of ATS texts among DHH readers. As an additional or alternative measurement, we secondarily recommend the use of a Likert-scale subjective judgement of the grammaticality of the text.
8.4.2 H2: Literacy Bias

As discussed previously, finding that a metric has a literacy bias does not necessarily mean that it is undesirable to use it within a study. It simply means that if a researcher were to use such a metric, then they should also report the literacy levels of the participants in their study, e.g., WRAT scores. Without information about the literacy level of the specific participants in a study included in a publication, it would not be possible for readers of that paper to compare the results to other published work, since the literacy characteristics of the participants may have influenced the scores.

In the case of our current study, we found that three metrics had a literacy bias: Likert-scale subjective judgements about the understandability of a text, comprehension questions written at a higher complexity level, and comprehension questions written at a lower complexity level. Since none of these three metrics had actually been effective in measuring text fluency in our study, our finding of literacy bias for these three metrics may be moot for researchers interested in measuring text fluency. However, since comprehension questions and subjective judgements of text complexity are used in other text-evaluation contexts, our findings may be of interest to such researchers. For comprehension questions (at both complexity levels), participants in the WRAT-H group scored significantly higher. For understandability judgements, WRAT-H participants gave texts higher scores. We speculate that the higher literacy of WRAT-H readers made it easier for them to understand texts and to answer comprehension questions.

For the remaining metrics (reading speed, score prediction, readability and grammaticality judgements, as well as judgements of whether the tool had done a good job), we did not find evidence of a literacy bias, i.e., there were no statistically significant differences between the responses from WRAT-H and WRAT-L readers. TOST confirmed that responses were statistically equivalent, within margins. This finding should be interpreted with caution: Prior work on evaluating the complexity of simplified texts had observed literacy biases in reading speed, score prediction, and readability judgements. In the general
case, we believe that it is reasonable that literacy biases may exist for these metrics, e.g. that higher-literacy readers read more quickly than lower-literacy readers. Our findings should be interpreted more narrowly: In the context of a study in which participants were asked to read texts that contained dis-fluencies introduced by ATS technology, our findings suggest that literacy bias for these metrics was less substantial. From the perspective of a researcher who is only evaluating fluency of texts, it may be less necessary to report the literacy level of DHH participants in the study. However, reporting of participant literacy level is generally recommended, especially when evaluating the complexity of the texts, as found in our previous methodological study presented in Chapter 7.

8.5 Limitations and Future Work

There were several limitations in our study, as well as several possible avenues for future work. First, we cannot guarantee that some of the metrics found to be ineffective in our study would not work in studies with larger sample sizes and more statistical power. Of course, since evaluations of ATS typically involve fewer participants than the number included in our current methodological study, our contribution is still useful to researchers. Namely, metrics that did not reveal significant differences in our study would be unlikely to reveal significant differences in evaluations with an equal or smaller sample size. Future work, however, could benefit from employing participants across broader ranges of literacy levels. Furthermore, our study only included texts in the science domain and thus future work could explore whether our findings are generalizable to other text domains.

In this study, we only evaluated one type of comprehension questions, namely, multiple-choice questions. Thus, our findings may not generalize to other types of comprehension questions, e.g., cloze tests or summarizing tasks. Further, while we did not recruit participants with fluency in American Sign Language (ASL) specifically, future work could also incorporate comprehension questions recorded in ASL as a possible metric for use among participants who are ASL signers. Other methods of evaluation, such as eye-tracking, could
be explored during in-person studies, which were not possible at the time this study was conducted due to COVID-19 restrictions.

Our current study specifically focused on how to evaluate the fluency of automatically simplified texts, which may be damaged as a result of grammatical errors being introduced. However, as mentioned during the related work, semantic errors may also be introduced in the simplification process. As those types of meaning-preservation errors are typically evaluated by asking expert readers to examine the original and simplified texts side-by-side, future work could explore whether this type of evaluation is possible among DHH readers, or whether some of the metrics employed in our study (e.g., comprehension questions) can effectively distinguish the loss of information caused by those errors.

Finally, it is our hope that future work will include DHH users in actual user evaluations of the fluency of the output of ATS technologies. Furthermore, future work can include explorations of the preferences of DHH readers for different interface parameters of ATS-based reading assistance tools, as well as their specific linguistic needs and how ATS tools can be adapted to better support those.

8.6 Conclusion

In this study, we conducted methodological research to evaluate several metrics in terms of their effectiveness for evaluating the fluency of automatically simplified texts among DHH adults across a range of English literacy levels, as well as potential literacy biases when using those metrics. Our findings revealed that reading speed and participants’ subjective judgements of grammaticality were effective for distinguishing text fluency levels among DHH participants across a range of literacy levels. Judgements of understandability and readability were only effective among participants with higher literacy. Literacy biases were only observed in comprehension questions scores and judgements of understandability, i.e., participants with higher literacy had more positive scores and judgements overall. Our findings provide methodological guidance for the design of studies with DHH participants
evaluating ATS technologies. Namely, we recommend the use of reading speed and judgments of grammaticality when evaluating the fluency of simplified texts among DHH adults at a range of literacy levels.
Epilogue to Part II

Part II of this dissertation focused on the methods for evaluating various characteristics of the output of ATS-based reading assistance tools among DHH adults at different literacy levels. After identifying a lack of prior methodological guidance for evaluating the complexity and the fluency of simplified texts, two methodological studies were conducted with DHH participants at different literacy levels to investigate what are the best metrics for evaluating those two key aspects of the output of ATS. More specifically, Part II of this dissertation explored the following research questions:

RQ2.1: When English texts of different complexity levels are evaluated by DHH individuals with (a) lower English literacy skill and (b) higher English literacy skill, which metrics reveal statistically significant differences among texts that vary in a specific characteristic of the output of ATS? We referred to this characteristic as the discriminative ability of the metric, and this is a desirable characteristic. We investigated this research question for two characteristics of the output of ATS:

RQ2.1.1. Which metrics exhibit a discriminative ability in evaluating the complexity of simplified texts among DHH adults at the different literacy levels? The results from our study summarized in Chapter 7 suggest that comprehension questions only work under a number of conditions
(when they’re used with participants with lower literacy, the questions themselves are written in language simple enough for participants to understand and the differences in complexity level of the texts are of several grade-levels).

**RQ2.1.2.** Which metrics exhibit a discriminative ability in evaluating the *fluency* of simplified texts among DHH adults at different literacy levels? The results from our study summarized in Chapter 8 revealed that among our DHH participants across a range of reading literacy, their reading speed and subjective judgements of a text’s grammaticality were effective at distinguishing between levels of fluency in automatically simplified texts.

**RQ2.2.** In an overall analysis of response scores for all texts varying in a particular characteristic of the output of ATS, when we compare response scores between DHH individuals in the higher-literacy and lower-literacy groups, do we observe a significant difference? We refer to this as the *literacy bias of the metric*. While this is not necessarily a problem with the metric that would prevent its use in evaluating text complexity, researchers who use this metric must ensure that they report the literacy skill level of their participants in a study, in order for results across studies to be comparable. Similar to RQ3.1, we explore this question for the same two characteristics of the output of ATS:

- **RQ2.2.1.** Which metrics exhibit a literacy bias when evaluating the *complexity* of simplified texts among DHH adults at the different literacy levels? Our results suggested that, when evaluating texts at different complexity levels, all of the metrics examined in our study exhibited a literacy bias, with participants with higher literacy scoring higher or providing more positive subjective judgements.

- **RQ2.2.2.** Which metrics exhibit a literacy bias when evaluating the *fluency* of simplified texts among DHH adults at different literacy
levels? Only comprehension questions (at both levels of linguistic complexity examined) and judgements of understandability revealed a literacy bias.
Part III:

Design and Evaluation of ATS-based Reading Assistance Tools
Prologue to Part III

There are numerous design parameters involved in the design of ATS-based reading assistance tools, including (but not limited to) the type of simplification, how much text is replaced at once, the placement of the simplifications, as well as how much autonomy is provided to the user. Thus, Part III of this dissertation focuses on investigating the benefits of specific types of simplifications, as well as the preferences of DHH adults for the various design parameters involved, to inform researchers working in the design of these technologies. Thus, the studies presented in Part III of this dissertation involve participants’ interactions with fully functional prototypes that allow customization of the various design parameters involved at different stages of the simplification process.

Part III begins with Chapter 9, which outlines prior work from the perspective of the user interfaces employed in their ATS-based reading assistance tools, as well as prior HCI work on user autonomy. Chapter 10 then describes an experimental study focused on investigating the benefits obtained by DHH adults from specific types of simplification, namely lexical simplification, as well as their preferences for the level of autonomy the tools provide. Then, Chapter 11 identifies the various design parameters and their settings that have been employed in prior work, and summarizes our work investigating the preferences of DHH adults for the various possible settings of the design parameters identified, when using specific types of text simplification. Finally, Chapter 12 summarizes our evaluation of an ATS-based reading assistance tool that reflects the current state of the art of ATS technologies, among DHH adults engaging in reading about science news.
More specifically, Part III of this dissertation explores the following research questions in the context of a reading-assistance software system for DHH users, capable of providing lexical simplification for words on webpages:

**RQ3.1:** Does providing a lexical-simplification reading-assistance tool affect DHH users’ perception of their ease in reading a text or perception of how well they understood it? (This question is explored in Chapter 10)

**RQ3.2:** Are DHH users’ subjective preferences about lexical-simplification tools affected by whether they are provided with greater autonomy (i.e. in requesting simplifications on-demand or seeing what words have been replaced)? (This question is explored in Chapter 10)

After investigating the specific benefits of lexical simplification, and users’ preferences for autonomy, in this part we also investigate the following research questions in the context of using a fully-functional prototype capable of providing both syntactic and lexical simplifications, and the combination of the two:

**RQ3.3.** When employing ATS-based reading assistance tools that provide syntactic or lexical simplifications, what are the preferences of DHH users for the various parameters involved in the design of these tools? (This question is explored in Chapter 11)

Finally, using the settings identified through RQ3.3, as well as the methodological guidance identified through RQ2.1. and RQ2.2, we conclude by exploring the following research question:

**RQ3.4.** When providing an ATS system that provides both lexical and syntactic simplification features, as well as hybrid approaches
that combine lexical and syntactic simplification to DHH users, how do users interact with and benefit from such system, and how do their preferences vary for the various simplification approaches? (This question is explored in Chapter 12)
Chapter 9

Background and Related Work

As context for Part III, this section summarizes work on the user interfaces of systems that simplify text or modify its appearance for readability. Prior HCI research on the role of autonomy is also discussed.

9.1 User Interface for Text Simplification

Aside from transforming the text itself, research on reading-assistance tools has explored systems that modify the layout or visual appearance of a page of text, including research on the user-interface design of such systems, e.g. inserting whitespace between sentences in a paragraph [139], highlighting portions of paragraphs [72], annotating text [26], or changing font size and line spacing [102]. This research indicates how visual design affects text readability.

Prior reading-assistance research with lexical simplification has used various user-interface designs, varying in the level of control users are provided and visibility as to what words have been replaced. However, most prior user studies on lexical simplification for reading assistance have focused their investigation on the quality and effect of different simplification methods, e.g. [92, 99, 105], rather than on the user-interface design. In one such
study, Rello et al. investigated the use of different simplification systems to assist people with dyslexia, and one of their tools provided simplifications on-demand. Participants felt that texts were easier to understand when using the version that provided simplifications on-demand, as compared to the original text and a version that automatically simplified text before the user saw it [101]. Thus, the authors indicated a need for further research to investigate the user interface of such systems. However, to the best of our knowledge, the effect of providing on-demand choice in the user-interface design of lexical simplification has not been further investigated.

9.2 User Autonomy

Prior HCI research has investigated the effect of providing users with autonomy, with the concept sometimes referred to by different terminology, e.g. agency, independence, or self-determination [22, 43, 45]. This research has found that fine-tuning the levels of users’ autonomy can increase motivation and their desire to use software [43]. However, there are tradeoffs: For instance, increased autonomy in an interactive system may also increase the complexity or cognitive load, and user autonomy may need to be limited in order to provide greater security or privacy [45]. A workshop at CHI 2014 explored the topic of autonomy, which the organizers defined as relating to a sense of agency or self-regulation [22]. The workshop identified four spheres of autonomy in HCI: (1) within a software design, which refers to the sense of control the user has when using software, and has been shown to impact motivation and engagement [43]; (2) the creation of assistive technologies to increase autonomy in daily life; (3) enabling users to design their own technologies; and (4) helping users develop a sense of autonomy through technology. While simplification as reading assistance may fall under assistive technologies to assist in daily life, our work focuses on investigating the effect of user autonomy when using text simplification within the software design.

In the accessibility field, Rello et al. [101] had speculated that providing greater control
with on-demand assistance explained some of the benefit of simplification for users with dyslexia in their study. ATS is an artificial-intelligence (AI) technology, and in a recent paper on the ethics of AI-based assistive technologies for DHH users, authors have advocated for enabling DHH users to make their own choices about when to deploy AI-based systems [62]. This sentiment intersects with recently published guidelines for human-AI interaction [7], which recommend that systems enable efficient invocation of AI services when needed. Other accessibility researchers have advocated moving away from paternalistic approaches in accessibility technology design, and instead towards approaches that provide users with more choice or that support “inter-dependence” [50]. Given prior work on the benefits of providing users with on-demand AI-powered assistive tools, we investigate whether DHH users of ATS reading-assistance tools prefer designs that provide them with greater control over which words are simplified or greater visibility of what words have already been modified.
Chapter 10

Automatic Text Simplification

Tools for DHH Adults: Benefits of Lexical Simplification and Providing Users with Autonomy

10.1 Introduction

We investigate applications that provide ATS for complex texts, in particular to support reading texts online. ATS can scan a webpage, find complex words or phrases, and replace them to make the text easier to understand. Prior work had evaluated syntactic simplification with DHH users (simplifying the structure of sentences) [73]. However, prior work has not evaluated DHH users’ perception of automatic reading-assistance tools based on lexical simplification (replacing a difficult word with a simpler synonym), nor of hybrid systems that combine both syntactic and lexical approaches. As a first step, this study investigates whether DHH users perceive a benefit from lexical simplification alone.

Prior reading-assistance work has investigated using ATS with various groups, e.g. peo-
ple with aphasia [34], people with dyslexia [99, 101], and second-language learners [9]. However, most of these studies have focused on improving the quality of the underlying ATS technology, with less focus on the human-computer interaction (HCI) aspects, e.g. how users actually interact with the software itself, what modifications they can make to the software, how much decision-making they have to do, and how much control they have over what the software automatically does for them. Having more ATS reading assistance may not necessarily be better: as with any assistive technology that uses artificial intelligence, there is risk that it may make errors or that users may experience a loss of agency, if they only see a filtered or modified view of the original text. Various prior work in HCI has investigated the issue of autonomy or related concepts [22, 44]. In the context of ATS, we define autonomy as how much control a user has over what words are simplified or visibility of what words have already been simplified. We speculate that providing DHH users with a desired level of autonomy may result in an improved user experience and motivation to use the software again.

10.2 Method

We first conducted a preliminary study with 12 DHH participants to evaluate a wide variety of prototype designs for the user interface of a reading-assistance system with lexical simplification. Our goal was to select a manageable subset of conditions for our second study. Next, we conducted an experimental study in which 25 DHH participants compared working prototype systems. That final experiment included a baseline system that did not provide any lexical simplification (in support of RQ3.1) and a system that automatically applied lexical simplifications before the users saw the text (for RQ3.2).
10.3 Prototype

To investigate RQ3.2, we needed a set of prototype variations for reading-assistance software that would enable us to explore the concept of user autonomy. Based on prior work on lexical simplification, we identified two principal design dimensions that appeared relevant to this issue: how the user initiates a request for a simplification (henceforth referred to as user initiative) and whether the system provides some visible indication that a simplification has occurred (henceforth referred to as change visibility). The context of our prototype is that it would provide reading assistance while the user is visiting web pages, and it would therefore consist of some type of web browser enhancement or plug-in, an approach recommended in [129].

Within the user initiative dimension, we identified five levels based on an analysis of prototypes in previous research on text simplification, with each level varying as to the degree of autonomy it provided to the user:

1. As a baseline, we provided users with a non-interactive prototype, in which the system automatically replaced all the words with simpler versions available before the user even saw the text [73, 101]. As illustrated in Figure 10.1, we refer to this level as automatic.
2. Providing some additional user initiative, the replace all version allowed the user to click on a button to simultaneously replace all words on the page for which simpler versions were available.

3. ATS software may not be able to provide a simpler replacement for every word in a text, or it might determine that the existing text is already simple enough for the reader. Researchers have therefore sometimes used text decoration, such as highlighting or text color differences, to suggest to the reader which words in the text the user may wish to simplify \[14, 86, 99, 101\]. We refer to this level as suggestions.

4. Given that this text decoration would be added by the system, we developed a variation of the suggestions level that allowed the user more autonomy by providing a button to activate or deactivate this decoration. In this way, the user would be able to experience the web page in its original intended appearance without any visual decoration, and the system would not be suggesting that they might want to simplify particular words. We refer to this level as toggle suggestions.

5. Researchers in prior work have also created reading-assistance tools in which the user was allowed to request a simplification on any word they felt was complex, without providing the user with any text decorations \[101\] to suggest that they may want to simplify particular words. We refer to this level as no suggestions.

In addition to design differences that influence user initiative over the simplification process (listed above), we identified a relatively orthogonal design dimension in which prototypes vary in the degree of change visibility, that is, whether the system indicates to users retrospectively which words have been modified. Along this dimension, we identified four levels of increasing autonomy, with the perspective that designs likely provide more autonomy if they clearly indicate to the user which words have been changed, making it easy to see what the original word had been, and minimize blocking the onscreen text:

1. Researchers in prior text simplification work have investigated providing simplified
versions of text in a separate region to the side of the window displaying the text [105, 114], thereby leaving the original word in the text visibly unchanged. We refer to this level as sidebar.

2. Researchers have also used pop-ups (small rectangles that appear floating above the text near the word that has been hovered-over) to provide the reader with a simpler word while still allowing the user to see the original text [86, 99, 101]. We refer to this level as pop-up.

3. While the two options above would actually leave the original word in place in the text, in this next design variation, the word is replaced with a simpler synonym, but some text decoration, e.g., a background highlight color, is used to convey visually that a change has happened [14]. We refer to this level as trace.

4. Lastly, we include a version, as a baseline, in which the word is replaced without leaving behind any visible indication that it had been replaced [17, 73, 101]. We refer to this level as no trace.

10.3.1 Simplifications

While the state of the art in ATS has been increasing in accuracy in recent years, for both the identification of complex words and the selection of appropriate synonyms (given the specific meaning of the word in context) [80], ATS is still not perfect, with many potential places for errors, which could be detrimental to reading rather than helpful [110]. Thus, to ensure that the quality was kept constant across texts, we used a Wizard-of-Oz approach to create the simplifications for our prototype by using human-produced simplifications aided by Par4Sim [138] (a semantic writing tool that automatically identifies complex words, highlights them and suggests potential simplifications) and SimplePPDB++ [80] (a resource containing lexical paraphrase rules with readability scores). A team of three (2 Deaf researchers and 1 hearing researcher whose native language is not English) used the following protocol to obtain lexical simplifications for stimuli:
1. The researchers input the text into Par4Sim to begin identifying complex words. While they used the complex words suggestions from the system, they also read through the text in case there were any complex words the system had missed. The first-hand perspective of the researchers as members of the Deaf community or as someone who had learned English as a second language was valuable in their identification of potentially difficult-to-read words in the text.

2. After identifying all complex words, they would look at the suggestions from Par4Sim.

3. If no Par4Sim suggestions were considered simpler or fitting to the context, then they would query the SimplePPDB++ to consider other suggestions.

4. If nothing above provided satisfactory simplifications, then they would perform a Google Search for a synonym, until a simpler synonym was found.

10.4 Preliminary Study

The factorial combination of both design dimensions, with their 4 or 5 levels, would yield 20 prototype variations, which was too many for comparison in an experimental study. So, it was essential for us to conduct a preliminary study to reduce this number. Thus, the main goal of our preliminary study was to identify potential design variations that may be preferred by people who are DHH for software to provide lexical simplification on-demand.

10.4.1 Materials

Reading Stimuli

To show our participants various designs for an ATS system, we needed to select a webpage with text that could be simplified. Using an article from the Smithsonian’s website Tween-Tribune, which provides news articles at various levels of reading complexity, we created non-interactive high-visual-fidelity prototype videos, demonstrating each level of our two
design dimensions discussed above. These videos were created using Safari in an iMac 27-inch desktop. Each video lasted approximately 10 seconds and showed a user requesting 3 simplifications on the first paragraph of the text. Given that this preliminary study was formative in nature, we did not attempt to display 20 videos to users, with all combinatorial possibilities, which may have been overwhelming. We instead provided a set of video prototypes for each design dimension. Thus, videos for the user initiative design dimension showed replacement of words when simplification was triggered, without any visual trace of which words had been replaced. The videos illustrating various levels of change visibility showed a word being clicked to trigger a replacement, without any visible decoration on the word beforehand to suggest replacing it.

Additional Demographic Questions

In addition to standard demographic questions, e.g. age, we also asked participants about how they identify themselves, e.g. Deaf, deaf, hard-of-hearing, etc. In addition, because there is great diversity among DHH individuals in the U.S., especially in regard to reading literacy skills, we also gathered additional demographic information about our participants: Specifically, we also asked our participants to respond to an instrument that measured their English literacy skills and to respond to a psychology questionnaire for measuring an individual’s inclination toward autonomy, since this was a focus of our work. Our intention was to provide future researchers with sufficient information about our participants, to facilitate any future replication of our study.

- English Reading Literacy: Following in the methodology of [12], we administered the sentence comprehension subtest of the Wide Range Achievement Test 4th Edition (WRAT4) [132] to measure literacy because it is brief, does not use audio stimuli, and has been previously validated with users who are DHH [92].

- Autonomy: After considering some relatively longer psychological instruments for measuring an individual’s inclination towards autonomy, e.g. [32], we chose to admin-
ister the relatively brief Index of Autonomous Functioning (IAF) [131], consisting of 15 Likert items.

10.4.2 Data Collection Procedure

This formative study used an interview-based methodology with a set of video prototypes, illustrating various levels of design dimensions that were discussed with the user. After the participant signed a consent form and answered demographic questions, the interviewer explained the basic concept of ATS tools and the scenario in which the user may be reading a web page that contains unfamiliar words. Next, the interviewer discussed each design dimension and showed the participant all the videos for each. After each video, participants were asked to discuss the advantages and disadvantages of each; the goal of this was to encourage the participant to carefully consider each option. After all the videos for each dimension were shown, the participant was presented with a screen with small pictures of each levels simultaneously, and the interviewer asked participants to rate how likely they were to use each version on a 6-point scale (using an even number to avoid a middle point during this preliminary study). At the end of the study, participants were asked to provide feedback on the various visual text decorations that had appeared in the prototypes, to determine whether further adjustment of the designs would be needed. Finally, participants completed the IAF and the sentence comprehension subtest of the WRAT test was administered.

10.4.3 Recruitment and Participants

Deaf or Hard-of-Hearing participants were recruited through email and social media; participants met with a research assistant for a 70-minute study in a private office and received $40 cash compensation. This study was approved by our university’s institutional review board for the protection of human subjects. The studies were conducted in English or American Sign Language (ASL), at the participant’s preference. We recruited a total of
Figure 10.2: Participants’ response to a 1-to-6 scalar question on how likely they were to use a system based on each level of the user initiative and change visibility design dimensions, from 1 (very unlikely) to 6 (very likely). This small study was insufficiently powered for statistical difference testing, but it was meant to guide prototype selection for the main study. Responses are displayed using a diverging stacked bar graph, as recommended in [104].

12 participants. Participants’ self-identified gender included 7 male and 5 females, mean age of 24 (SD = 1.5). While 8 participants identified as culturally Deaf [91], 3 identified as Hard-of-hearing and 1 as deaf. Participants’ standardized average WRAT score was 84.1 (SD= 14.03); the average WRAT score for U.S. adults is 100 [95]. The participants’ average IAF score was 3.8 out of 5 (SD = 0.32).

10.4.4 Results and Discussion

The goal of this formative study had been to help us to narrow our set of prototypes to be included in our final experimental study. We knew in advance that, in our final experimental study, we wanted to include two “end points” of the design space: (1) a baseline condition in which users see a webpage without any text-simplification assistance provided and (2) a version in which the complex words are automatically replaced before the user sees the text without any visual indication of which words have been changed (thereby providing
users with the least autonomy). Thus, our goal was to identify 1-2 additional prototypes to include in our final experiment. Notably, this formative study included few participants, and it was not designed to be sufficiently powered for statistical difference testing. Our goal was simply to avoid making an arbitrary choice about which prototypes to include in our final experiment.

As Figure 10.2(a) illustrates, for the user initiative design dimension, users expressed most interest in the toggle suggestions prototype, while the no suggestions prototype was a close second. The replace all and automatic were the least preferred options. DHH users were most interested in designs that did not provide visual suggestions about which words can be simplified or enable users to toggle these visual suggestion indicators on or off.

As for the change visibility dimension, in Figure 10.2(a) the trace was first place, with pop-up second, no signifier in third, and sidebar in fourth. Thus, users were interested in designs that conveyed which words had been changed by leaving a visual trace (e.g. some decoration on the text) or by conveying simpler synonyms for a word using a pop-up (without replacing the word in the text itself).

While we had found the top options for user initiative (toggle suggestions and no suggestions) and for change visibility (trace and pop-up), our goal had been to identify two prototypes for our later experiment. So, we decided to combine design elements as follows: (a) toggle suggestions combined with trace and (b) no suggestions combined with pop-up. Our rationale was that since the toggle suggestions option already introduced text decoration into the interface, it was more natural for that to be paired with trace (which also used text decoration to indicate modified words).

Based on feedback from DHH participants in this formative study, we selected to use gray background color highlighting to indicate suggestions and yellow background color highlighting to indicate words which had already been replaced. As part of our preliminary study, we also discussed various design options with participants for how various elements of the user interface might appear, and we asked participants to share their thoughts about these alternatives. Participants responded negatively to the concept of using squiggly un-
Figure 10.3: One of the articles used as stimuli in our experimental study, along with zoomed-in views of how text appeared under the four different conditions, e.g. clicking on a word in the *decoration* condition or hovering over a word in the *pop-up* condition.

derlines, e.g. similar to the red underlines used in Microsoft Word to indicate spelling problems, because participants felt like such decoration indicated errors on the page. In addition, traditional underlining or font color changes were not preferred since they may look similar to hyperlinks on a web page.

### 10.5 Experimental Study

The goal of our main experimental study was to answer our research questions. RQ3.1 considered whether DHH users report subjective improvements in the readability or understandability of texts when using automatic reading-assistance tools with lexical simplification, and RQ3.2 considered whether DHH users’ subjective impressions are affected by the degree of autonomy provided by the system. We examined several subjective measures that had been used in prior work on reading assistance tools, which are discussed below. In contrast to our preliminary formative study above, in this subsequent experimental study, users were provided with interactive prototypes. Guided by the results of our preliminary study, we implemented the following four interactive prototypes:

1. A webpage with all the text in its original form without any simplifications applied nor available was provided as a baseline. We refer to this condition as *original*.
2. We provided a version in which all complex words were already replaced, without any visible trace, before the user saw the text. In terms of the autonomy provided to the user, this was our least autonomous condition. We refer to this condition as automatic.

3. We combined the toggle suggestions level for user initiative with the trace level for change visibility, to produce a system that provided decorations to suggest complex words or to show words that had been simplified. Since the system decorated and transformed text upon user request, this was our medium autonomy condition. We refer to this condition as decoration.

4. We combined the no suggestions level for user initiative with the pop-up level for change visibility, to produce a system that provided a temporary pop-up near any word that can be simplified if the user hovers their pointer over it. This was our most autonomous condition given that the user could select any word and the text was never replaced. We refer to this condition as pop-up.

Figure 10.3 illustrates each prototype. Readers should note that the name “pop-up” had referred to a level of change visibility in our formative study and now as the codename of one of the four prototypes in our final experimental study. Similarly, “automatic” had referred to a level of user initiative in our formative study and now as a codename for a prototype in our final experimental study. Given that the prototypes corresponding to each of these labels were essentially identical in both the formative and experimental studies, these seemed to be the most logical codenames for the prototypes, but we do note this identical nomenclature.
10.5.1 Materials

Reading Stimuli

In this study, we selected four texts from the ScienceDaily website, as they were freely available and were suitably complex, with Flesch-Kincaid grade levels ranging from 12.3 to 17.6. All of the articles consisted of research news about scientific studies concerning flying creatures (honey bees, birds and fruit flies). All four texts had a similar length: between 617 and 682 words, with an average of 654. After identifying complex words and their respective simplifications using our Wizard-of-Oz approach as described above, between 40 and 70 complex words were identified in each text. After applying all simplifications, the Flesch-Kincaid grade levels ranged from 11.6 to 16.4. These articles originally contained a summary or abstract at the beginning, which we omitted to prevent our participants from simply reading the summary.

Subjective Evaluation Questions

To collect subjective feedback from our participants about each of the prototypes, we asked them to respond to three 5-point Likert-type questions for each prototype. These questions were selected based on having been used previously in prior work that had evaluated automatic text simplification systems. Specifically, in our study, participants indicated their agreement with each of the following:

- “This text was easy to read.” Several prior studies have used a question like this, e.g. [101, 139];

- “I was able to understand this text well.” Prior studies asked participants to indicate their perception of how well they understood the text, e.g. [101, 102, 139].

At the conclusion of the entire study, participants were reminded of the three prototypes they had used that had provided some text simplification, and they were asked to indicate whether they “would be likely to use” each. Our rationale for not asking this question
about the baseline “original” condition was that asking the user how likely they would be
to use a tool, after simply showing a user a web page without any text-simplification tool
provided, would have likely led to confusion among participants.

Comprehension Questions

To encourage participants to pay attention to the texts that were displayed during the study
and to attempt to use each prototype to understand the content, we informed participants
that there would be a quiz after each text. We developed 3 comprehension questions for
each text. These questions were modeled based on [35], using a main idea, main fact and
incidental question. However, considering that our previous work presented in Chapter 7
had suggested that subjective questions were effective for evaluating complexity among DHH
adults, the comprehension questions were mostly intended to keep participants engaged with
the text.

10.5.2 Data Collection Procedure

Participants first signed consent forms and responded to demographic questions. Then,
they were shown a set of instructions describing the operation of each prototype. While
the four texts were presented in the same order, our four conditions were rotated using a
balanced Latin-Squares schedule to ensure that neither variations in the complexity of the
texts nor the order of presentation of the conditions affected our results. At the end of
each text, participants responded to the two subjective questions (listed above) and then
answered three comprehension questions about the text. At the conclusion of the study,
participants were shown images of the three conditions that involved text simplification
(i.e. automatic, decoration, and pop-up) in the same order that each individual participant
had seen them. Then, we asked them to rate how likely they were to use each one of the
conditions using a 5-point Likert-type question. After providing their response for each
condition, they were asked open-ended questions about the rationale for their responses.
At the end of the experiment, participants were administered the IAF and WRAT tests using an online form, and then informed of the Wizard-of-Oz nature of our system as part of a debriefing session.

10.5.3 Recruitment and Participants

This study was approved by our university’s institutional review board. Participants were recruited through email and social media, and the advertisement included our screening criteria: identifying as Deaf or Hard-of-Hearing. Participants met with a research assistant for a 70-minute study in a private office and received $40 cash compensation. The studies were conducted in English or American Sign Language (ASL), at the participant’s preference. We recruited 25 participants with mean age of 23.5 (SD = 2.27), including 10 self-identified as male, 14 as female and 1 as non-binary. While 16 participants identified as culturally Deaf [91], 3 identified as Hard-of-hearing and 6 as deaf. The participants’ average IAF score was 3.68 out of 5 (SD = 0.36), with an average WRAT score of 83.04 (SD = 12.38).

10.5.4 Results

Figure 10.4 displays participants responses to the subjective questions in our study that compared all four conditions, including the baseline condition. In this graph, significant pairwise differences are indicated with asterisks as follows: *** if $p < 0.001$, ** if $p < 0.01$, or * if $p < 0.05$. The statistical analysis performed for each question is described below.

After reading each text, we asked participants to indicate how much they agreed with: “This text was easy to read.” A Friedman test indicated a significant difference ($\chi^2 = 7.9682$, $p = 0.047$), yet post-hoc pairwise comparison using Wilcoxon Signed Rank tests with Bonferroni corrections did not reveal any significant pairwise differences.

We asked participants to indicate how much they agreed with: “I was able to understand this text well” for each condition. A Friedman test indicated a significant difference ($\chi^2 = 16.261$, p-value = 0.001), and post-hoc pairwise comparison using Wilcoxon Signed
Figure 10.4: Participants’ responses to questions about all four conditions in the experimental study, including subjective Likert-scale responses for (a) the text was easy to read and (b) I was able to understand this text well, with significant pairwise differences marked with asterisks (* p <0.05). In (c), analysis on objective comprehension questions did not reveal any significant differences between the four conditions.

Rank tests with Bonferroni corrections revealed significant pairwise differences between the following: original / pop-up (Z = -2.036, r = 0.407, p = 0.042) and original / decoration (Z = -2.526, r = 0.505, p = 0.012).

While our comprehension questions in this study were intended to keep participants engaged with the text, we include participants’ scores for interested readers. The average scores for these questions across the four conditions are shown in Figure 10.4(c). One-Way repeated measures analysis of variance (ANOVA) did not reveal any significant differences between the conditions.

Next, participants were reminded of the 3 simplification conditions (with pictures of
Figure 10.5: Participants’ agreement to a Likert-scale question, presented at the end of the study, as to whether they would be likely to use each of the three simplification conditions, with asterisks marking significant pairwise differences (** p < 0.01, *** p < 0.01).

10.6 Discussion

RQ3.1: Perceived Benefit from Lexical Simplification

Users indicated stronger agreement with the statement “I was able to understand this text well,” for both the pop-up and decoration conditions, compared to the original condition (without any lexical simplification provided). This finding is in line with prior research on lexical-simplification tools for users with dyslexia, in which participants reported subjectively that simplification provided on-demand made texts easier to understand [99]. These subjective benefits perceived by our participants thereby motivate future computational lin-
guistic research on lexical simplification technologies for DHH adults, specifically methods for tailoring the output of automatic simplification tools for these users, which has been important for other groups [86].

10.6.1 RQ3.2: Providing Simplifications On-Demand

In our results above, the pop-up and decoration conditions received higher scores on the “likely to use” question, in comparison to the scores for the automatic condition. Thus, DHH users believed they would be more likely to use reading-assistance tools that provide simplifications on-demand, suggesting that users preferred greater autonomy in such tools. This finding aligns with speculation by authors in [99] that some of the benefits they observed from simplification tools for users with dyslexia were due to one of their prototypes providing assistance on-demand [99].

When asked about the reasons for their responses in our study, DHH participants mentioned a variety of reasons for why they preferred the on-demand forms of simplification (pop-up or decoration), including: a desire to have a choice, a desire to see the original text, and a desire to learn new words. When commenting on their preference for the pop-up condition, for instance, participant P13 commented:

“I may already know the word and may choose to skip the simpler word conversion but that choice wasn’t made for me already. I can choose whether to do [it] or not.” - P13

In terms of the desire to see the original text, participant P24, e.g., expressed concerns about missing important terminology for a class if the text is automatically replaced:

“I am not so sure about this method because if I am reading this article for a class, I would prefer to know all the terms that will be brought to class from this article.” - P24

In particular, the issue of learning new vocabulary was the most frequently mentioned reason by participants, e.g.:

“I like to know what words were originally used. I have no way of know what words were used when it was automatically simplified, therefore no way to learn.” - P21
Our results expand on the HCI literature on autonomy by illustrating that, in the context of lexical simplification, DHH users indicated being more likely to use designs that provide control in requesting simplifications on-demand and in seeing what portions of text have been transformed. This finding is in line with prior research that had found that increasing users’ sense of autonomy within a software environment may increase their engagement with it [43]. Our DHH participants indicated they would be more likely to use tools providing them with greater autonomy.

Our results also have implications for researchers in the field of ATS who want to test the quality of their simplification methods, by highlighting that whether those simplifications are provided on-demand may have an effect on the subjective judgements of their users.

10.6.2 Tradeoffs between On-Demand Prototypes

While DHH users preferred the two on-demand prototypes (pop-up or decoration), as compared to the other conditions (original or automatic), we did not observe any significant difference between their preferences for pop-up and decoration. In open-ended comments, DHH participants did mention some trade-offs between these two conditions, which relate to the different levels of autonomy provided by each. For instance, when discussing the pop-up condition, participants mentioned that the pop-ups could be distracting, and some participants disliked having to hover over a word to find out whether the system was capable of providing any simplifications for it. However, other participants indicated that they appreciated that it kept the original text and allowed them to see both the complex word and its simpler synonym at the same time. For instance, participant P9 commented:

“I like [the pop-up] more than the highlighting option, because it allows me to basically know an easier synonym of a difficult word, but allows me to reread the text at the level its at rather than a simplified version since I will know what the difficult word means.” - P9

As for the decoration condition, participants found it to be less distracting than the pop-up condition, and they mentioned that it was helpful to know which words they could
click on. Participant P16, for instance said:

“I actually like this method because it allowed me to learn more vocabularies and it was not distract at all like pop up. If there is a word that I am curious to know the meaning and I would click it to see it. Also, unlike pop up, you don’t have to move mouse around to find which word that has pop up because highlight is already there and it doesn’t bother me when I read the article.” - P4

However, some participants commented that clicking may take longer than hovering over words, and some were concerned about not being able to see both the original word and its simpler replacement word simultaneously.

10.7 Limitations and Future Work

There were several limitations in our study, which may limit the generalizability of this work or suggest future research:

- We have only investigated one type of text simplification (lexical), and it is important to clarify that we are not implying that lexical simplification alone is sufficient for addressing text readability challenges among DHH users. Instead, we chose to investigate lexical simplification in this study because we had perceived a gap in the literature, i.e. with prior computing research having previously investigating syntactic simplification among adults who are DHH [73]. In Chapter 12, we explore the preferences among these users for lexical, syntactic, and hybrid/combined simplification tools.

- While we have provided information about the level of difficulty of the texts shown in our study, which were rotated across conditions, these texts did not have identical Flesch-Kincaid scores. To further mitigate this lack of control over the texts’ difficulty, we shall use a Graeco-Latin square design to rotate conditions in future work.

- Our study did not include a baseline condition that displayed original text with decora-
tion around complex words, without any text simplification provided. Such a baseline could have revealed whether the observed effect on perceived readability might arise from text decoration alone, which could have led to response bias [121], rather than from actually providing lexical simplification.

- The responses gathered in our study may not be representative of a group of DHH individuals with a different distribution of ages, genders, identity (Deaf, deaf, or hard of hearing), etc. To enable replication of our work or for readers to better interpret our findings, we have provided these demographic characteristics, as well as WRAT scores (reflecting English reading literacy level) and IAF (reflecting the personality characteristics of the individuals in the study, in regard to autonomy). A further study with a larger group of participants or individuals with a different range of such characteristics may be needed to understand the range of opinions across an even more diverse set of DHH users.

- We used a Wizard-of-Oz method for identifying complex words and providing simplifications; there is a risk that the output provided may differ from an automatic system. Specifically, as highlighted in Chapter 8, an automatic system may sometimes provide erroneous word replacements (that change the meaning of the text), and a future study would be needed to investigate DHH participants’ evaluation of on-demand simplification systems with some errors in the simplifications provided.

- We used an initial study to select a subset of possible levels of user initiative and change visibility to include in our final study, but that first study was underpowered and used formatively in this work to avoid our arbitrarily selecting prototypes for comparison. In future work, a full factorial experiment with a larger sample size may enable us to determine the best combinations of both user initiative and change visibility. A future study could also focus on how the trade-offs of varying the level of user autonomy may make certain variations more appropriate for specific conditions (e.g. different kinds of webpages or tasks). Finally, future work should also ensure
that the color coding for complex and simplified text used for prototypes deployed in
the wild is accessible for users with color blindness.

10.8 Conclusion

After conducting a preliminary study to inform our design of a set of prototypes, we con-
ducted an experimental study with DHH participants who read four articles using various
prototypes, responded to subjective questions about each prototype, and shared feedback.
Participants reported that they perceived a benefit from lexical simplification tools, and
they preferred prototype designs that provided them with greater autonomy. Participants
also discussed trade-offs between their highest-rated prototypes in this study.
Chapter 11

Exploring The Design Space of ATS-based Reading Assistance Tools with DHH Adults

Most research studies on the use of ATS to provide reading assistance to different user groups described in previous chapters have made design choices in regards to various design parameters of these tools. These parameters relate to what and how simplifications are requested, as well as how they are presented. However, with the exception of some small informal studies (e.g. [14]) or a study that incorporated different designs for displaying the output of different systems [99], little work has examined the preferences of the target user groups for those design parameters.

Recent work in accessibility has advocated for the creation of taxonomies of the design space of technologies, highlighting benefits such as supporting participation of users in the design of the technologies, as well as systematizing the choices for designers and researchers working on those technologies, which can also help in identifying gaps in the literature and potential new opportunities for solutions [88]. Explorations of design spaces have been conducted for several decades, including work on input devices [23], and gaze interactions in
head-mounted displays [52]. Researchers have thus argued for the value of exploring design spaces when there is a need to structure existing solutions [88], but this also helps identify potential solutions and the preferences of users for particular design choices.

Thus, considering the lack of reported explorations of the design space in the context of ATS-based reading assistance tools, this chapter begins by outlining the various categories of parameters, and their options, that have been employed in prior work on the use of ATS to provide reading assistance. Then, this chapter outlines our work focused on the exploration of the design parameters through an exploration of this design space with DHH adults, in which we examine their preferences for the various design parameters, and the rationale for their preferred choices. Our primary goal with this exploration is to identify good settings for the various design parameters in order to construct a prototype of sufficient usability for use in an experimental studies evaluating how people interact with and benefit from such prototype, which is described in Chapter 12.

11.1 Research Questions

With this context, and with the goal of identifying which design combinations to include in a final prototype for the evaluative work of Chapter 12, we explore the following research questions:

RQ3.3. What are the preferences of DHH users for the various design parameters of ATS-based reading assistance tools? This overarching research question can then be broken down into two sub-questions:

RQ3.3.1. In a study with DHH users, what are their preferences for the various design parameters for each of the choices for the type of simplification? Do DHH users’ preferences among design parameters vary depending upon the type of simplification that is provided? Considering that prior work has identified benefits from both lexical and syntactic simplification for DHH adults, it is useful to determine how the preferences for
the design parameters change when employing those types of simplification, or combining both.

RQ3.3.2. What factors do DHH users prioritize when considering their preferences for the design parameters of ATS-based reading assistance tools? In open ended comments about their experience, it is useful to understand the values, goals or trade-offs that participants consider as they decide their preferences for various parameter settings.

11.2 The Design Space

An analysis of systems employed in prior work on the use of ATS as reading assistance tools revealed several parameters involved in the design of these systems. Most of these parameters can be fit into three categories by breaking down the process users go through when engaging with these systems.

First, users have to make the initial decision for what to request, which is the first category and includes the parameter related to the type of simplification requested.

- The type of simplification is the parameter that has been more clearly defined in prior work, given that, as outlined in Chapter 2, this relates to how the simplification itself is actually done. In other words, this relates to the type of linguistic transformation being performed on the text itself during the simplification process. As summarized in Chapter 2, there are typically three types of simplification: lexical, syntactic, or the combination of both (henceforth referred to as lexical + syntactic).

Then, the user needs to request the simplification. Thus, a second category of parameters is how the simplifications are requested, which includes the categories of user initiative and scope.
• As summarized in chapter 10, the user initiative parameter relates to whether users are allowed the initiative to request a simplification, or decide what text to request a simplification for. In Chapter 10, we had identified that allowing the user the autonomy to request simplifications is important in the design of these technologies. Thus, in the study presented in this chapter, we only explore the possible settings of this parameter that provide autonomy. These settings include giving the user full initiative of requesting simplifications for the text they decide, without any suggestions (e.g. [34]), or adding decorations to complex text as a suggestion to users for which text to request simplifications for (e.g. [14, 66, 101]). However, to the best of our knowledge, the latter (providing suggestions) has only been explored in the context of word-level simplifications.

• The scope refers to how much text the user requests to simplify at once. While it is related to the type of simplification requested, it is separate from that parameter given that a user may request lexical simplifications (i.e., replacing complex words) for a sentence at a time, for instance, or syntactic simplifications for all sentences in a paragraph. Varying the scope of how much text a user requests at once has not been explored in prior work as most systems presented focusing on lexical simplification only allow requests for one word at a time (e.g. [14, 101]), while the systems providing syntactic simplification typically provide those at the document level as opposed to one sentence at a time (e.g. [17, 105]). However, this was a parameter suggested by participants in our study presented in Chapter 5 as a way to interact quicker with a prototype, which may potentially be helpful for social accessibility issues.

Finally, the system has to present the simplifications to the user. Thus, there is a third category of how the simplifications are presented, which includes the parameters of placement, replacement indicator and permanence of the simplifications.

• The placement refers to how the simplifications are provided to the users and where they are placed in relation to the original text. Various options for this parameter
have been summarized in Chapter 10, which include replacing the text in place (e.g. [14, 27, 66]), as a “tool-tip” or pop-up that appears above or below the text (e.g. [86, 101]), as well as placing it elsewhere in the page, such as on a sidebar (e.g. [105, 114]). Finally, it is important to note, that to the best of our knowledge, the use of pop-ups have not been explored when combined with settings beyond the word-level for neither scope or type of simplification.

• The replacement indicator refers to whether there is a visual indication that a portion of the text has been replaced. Prior work has employed systems with (e.g. [14, 27, 66]) or without visual indications (e.g. [17, 34]), and in Chapter 10, we had identified a preference for using highlights for these visual indications for lexical simplifications. However, to the best of our knowledge, this parameter has not been explored either beyond the scope of word-level simplifications.

• The permanence of the simplification relates to how long the replacement remains visible to users, and the systems employed in prior work have also varied in how they approach this. Some prior work has explored the use of temporary replacements, such as while the user hovers over the text (e.g. [101]). Other systems have employed designs that allow the user to see the replacement until they dismiss it (e.g. [14]), while others provide the replacements permanently given that the user never sees the original text (e.g. [73, 101]).

One more parameter that emerged from our analysis of prior work is the naming of the tool, or how it is characterized. Unlike the parameters listed above, this is not a visual aspect of the interface of an ATS-based reading assistance tools nor fits into the process of users requesting simplification. As we saw in Chapter 5, the way these systems are named or characterized may be important for the adoption and social acceptability of these tools. Prior work’s labeling of these technologies include names such as “text simplification” (e.g. [14, 105]), “text adaptation” (e.g. [21, 137]), “facilita” which is Portuguese for “facilitate” (e.g. [129]), or “simplifying jargon” (e.g. [114]). The suggestions we had identified in
Figure 11.1: Screenshots of the prototype browser extension. Image (a) shows the prototype when lexical simplification is selected. Image (b) shows the option for “words” in the options for Quantity greyed out as it is not compatible with both and syntactic simplification.

Chapter 5, however, mostly focused on easiness, speed or productivity.

11.2.1 Prototype

To support our investigation of the preferences for these design parameters, we built a prototype of a browser extension that supports the configuration of all the various design parameters identified above, as shown in Figure 11.1. The order of the parameters in the prototype is based on the process outlined in the previous subsection, beginning with the parameters related to what is being requested, followed by how the simplifications are requested, and finally, how they are presented. Figure 11.2 shows examples of 3 combinations
Figure 11.2: Screenshots of portions of a ScienceDaily article showing 3 combinations of the various parameter settings. Figure (a) shows lexical simplification being employed at a word-level, using all highlights and providing replacements permanently in place. Figure (b) shows lexical + syntactic simplifications provided as a temporary pop-up for sentences, without any highlights. Figure (c) shows syntactic simplification alone, applied at a sentence-level, provided manually on the side with highlights only for suggested sentences.

The prototype begins with the simplification type parameter, with the options lexical, lexical + syntactic (labeled as “both” in the prototype) or syntactic simplification. Then, the scope parameter is labeled as “quantity,” with the options words, sentences, paragraphs or documents. As illustrated in Figures 11.2a and 11.2b, as users adjust this setting, hovering over the selected quantity would reduce the opacity of that respective portion of the text (e.g. if the option for sentences is selected, hovering over a sentence would reduce the opacity of that sentence). This change in opacity would indicate to the user what portion of the text would be modified when using the various settingDs. To conclude the parameters related to how the simplifications are requested, a toggle labeled “Highlight Complex” is provided to activate the highlights for suggestions on demand. When this toggle is selected, the appropriate complex text is highlighted by adding a gray background, as illustrated in Figure 11.2. This color was selected based on our pilot study in Chapter 10.

Then, the options for the placement parameter are grouped under the label “Location,” and include the options in place, pop-up and side. Pop-ups are presented as a small tool-tip
above the requested text, which are illustrated in Figure 11.2b. As we designed the option for pop-up, we decided to place pop-ups for sentences above that sentence’s paragraph to prevent obstruction, as shown in Figure 11.2b. Thus, requesting another pop-up replaces any current pop-ups, which in effect means that only one pop-up would be visible at a time unless the option for permanent replacements was selected. For consistency, we preserved this behavior regardless of the option selected for the scope parameter. The side option presents the replacements inside of a card on the right side of the screen, which is illustrated in Figure 11.2c. As the user hovers over the original text, the respective card changes in opacity to indicate which card belongs to which text, and vice-versa. When the settings allow for the user to revert the change, an x is shown at the top of the card to dismiss it.

Another toggle is provided with the label “Highlight Replaced” to activate the replacement indications. When the highlight replaced option is selected, text that has been replaced is highlighted with a yellow background, which was also selected from our pilot study presented in Chapter 10 and is illustrated in Figure 11.2a.

Finally, the options for the simplifications’ permanence are grouped under the label “Duration” and included temporary, manual and permanent. The temporary option shows the replacement while the user hovers over the desired text, and reverts to the original text as the user moves the mouse away from the replacement. Thus, it remains only temporarily as the user hovers over the text. The manual option provides replacements when the user clicks on the desired text, and allows the user to revert the change by clicking again on the replaced text. Lastly, the permanent option also allows the user to request the replacement by clicking on the desired text, but there is no option to revert the change.

Given that this prototype is intended to allow users to explore different settings, changing the options for the settings labeled simplification type, quantity, location or duration would revert all existing replacements, reverting to the original version so that only one option for each of those settings could be used at a time. The settings for highlights, however, would be preserved and toggling either highlights would not revert any replacements.

A careful reader may notice that ATS approaches that involve syntactic simplifications
(which would include the settings labeled as “syntactic” and “both” for \textit{Type of Simplification}) are only applied for scopes longer than a word. Thus, we still include the option for replacing individual words in the setting labeled as \textit{quantity}, as this is the approach most commonly used setting for the \textit{scope} parameter when providing lexical simplifications. However, given that approaches that involve syntactic simplification do not apply to individual words, this option is not compatible for simplifications that include syntactic rewritings. Thus, when the options labeled as “syntactic” or “both” are selected for the \textit{Type of Simplification}, the option for “words” in the quantity setting becomes unavailable, as illustrated in Figure 11.1b.

11.2.2 Implementation

The prototype was implemented as a browser extension for Google Chrome, using the developer guidelines for those extensions\textsuperscript{1} under their Manifest V2\textsuperscript{2}. As such, the prototype was implemented using the basic coding stack for web technologies: HTML, CSS and JavaScript.

The extension identifies sentences in a page, and sends these through an API call to a local server, which returns the simplifications at a sentence level to the extension. For this experiment, this API was run locally using Django\textsuperscript{3}, which is a Python-based web framework for building back-end servers. The server returns sentences based on the aligned sentences described in section 11.3.1, which were stored as JSON objects aligning each original sentence to the possible simplifications for each type of simplification. The extension would then mark up complex words and sentences (i.e. words and sentences that had a replacement) which would add the highlighting and replacement functionality to those words and sentences.

\textsuperscript{1}https://developer.chrome.com/docs/extensions/mv3/
\textsuperscript{2}It is important to note that at the time of this writing, Google is deprecating support for the Manifest V2 to favor the recently-released Manifest V3.
\textsuperscript{3}https://www.djangoproject.com/
11.3 Method

Our exploration of the design space was conducted using a structured usability-testing methodology that guided participants through explorations of various parameter settings. Given the large number of combinations possible with all the possible settings of the various parameters, it was not feasible to have participants go through every single possible combination of those settings. Considering that a key goal of this dissertation is to inform researchers working in the underlying NLP technologies that can provide the various types of simplification (e.g. lexical or syntactic), we privileged this particular parameter for exposure to participants to explore how their preferences for the rest of the design parameters varies when employing each type of simplification. Thus, we asked participants to explore the rest of the design parameters when using each option for the types of simplification. In other words, participants would read a text with each simplification type (i.e. lexical, lexical + syntactic, and syntactic), and explore their preferences for the other parameters how simplifications are requested and presented. Following an analogy suggested in [88], this may feel similar to an optometry test where the user adjusts the settings until an optimal setting is identified.

Immediately following the exploration stage for each type of simplification, and after identifying their preferred settings, participants then engaged in a task-based reading activity where the task consisted of answering comprehension questions about the text after reading each text. This task-based activity was done using the identified preferred settings to ensure participants focused on a reading task (as opposed to exploring the settings) using the preferred settings identified for each simplification type. At the end of these task-based activities, participants provided subjective judgements of the texts and the tool, and the research assistant asked participant questions about their rationale for their choices for the various settings, and how useful they found the support obtained using those settings during the task-based activity.

After having explored the settings and engaged in a reading task with all of the sim-
plification types, the research assistant then asked participants’ questions about whether there were other aspects of the process they would like to adjust, as well as how they would name a system that provides each type of simplification with their preferred settings, and why they would name it that way.

11.3.1 Materials

We selected six articles from the science news website ScienceDaily \(^4\), which we had previously used as a source for stimuli in Chapter 10. To select these, we first started with a selection of 15 texts that had Flesch-Kincaid grade levels between 13 and 15, and were between 550 and 750 words. Then, a team of two DHH research assistants judged each of the 15 texts in terms of whether they required special knowledge to be able to understand the content. Then, we discarded 5 articles that both research assistants indicated would need special knowledge to understand, leaving us with 10 candidate articles to use as stimuli. In this study, we used the first six articles in alphabetical order in the list. Links to the articles are included in Appendix F.

In their original versions, the six articles included in this study had an average Flesch-Kincaid grade level of 13.9, with an average word-length of 632 words and contained 27 sentences on average. It is important to note that ScienceDaily provides a summary of each article at the top of the page. These summaries were removed for this study to prevent participants from just reading the summary, and thus they are not included in these text statistics.

Simplifications

To obtain simplifications for our study, we processed the articles through a state-of-the-art lexical simplification pipeline presented in [80], which involves the typical lexical simplification stages of identifying complex words, as well as generating, selecting and raking

\(^4\)https://www.sciencedaily.com/
replacement candidates. We used a high precision threshold for the pipeline, which provides less substitutions overall, but leads to fewer errors. The syntactic simplification for this study was performed using a state-of-the-art pre-trained language model [98], with a sentence length threshold of 10 words. For the lexical + syntactic simplifications, the texts were first processed syntactically, and then lexical simplification was applied.

After applying lexical simplification, the average Flesch-Kincaid grade levels of the texts was reduced to 13.18 while the average word and sentence counts remained the same. The syntactic simplification reduced their average Flesch-Kincaid grade levels to 9.67, while the average word-length became 647 and the average sentence count, 48. For the lexical + syntactic simplifications, the word and sentence counts remained the same as the syntactically simplified versions, but the average Flesch-Kincaid grade levels were further reduced to 9.05.

Alignment of Simplifications

While the interface and front-end for our prototype was fully-functional, the back-end consisted of the original texts paired with predetermined simplifications obtained using the methods specified above. More specifically, our process consisted of first processing our articles through ATS systems that provide lexical or syntactic simplifications (or both). Then, using a Python script, these articles were processed on a sentence-by-sentence basis. According to the type of simplification obtained, the words or the sentences that could be replaced were aligned to the original words and sentences of the original article.

This alignment allowed us to provide users with the corresponding simplifications for the words or sentences they requested (by querying the aligned simplified words or sentences). When the parameter configurations supported switching back to the original (i.e. when simplifications were provided temporarily or until the user dismisses them), our structure also allowed us to swap simplifications back to their original version on demand.
11.3.2 Procedure

Participants met remotely over Zoom with a research assistant who is fluent in American Sign Language (ASL), after signing a consent form. The data collection website, which was built using the jsPsch library [31], and the prototype were running on the research assistant’s computer. Considering that a browser extension would need to be available on Chrome’s web store for participants to download, and could come into conflict with existing extensions on participants’ browsers, we opted for this approach to keep the extension local to the research assistant’s computer to ensure that the participants’ experience with the extension did not conflict with other extensions. Thus, participants interacted with the prototype using the remote control function on Zoom as the research assistant shared their screen.

The experiment began with a set of instructions introducing the experiment to participants, which introduced the goal of the study as the “evaluation of text simplification tools that make complicated text easier to read.” This initial screen also provided participants with an overview of the study. Then, participants were walked through instructions of where to find the extension, what the settings were and what each of the settings did. Below the description for each of the options for each setting, participants were also provided with a short video of the option in action, which were played side by side simultaneously. Participants also had the option to click on any individual video to enlarge it.

Then, participants read a total of 6 texts, which were described in detail above. For each simplification type, participants were guided through an exploration stage with one text, followed by a task-based reading activity with another text as was also described above. After reading both texts for each simplification type, participants discussed their experiences and preferences with the research assistant. The order of the texts and simplification types was counterbalanced using a Graeco-Latin square design to control for potential order effects. Finally, after having read all six texts, participants were asked questions about their overall experience.
At the end of the study, participants filled out a demographics form, and completed the sentence-comprehension subtest of the Wide Range Achievement Test 4th Edition (WRAT4) [132], which has been used throughout this dissertation as a measurement of their literacy skill. Participants were then compensated with US$40 for their participation.

### 11.4 Participants

Participants were recruited through social media, through Facebook groups and Reddit (on subreddits such as r/deaf or r/hard-of-hearing), and through advertising on the RIT campus. The recruitment criteria was self-identifying as Deaf or Hard-of-hearing and being at least 18 years of age.

We had a total of 15 participants in this study. Participants’ average age was 24.6 (range 18 to 36, SD = 5). A total of 8 participants identified as female, with 6 identifying as male and one preferring not to say. Participants average WRAT4 scores were 88.87 (range 68 to 108, SD = 12.76).

### 11.5 Data Analysis

We followed an approach similar to [88], where researchers ensured that participants indicated their optimal preferences, and then counted the number of participants who selected each choice for each design parameter. In their analysis, they identified that for some parameters, most participants made the same choice and thus those they concluded that particular choice may be preferred. For other parameters, however, their participants had provided diverse responses, which suggested that the preferences for those parameters may be more user-dependent. For those parameters that were more user-dependent, understanding the rationale for participants’ choices was more important.

While our method in this study does not support conducting difference testing between participants’ choices for each parameter, it was not our goal to identify significant differences
between participants’ choices nor conclusively determining the optimal choices. Instead, the goal of our analysis was to identify, from a formative perspective, whether there are certain settings that most participants would prefer for each of the parameters when using each type of simplification. These preferences would then be used in a subsequent study (presented in Chapter 12) in which participants would engage in reading tasks using the preferred settings identified in this study.

For the qualitative data obtained related to participants’ rationale for their preferences among the design parameters, whether there are other aspects they would like to adjust, and how they would name the system, we conducted thematic analysis following the approach outlined in [19]. The data was first interpreted and transcribed by the same research assistant who conducted the interviews. The transcriptions added up to a total of 11,775 words, for an average of 841 words per participant. Then, one researcher coded participants’ responses using an inductive open coding approach, followed by axial and selective coding to search for themes in their responses related to their rationale for their choices. Considering the results from previous chapters of this dissertation, we expected that the rationale may relate to issues of productivity (e.g. saving time) as presented in Chapter 4, social accessibility (as presented in Chapter 5) or autonomy (as presented in Chapter 10). However, considering that participants were interacting with a prototype that had not been employed in our prior studies, we still employed an inductive approach as new themes could emerge from these interactions.

11.6 Results

In this section, we first present participants’ preferred options for the various parameters, for each of the simplification types, which are also illustrated in Figure 11.3, as well as the choices that were not selected by any participant. Then, we discuss our findings from the open ended data, which complement participants’ choices for those parameters.
Figure 11.3: Bar charts showing the number of participants that selected each of the options for each of the settings in the prototype (i.e. quantity (a), highlight complex (b), location (c), highlight replaced (d) and duration (e) for each simplification type). The x axis in each chart shows the three simplification types (with lexical + syntactic labeled as “both”, and the bars are grouped accordingly using shades of grade to indicate the options for each parameter.

11.6.1 Parameter Preferences (RQ3.3.1)

Lexical Simplification

When using lexical simplification as the simplification type, most participants preferred words as their choice for the scope parameter (N = 7), with pop-ups being the most popular choice for their placement (N = 11). Most participants also preferred having highlights to suggest complex words (N = 14), as well as to signify words that had been replaced (N = 13). Lastly, most participants preferred to have the simplifications remain until the user dismisses them (N = 11).
Syntactic Simplification and Lexical + Syntactic Simplification)

The majority’s choices for lexical + syntactic and syntactic simplifications coincided for all parameters. For these types of simplification, participants preferred sentences as the scope to request replacements (N = 10 for lexical + syntactic, and N = 8 for syntactic), which for these types of simplification was the shortest unit possible given that words was not available. Participants preferred having the replacement of sentences sentences happen in place (N = 10 for lexical + syntactic, and N = 10 for syntactic). As for the highlights, participants also strongly preferred having highlights for suggesting complex sentences (N = 14 for lexical + syntactic, and N = 13 for syntactic) and indicating replaced sentences (N = 13 for lexical + syntactic and syntactic). Lastly, participants also preferred to have the simplification be displayed until the user dismisses them (N = 12 for lexical + syntactic and N = 8 for syntactic).

Undesirable Parameter Options

While the results above present the top preferred choices for each simplification type, those findings do not reflect that there were some options that no participant selected as their preferred option across simplification types. Those options included the option for requesting the full document as the scope of the request, and the option of placing the simplification off to the side of the original text.

11.6.2 Participants’ Rationale for Preferences (RQ 3.3.2)

Through our qualitative analysis of participants’ open ended comments, we identified a total of six themes. The first theme relates to participants’ goals and values when using the tool, while the subsequent four themes explore how these values apply as rationale to support their preferences for the various design parameters. The sixth theme explores participants suggestions for additional adjustments or features. And finally, while this was a direct response to our question and not a theme identified through the analysis, we discuss
participants’ suggestions for naming the tool.

**Participants’ value their reading fluency and efficiency, and having control over the experience**

Across participants’ discussions of their choices for the different UI settings, there were a few common values that seemed to justify those choices. First, participants valued maintaining their reading flow as natural as possible. Thus, UI choices that obstructed existing text, distracted by adding additional “clutter” or altered their reading fluency were often seen as undesirable, and vice versa. For instance, when describing the use of in-place replacements, P15 said “I feel the tool is efficient when it substitutes the text for easier reading as it provides less clutter”.

Second, participants also valued their reading efficiency, which was often referred to as the time spent reading. For instance, when discussing the benefits of the tool, P7 said “It did help me understand some of the sentences and saved me a few seconds of thinking about the meaning”. Thus, UI choices that were quick and easy to use were seen as preferable, whereas any choices that would add unnecessary time were seen as undesirable. P1, when discussing the benefits of requesting words for lexical simplification, described it as “smoother and faster.”

And finally, participants wanted control over the experiences. Thus, choices that supported them in having control over when changes happened or undoing changes were also preferable. In the words of P3, when discussing the duration setting, “I would keep the [manual] duration because it gives me a little bit more freedom to control when I want to see something simplified and when I don’t.” Control was also related to the idea of being able to compare the original text. For instance, P14 discussed liking temporary replacements because “I like to compare what’s the difference between translation and article, plus I want it to go back to normal after I move the mouse away.” The control over being able to compare the texts also allowed participants to not only learn new ways of saying things,
but also served as a way to check the quality of the replacement with a few participants expressing concerns about potential grammatical and semantic errors, such as P2, who said “I’m judging the grammar a bit.”

**Shorter units of replacement may be more efficient**

Participants displayed a tendency to prefer requesting the shortest unit possible for each simplification type (i.e. requesting words for lexical simplification, or sentences for the types that involved syntactic rewritings). This may seem paradoxical when considering that participants highly value efficiency and saving time, but requesting replacements for shorter scopes may require an extra amount of time to request the simplifications by clicking on more things. This concern was still expressed by some participants, such as P8 who said “I don’t like clicking on each individual word.”

However, when combined with the desire to maintain their reading fluency and the ability to compare the original to the replacement, participants’ desire for efficiency was also the reason most participants preferred the shortest unit as the scope of their requests. This is because participants reported wanting to reread the text once it has been replaced, and trying to identify what changed. P1, for example, when discussing the use of syntactic simplification for paragraphs commented “It took me longer because I was reading the paragraph, then clicking to translate, and then rereading the whole thing again.” Thus, rereading longer units of text may actually take longer than requesting shorter units of text that are only requested when needed.

**Pop-ups and in-place replacements provide a better experience in different contexts**

Most participants concurred that providing simplifications off to the side of the original text was not the best option. Participants complained that it felt out of place, was hard to read, and took not only additional space, but also time to look to the side and back to
the original text. For example, P3 commented “I didn’t like it on the side, I felt confused with it being out of place, so it didn’t help me read”. Thus, there was a preference towards pop-ups and in-place replacements, but participants also discussed trade-offs among these two options.

Pop-ups seemed good for comparing the original text and its replacement, which could support quicker rereading. In the words of P11: “I like them separate but close by so I can reread exactly what I need.” One participant also found them beneficial to actually read the replacements because otherwise it could be easy to skim over the complex words. However, when using pop-ups for longer units of text, then they can become obstructive as they will block the text behind them. P2, for example, complained that “the pop-up was blocking the text and interrupting my flow of reading.” Thus, participants expressed an inclination towards in-place replacements when working with longer units of text. In that context, in-place replacements reduced the potential distraction that pop-ups could provide, while still maintaining a coherent reading experience. P8, for instance, commented “With [in-place], I feel like if I click on it, I can focus on that. But if it’s in two different places, it’ll be a bit confusing for me. So, that’s why I like [in-place].”

Control is a key factor for duration

Most participants agreed in their preference for the manual option for duration, with most citing control over being able to revert the change as the main reason. When discussing their choice for manual duration, P3 said “I liked that I could choose what I wanted to look at and what I didn’t.” While the temporary option also allowed a certain level of control, some participants disliked that it interfered with any mouse movement, which they might use to help them focus as they read to a text and thus may lead to replacing or undoing simplifications involuntarily. This was the case for P2, who explained that “sometimes I like to move the mouse along the words so I can read along, which helps me focus, but [the temporary] gets in the way so I moved it to manual which makes things easier.” Finally, while some participants thought that permanent option was efficient, the
potential for obstruction when using permanent pop-ups as well as the lack of an ability to revert a change lead to a more prominent preference towards the manual option. P13 put it as “I don't like the idea of clicking something and not being able to change it back.”

Highlights may help with focus, but also have the potential to distract

At a surface level, participants expressed a desire to adjust the highlight colors for both complex text and their replacements, as participants’ preferences for colors varied. However, there was more nuance to participants’ discussions of highlights, as they seemed to have high potential to support the reading experience, but also to hinder it.

On one end, some participants confused the highlighting with highlighting important words, as that is how they regularly use highlights. For instance, P3 commented that “[highlights] help me figure out where to focus my attention” and similarly, P6 commented “I like the idea of using the highlights there so I know which part is important to read.” Other participants explicitly saw having highlights for complex words as a benefit so that those are not skimmed over and thus, the highlights helps them focus. P11, for instance, described it as “I feel like lexical helps me catch important and complicated words. Sometimes I would miss some of them by reading too fast so when they're highlighted like that, they’re being emphasized to my eyes so I feel like I remember the words better.”

Several participants found that highlighting longer units of text (e.g. sentences or paragraphs) could be distracting or overwhelming, could make it hard to identify what had changed after the complex text is replaced, or could block existing tools. P4 indicated being ”a little overwhelmed with the amount of highlights,” while P9 complained that they couldn’t use the Apple dictionary feature\(^5\). Some participants also expressed confusion over the meaning of the highlights when using them for the simplification types that involved syntactic rewritings (which would, in effect, highlight units of text longer than a word). For

\(^5\)This may have been because of the force touch feature in MacBook trackpads often used to query Apple’s dictionary, as it could trigger a simplification since it registers a click. However, the dictionary works in our testing, but we wanted to stay true to participants’ experience in our reporting.
instance, P5 thought that this was still signifying words that could be replaced and thus it was hard to identify which words would actually be replaced. P6 also found the highlights in this context to be overwhelming because of the number of “complex words” and P11 was concerned that the complex words could be overlooked if all words are highlighted.

Additional requests and adjustments

Many participants complained that the slider design for choosing among the options in the prototype extension were too small to click on and thus should be made larger. P6, for instance, said “one thing I’m struggling with is selecting and moving the options, they’re very small.” P11 also suggested that the “document” choice for scope should be renamed to “page” in the context of a browser extension as this will be used on websites and it might be confusing to refer to a document. Lastly, P11 also highlighted the importance of providing a “dark mode” version of the extension UI (i.e. light text and UI elements on a dark background) so that a person’s reading experience is not affected by the bright white background if reading in a dark environment or using a dark mode setting on their device.

Participants also suggested other ideas to support the reading experience. For instance, P15 suggested that the tool could provide a “word bank” feature where all of the words a user has requested are stored in the extension, while P5 suggested that having a thesaurus would also be helpful.

Participants’ Naming Suggestions

Across participants’ naming suggestions, there were four categories that we identified. One category of names just involved the emphasis on the word ”simple,” which P3 thought would be the most important thing (“I think the most important thing is to include ‘simple’.”)

These names included “simplification” (P1); “simple sentences,” for which it is important to note the alliteration this creates in ASL as the signs for the words “simple” and “sentences” are similar; and “simple text” (P7).
The second category focused on the fact that the text is being modified, and thus involved the word “modification,” with P9 suggesting the name “modification tool” because “you change how it looks, to make it simple or whatever you want.” P2 also suggested the names “TextMod” or “ReadingMod” because “Mod means modification so combining that with any textual phrase would be helpful.”

The third category focused on reading, with suggestions including “reading helper” (P13), “reading assistant” (P14), or “readify” (P5) because “it makes sense to combine ‘read’ with ‘ify’ to make it trendy.”

Lastly, the fourth category involved what participants saw as the purpose of the tool, with the names including simplifying for the layman (“layman’s words”, P7), studying (“study tool”, P12), learning vocabulary (“vocabulary wordlist”, P15) or just making it simple to read (“that tool that helps make it simple to read, look look!” P7).

11.7 Discussion

The values we identified as important for participants’ preferences in this work, which included saving time or being efficient, as well as maintaining reading fluency and control over their interactions, have some overlap with themes that had been discussed in previous work presented in this dissertation, but incorporate additional insights.

For instance, participants’ comments on saving time related to the themes identified in Chapter 4, where participants seemed to quantify text complexity in terms of the time it takes to understand a text. This topic of time had also briefly surfaced in Chapter 10, where participants discussed the trade-offs between pop-ups and in-place replacements in the context of using lexical simplification, where participants viewed the fact that the interaction for pop-ups could take longer as a downside of that design choice. However, in the work presented in this chapter, we found a third time-related dimension that had not been explored in prior work: the time it might take to reread the text after it has been simplified as it was important for participants to compare the original and simplified texts.
for the various reasons discussed in the results. Prior work had also identified rereading as a reading comprehension strategy that DHH readers might use more frequently than hearing readers, especially DHH readers with lower reading skill [8, 10]. Our results suggest that this strategy may still be employed in the context of having texts being simplified and becomes an important consideration in the design of these tools. Thus, while in Chapter 5, participants had suggested the replacement of longer amounts of text at once as a way to potentially mitigate social accessibility issues, participants in this study leaned towards replacing shorter units of text because of the time it takes to reread the replacement.

Similarly, participants value for control over their interactions had also been the main focus of our work presented in Chapter 10, where participants preferred designs that provided them with greater autonomy. However, while some of the motivations for having greater autonomy discussed by participants in that study did include the ability to learn new vocabulary, our work in this chapter provided further insight into participants’ desire for control and for reducing distraction. Namely, participants wanted to be able to compare the original and simplified texts, but without the potential obstruction or interruption that choices that provided less control could introduce (e.g. temporary or permanent duration). These ideas of comparing without obstruction or interruption also related to participants’ value of maintaining their reading flow, which we identified as an important factor in participants’ decisions for their design preferences.

Considering that prior work exploring design parameters in which participants actually engaged with simplification tools had mostly focused on lexical simplifications, and that our prior work exploring participants’ attitudes towards other forms of simplifications had not involved actual interactions, this finding highlights the importance of having participants actually engage with tools as other important values, such as participants’ prioritization of being able to compare the texts while still being efficient and maintaining their reading flow, may be identified.
11.7.1 Visual Strategies and User Interface

Our work also sheds light into how the incorporation of user-interface elements in the context of reading assistance technologies should carefully consider readers’ current behaviors, especially their visual strategies. For instance, our findings suggest that participants leaned against temporary replacements while the user hovers over a portion of the text, potentially because it may interfere with the strategy of using the mouse to read along. This strategy, also known as pointer-assisted reading (PAR) has been observed to correlate with the practices of skimming and rereading [69, 70, 71]. However, to the best of our knowledge, no work has explored the prevalence of this strategy in the context of DHH readers.

We also observed a nuanced relationship with participants preferences towards highlights. Some participants seemed to confuse the purpose of the highlights as signifying important words (as opposed to words that could be replaced), as that is the purpose they most often used them for. This aligns with work focusing on highlighting important words in the context of captioning, where participants found the highlights to be useful [63], but there were limits to the amount of highlights that could be included before participants perceived them as distracting. The potential for distraction was also brought up as a concern by participants in our study. This may be more relevant for DHH readers, as research has observed that DHH readers may have a broader perceptual span as compared to hearing readers [38], and may read quicker [24] employing a higher degree of visual word processing [38]. This increased importance on the visuals of how text is presented, coupled with the habit of using highlights for important text, may explain participants’ concern for visual decorations being distracting when employed to indicate confidence in automatic captions [11]. While in our findings participants did express some concerns for the highlights’ potential to distract, however, the conflicting result was that nearly all of the participants still selected having highlights for both suggestions and replacement indications as their preferred options for the highlight settings. Thus, the use of highlights in this context merits additional investigation.
11.8 Limitations and Future Work

There were several limitations with this work, but it also opens several avenues for future work. A key limitation was our inability to explore participants’ preferences for every possible combination of the design parameters explored in this study. Thus, our study did not support identifying significant differences in their preferences or conclusively identifying the optimal settings for the various design parameters. Our goal in this chapter, however, was to identify possible preferred combinations of the parameters that relate to the user interface of the tool for each of the type of simplifications that NLP researchers typically provide (e.g. lexical, lexical + syntactic, and syntactic simplifications). Thus, now that we have established the design space of these tools, future work can more thoroughly explore the combination of all the possible design parameters.

As our results in this study identified, the visual strategies employed by DHH readers may also be important in participants’ preferences towards the various UI settings. Thus, future work could more explicitly investigate those strategies with DHH readers at various literacy levels, as well as other user groups who may benefit from automatic text simplification, to explore how the variations of the design parameters explored in this study may affect or support those visual reading strategies.

Prior work has also varied along one additional design parameter, which is the platforms that they use to allow the users to interact with ATS tools. These include the use of applications (e.g. [17, 129]), browser extensions (e.g. [14, 27, 101]), as well as website with text boxes to submit the text itself or URLs containing the text (e.g. [34, 86, 105]). However, given the number of parameters and how the number of combinations multiply with each parameter that is added, we did not consider the parameter of the platform. Instead, we opted for using a browser extension only in the study, for two reasons. The first reason was practical, given that a browser extension was easier to deploy for this study than other options such as an application, but it may also be used during more realistic reading tasks as opposed to copying and pasting text on a separate website. The second reason was...
that some of our findings may generalize to the other two options of platform explored in prior work; the use of ATS within applications (e.g. a Kindle application, or a news-reading application) will still require users to engage with the text within their natural reading environment, and thus the way they request simplifications or the way those simplifications are presented, while using a separate website may still require user input after processing the website. However, we acknowledge this generalizability of our results will still have to be investigated in future work.

In this study, we also were not able to investigate an additional design parameter which relates to the number of simplifications provided to the user. The systems that we used to process our simplifications do not have the capability yet to provide various alternatives to complex text. Thus, future work should explore how providing more than one alternative may affect participants’ preferences towards the various design parameters explored in this work.

As texts in different domains may present specific challenges to readers, their preferences may vary when reading texts from different domains. Thus, considering that we employed texts related to science news, future work can explore the preferences of DHH adults for these design parameters when reading texts in other domains. It is also our hope that our identification of the various parameters of the design space of ATS-based reading assistance tools may be used by other researchers working in the use of such tools for other user groups, to identify how their preferences for the various design parameters may differ from those of DHH readers.

Finally, a key avenue for future work, which is explored in Chapter 12 consists of conducting experimental evaluations of the designs identified in this chapter in order to explore how participants interact with the various types of simplifications when employing the combinations identified in this chapter, as well as how participants benefit from using a tool that provides those.
11.9 Conclusion

In this study, we conducted an exploration of the design space for an ATS-based reading assistance tool with DHH readers. The main goal of our study was to identify parameter settings to employ in our evaluation of a tool in Chapter 12. As such, based on the results from this study, we select the following settings: for lexical simplification, we will employ word-level simplifications provided as a pop-up, while for lexical + syntactic and syntactic simplifications, we employ sentence-level simplifications provided in place. For all types of simplifications, the replacements will be provided manually until the user dismisses them, using highlights for both suggestions and replacement indications to allow further explore the conflicts observed in this study in participants’ preferences towards highlighting. In addition to these parameter settings, we identified participants’ rationale for their choices, which provided additional insights to our findings from prior chapters on the themes of saving time, maintaining reading fluency and control over their reading experiences.
Chapter 12

Evaluation of an ATS-Based Reading Assistance Tool with Deaf and Hard-of-hearing Readers

After having narrowed down the design candidates for each type of simplification to include in a final prototype in Chapter 11, in this final chapter we aim to conduct a prototype evaluation focused on identifying the benefits of providing a fully-functional prototype of an ATS-based reading assistance tool as well as identifying how users interact with the prototype during task-based reading activities.

12.1 Research Questions

Using our fully-functional prototype developed based on the findings from Chapter 11 of this work, we investigate the following overarching research question:

RQ3.4. When providing an ATS system that provides both lexical and syntactic simplification features to DHH users, how may they benefit from and interact with such system? This research question can then be broken down into 3 research
RQ3.4.1. Do participants perceive benefits from providing DHH adults with an ATS system that can provide lexical or syntactic simplifications? This question focuses on whether an ATS system provides measurable benefits to DHH users, as measured by using metrics identified in our methodological work (e.g. Chapter 7), or observed through participants’ open ended comments.

RQ3.4.2. What are participants’ overall impressions of the usefulness of the tool? Upon interacting with a system that provides lexical or syntactic simplifications, do participants display a preference for one, the other, or both combined? This question would focus on participants’ subjective judgements of whether participants find the tool useful after engaging with it, as well as their preferences towards particular types of simplifications.

RQ3.4.3. How do participants interact with the different features of an ATS system that can provide both lexical and syntactic simplifications? This question, in turn, focuses on the behaviors of DHH users while engaging with a system that provides lexical and syntactic simplification.

12.2 Prototype

The implementation of our prototype for this study was based on the prototype used in Chapter 11, and employed our findings from that study. Thus, it remained as an extension for the Google Chrome web browser that uses traditional web-based technologies to interact with the text on a web page, and API calls to a local server to request simplifications from our aligned database of pre-processed simplifications obtained using state-of-the-art ATS systems.
In this study, however, the number of settings was reduced to a single setting for changing the simplification type between the options for lexical simplification, which is labeled as “Lexical;” lexical + syntactic simplification combined, which is labeled as “Both;” and syntactic simplification alone, which is labeled as “Syntactic.” Considering participants’ comments from our exploratory study, we increased the size of the selection boxes for this setting by using a grouped checkbox component from Bootstrap\(^1\), which is illustrated in Figure 12.1. Furthermore, considering that in this study participants would be asked to read some texts without using the tool, we added an additional setting to fully disable the extension, which consisted of a Bootstrap toggle component labeled as. As shown in Figure 12.1b, disabling the extension would gray out the settings for the simplification type.

Figure 12.2 illustrates how the visual elements of the tool look when using the 3 settings for the simplification type, which were based on our findings from Chapter 11. The *lexical* option provides simplifications as a pop-up, using highlights to indicate words that had a simplification available as well as words that were replaced (i.e. the highlight color changes while the pop-up is visible). Requesting a simplification for a different word closes any

\(^{1}\)Bootstrap (https://getbootstrap.com/) is a front-end development toolkit which provides ready-made UI components.
CHAPTER 12. EVALUATION OF AN ATS TOOL

The way people interact safely with robots is at the forefront of today’s research related to automation and manufacturing, explains Debasita Mukherjee, a doctoral student and lead author of a recently published study. She is one of several researchers at UBC’s Advanced Control and Intelligent Systems Laboratory who are working to develop systems that allow humans and robots to interact safely and efficiently.

**Lexical**

The way people interact safely with robots is at the forefront of today’s research related to automation and manufacturing, explains Debasita Mukherjee, a doctoral student and lead author of a recently published study. She is one of several scientists at UBC’s Advanced Control and Intelligent Systems Laboratory. They are working to develop systems that allow humans and robots to interact safely and effectively.

**Both**

The way people interact safely with robots is at the forefront of today’s research related to automation and manufacturing, explains Debasita Mukherjee, a doctoral student and lead author of a recently published study. She is one of several researchers at UBC’s Advanced Control and Intelligent Systems Laboratory. They are working to develop systems that allow humans and robots to interact safely and efficiently.

**Syntactic**

Figure 12.2: Screenshots of a paragraph from a ScienceDaily article when using the 3 simplification types for one word or sentence. Gray indicates suggestions, while yellow indicates replacements. The *lexical + syntactic* option is labeled as “Both”.

existing pop-ups, so only one pop-up can be seen at a time. The other two options (i.e. lexical + syntactic and syntactic) also use highlights to indicate sentences that can be or have been replaced, but the simplifications happen in-place (i.e. the original sentence is replaced with the simplified version in the text). Switching between the settings reverts the text to its original form. Disabling the extension, in turn, also reverts the text to its original form and removes any visual elements introduced by the tool.

12.3 Method

Our study employed a within-subjects design in which participants read through texts both with and without the fully-functional prototype of an ATS-based reading assistance tool we developed with the preferences identified in Chapter 11. We obtain reading metrics, participants’ judgements about the texts and the tool, as well as behavioral data about their interactions with the tool. We also obtain open ended comments from participants to better understand their experience with and preferences for the tool. The following subsections include details about the materials used, our procedure and participants.

12.3.1 Materials

Reading Stimuli

In this study, participants engaged in a total of 4 task-based reading activities. For these activities, we employed a total of 4 articles which were obtained from ScienceDaily using the same procedure as outlined in Chapter 11. The average word-length of these 4 texts in their original versions was 664 words, ranging from 574 to 732. The average number of sentences, in turn, was 26.75, ranging from 23 to 35. These original articles had an average Flesh-Kincaid grade level of 13.8, ranging from 13.2 to 14.2. Direct URLs for all of the original articles are included in Appendix F.

\[^2\text{https://www.sciencedaily.com/}\]
As participants also engaged in an exploratory task of familiarizing themselves with the tool, we employed a fifth article from ScienceDaily to use exclusively during this exploratory task, with a word length of 578, 27 sentences and a Flesch-Kincaid grade level of 14.2. Similar to Chapter 11, the summary at the top of these articles was removed for the experiment.

Simplifications

The simplifications for these articles were also obtained following the same procedure as Chapter 11, using the state-of-the-art systems presented for lexical [80] and syntactic [98] simplification, and combining both to obtain the lexical + syntactic simplifications.

The total number of lexical simplifications suggested by the system was, on average, 47.25 (ranging from 36 to 80). When all lexical simplifications are applied, the average Flesch-Kincaid grade level of the articles was reduced to 13.33. In turn, an average of 22 sentences had simplifications available after processing the articles using the syntactic and hybrid approaches (ranging from 16 to 30). When applying all of the syntactic simplifications available, the average Flesch-Kincaid grade level would be reduced to 9.35, while the average word count and number of sentences increased to 682.8 and 49.8, respectively. Lastly, when all of the lexical + syntactic simplifications were applied, the average Flesch-Kincaid grade level would be further reduced to 8.6, with the average word count increasing to 682.8, and the average number of sentences increasing to 48.7.

Metrics

To address our research questions, we obtained participants’ responses to various question items and captured various metrics. First, to keep participants on task during the task-based reading activities, we authored comprehension questions for all of the articles and provided these as a single quiz at the end of each reading activity. These questions consisted of 3 multiple-choice questions for each articles, that asked about main facts of the
reading. Following the recommendations from Chapter 7, they were written at a low level of linguistic complexity. The average Flesch-Kincaid grade level of the question items was 5.4.

We also captured participants’ average reading speeds by measuring the time participants spent reading the articles and normalizing it over the number of words in each respective article.

After reading each text, participants were also asked to provide subjective judgments of readability and understandability, which consisted of the Likert-scale indicating their agreement with the statements “I was able to understand this news story well.” and “This new story was easy to read”. Participants were also asked to rate how useful they found the tool overall by indicating their agreement to the item Likert-scale item “The tool was useful for reading the news story.”

Lastly, after completing all of the task-based reading activities, participants were asked to indicate how useful they found each simplification type. Using a Likert-scale, participants indicating their agreement to the statement “The [type] simplification type was useful for reading the news stories,” where [type] was each of the simplification types (i.e. “lexical,” “both” (for lexical + syntactic) and “syntactic”).

Open Ended Questions

After each task-based reading activity in which participants engaged with the tool, participants were asked about their overall experience with the tool, and, depending on whether they switched settings throughout their reading activity, their rationale for doing or not doing so.

At the end of all reading activities, participants were asked about their thought process for rating the usefulness of each of the simplification types, as well as their thoughts about each of the types. Participants were then asked about their impressions of the labels used for the simplification types, as well as the visual elements employed by the tool (e.g. the highlights and pop-ups). Lastly, participants were asked about what they might change
about the tool if they had control over the design.

### 12.3.2 Procedure

After completing a consent form, participants met remotely with a research assistant who is fluent in ASL over Zoom for this IRB-approved study. The data collection website was built using the jsPsych library [31] and was executed locally on the research assistant's computer, which also had our prototype extension installed. Thus, participants interacted with the data collection website and prototype using the remote control feature from Zoom as the research assistant shared their screen.

The data collection guided participants through the entire procedure, which began with details and instructions about the experiment. Participants were then introduced to the tool with a short screen recording that demonstrates how to activate the tool, followed by a detailed explanation of the 3 settings for the simplification type which explained the function and look of the settings, accompanied by visual illustrations of the functionality.

Participants were then invited to an exploration stage in which they read an article and were encouraged to use the tool freely to familiarize themselves with it until they had a good sense of how it worked. Then, participants engaged in 4 task-based reading activities. For 2 of these 4 tasks, participants were encouraged to use the tool, while for the other 2 they were asked not to use the tool. The order of the texts and conditions was rotated to counterbalance these across participants\(^3\). Before these tasks-based reading activities, participants were instructed that there would be a quiz at the end and an opportunity to provide feedback about the text and, when applicable, the tool. Thus, after reading each text, participants responded to the metrics and judgements outlined above. After

\(^{3}\)Upon careful review of the data obtained, we identified that our data-collection website did not counterbalance the combination of texts and conditions evenly, with 17 participants being assigned one combination and the other 10, the other. However, the rotation of the conditions (i.e. reading with or without the tool) did take place for all participants. To rule out the potential effect of the texts on our results, we plan to collect data from an additional 7 participants to achieve an even balance.
each activity in which participants were encouraged to engage with the tool, the research assistant then asked some open ended questions about their experience.

After reading all the texts, participants were asked to provide their judgements for the usefulness of each of the simplification types. Then, the researcher asked participants open ended questions about their experience, the trade-offs of using the different features of the prototype, as well as their thoughts on the value of using ATS-based reading assistance tools.

Finally, at the end of the study, participants completed the sentence comprehension subtest of the WRAT4, as well as a demographics form. Participants were compensated with US$40 for their participation.

12.3.3 Participants

We recruited participants through social media, including posts on Facebook and Reddit, and through advertising on the RIT campus. Our recruitment criteria included identifying as Deaf or Hard-of-hearing and being 18 years old or older. We also invited participants from our study exploring the design space of the tool described in Chapter 11.

We recruited a total of 27 participants, which included 5 who had participated in our previous study. Participants reported an average age of 30.4, ranging from 20 to 54 and with a standard deviation of 11.3. In total, 14 participants identified as female, 12 as male and 1 preferred not to say. In terms of participants’ reported hearing identity included 18 participants who identified as culturally Deaf [91], 5 as hard-of-hearing, 2 as deaf and 1 as Deaf/Blind (who was able to adjust the screen to their preference). Participants average WRAT4 scores was 78.8, ranging from 59 to 102 and with a standard deviation of 11.9.

12.4 Data Analysis

While our study employed a within-subjects design, each participant engaged in more than one reading task for each conditions. Thus, when conducting difference testing for the
metrics obtained we used we use between-subject tests of statistical difference. Shapiro-Wilk tests of normality indicated that the distribution of the responses and measurements for all of the metrics (i.e. comprehension questions, subjective judgements of readability and understandability, and reading pace) deviated significantly from a normal distribution. Thus, we employ the non-parametric Mann-Whitney $U$ test for these metrics.

For participants’ judgements of the overall usefulness of the tool, we simply provide descriptive statistics (e.g. what percentage of participants responded to each option). However, to evaluate whether there is a significant difference between participants’ judgements of the usefulness of each of the simplification types, participants only provided this judgement once at the end of the study. Thus, we employ the non-parametric Friedman test for within-subjects responses. If a significant difference was observed, we followed up using pairwise Wilcoxon Signed Rank tests using Bonferroni conditions to determine between which types the statistical difference was observed.

For those analyses, we excluded the data from 3 instances in which participants’ reading speed was observed to be higher than the median plus 3 times the inter-quartile range (IQR) which may suggest that a participant might have been skimming or skipped a step in the data-collection website.

The behavioral data logged by the prototype included participants’ requests for simplifications, which included which type of simplification the participants requested on which portion of the text. The prototype also logged when participants requested to undo a simplification, which also stored the text it was requested on. Lastly, the prototype logged when participants changed the setting for the simplification type, also storing which simplification type the participant changed to. We analyze the overall number and average percentages of occurrences of these interactions. When looking at the requests for simplifications, we also separate the number of simplifications requested for unique portions of a text versus repeated requests for the same portion of a text, as we had identified rereading as a potential behavior in Chapter 11.

The open ended data obtained at the end of the experiment was transcribed by a
research assistant who is fluent in ASL. The transcriptions added up to a total of 12,236 words, for an average of 453 words per participant. We then conducted a thematic analysis of participants’ responses, following the approach described in [19]. More specifically, a team of two researchers coded the data by independently performing a round of open coding, followed by a round of axial coding. Then, the researchers met to discuss potential semantic themes using an inductive approach, which were then refined to distinguish themes and subthemes.

12.5 Results

In this section we present the results of our evaluation metrics, as well as participants’ judgements of the usefulness of the tool and the various settings provided. Then, we describe the results obtained from the observational data of participants’ interactions with the tool. Lastly, we present the results from our analysis of participants’ open ended comments at the end of the experiment.

12.5.1 Reading Metrics and Participants’ Judgements

Participants’ average comprehension scores after reading the texts with the tool was 76%, and 71% when reading without the tool. However, a Mann-Whitney U test did not reveal any significant differences between the conditions. Similarly, Mann-Whitney U tests did not reveal any significant differences in participants’ subjective judgements of understandability or understandability between the conditions of using and not using the tool. As illustrated in Figure 12.3, participants’ average reading speed, in turn, was 205 wpm when using the tool, compared to 238 wpm when reading without the tool. A Mann-Whitney U test revealed a significant difference in participants’ reading speed between these conditions ($Z = -1.96$, $r = 0.27$, $p < 0.05$).

After each reading task using the tool, we had asked participants to indicate how useful they found the tool as a whole. Overall, participants indicated agreement to the Likert-scale
Figure 12.3: Participants reading speed under the conditions of using the tool and not using the tool (\(^*\) = p < 0.05)

item “The tool was useful for reading the news story” after a majority of the reading tasks (58.5%). These judgements are illustrated in the bottom bar of Figure 12.4, which uses divergent stacked bar charts as recommended to visualize these Liker-scale responses [104]. More specifically, participants selected “agree” 31% of the time, and “strongly agree” 27.5% of the time. Participants selected the neutral option “neither agree nor disagree” 29.5% of the time, and “disagree” 12% of the time, with zero instances of “strongly disagree.”

Finally, after participants had completed all 4 reading tasks, we asked them to indicate how useful they found each type of simplification. The top 3 bars in Figure 12.4 summarize participants’ usefulness judgements for each simplification type. A Friedman test revealed a significant difference between participants’ usefulness judgements for the 3 types of simplifications (\(\chi^2 = 16.987, p < 0.001\)). Post-hoc comparisons using pairwise Wilcoxon Signed Rank tests with Bonferroni corrections revealed significant differences between participants judgements for lexical simplification when compared to both combined (\(Z = -2.34, r = 0.45,\)
Figure 12.4: Divergent stacked bar charts [104] summarizing participants’ judgements of the usefulness of the tool. The top 3 bars summarize participants’ judgements for each simplification type, while the bottom bar summarizes participants’ judgements for the tool as a whole, which were provided separately. (** indicates $p < 0.01$, * indicates $p < 0.05$).

$p < 0.05$), and syntactic alone ($Z = -2.9$, $r = 0.56$, $p < 0.01$).

### 12.5.2 Participant Interactions

By looking at all of the interactions across all participant interactions during the task-based activities, we were able to observe participants’ average number of requests for simplifications when using each of the simplification types. We also looked at the amount of words and sentences, out of the ones that had simplifications available, received at least one simplification request. Then, we look at the number of participants who engaged in rereading (i.e. requesting a simplification and undoing it repeatedly), and the proportion of simplification or undo requests that this rereading behavior accounted for. Lastly, participants were provided with a user-interface control that allowed them to switch the tool between different settings, i.e., from lexical, to syntactic, to “lexical + syntactic” settings. Thus,
we also look at the number of times participants switched to each of those settings. The following subsections summarize our observations.

**Requests by Simplification Type**

When using the *lexical* simplification setting, we observed an average of 9.3 simplification requests per participant (SD = 9.3, range = 0 to 37). The average for undo requests, in turn, was 3 undo requests per participant (SD = 5.1, range = 0 to 20). Notably, there were 2 participants who did not make any simplification requests for *lexical* simplifications.

When a participant selects a text using the setting for lexical + syntactic simplification, then not only is the text simplified syntactically but lexical simplification also occurs. When using this setting, we observed an average of 10.1 simplification requests per participant (SD = 9.8, range = 0 to 37). We observed an average of 3.6 undo requests when using this setting (SD = 5.1, range = 0 to 15). And, we also note that 7 participants did not make any simplification requests when using this setting.

Finally, when using *syntactic* simplification, we observed an average of 5.7 simplification requests per participant (SD = 7.1, range = 0 to 22). These were undone, on average, 2.9 times per participant (SD = 3.9, range = 0 to 12). Notably, a total of 11 participants did not make any requests for syntactic simplifications.

**Words and Sentences that Received Requests**

Across all of the text articles shown in the study, there were 189 words that were indicated through highlighting that lexical simplifications were possible. Some of these words were never clicked on by any participant. There were 106 unique words that were clicked on by at least one participant to request lexical simplification.

When using the settings for *lexical + syntactic* simplification or *syntactic* simplification alone, there were 88 sentences that were highlighted to indicate that could be applied when using this setting. Out of these 88 sentences, 84 were clicked by at least one participant to
request *lexical and syntactic* simplification, whereas 62 sentences were clicked by at least one participant to request *syntactic* simplification alone.

**Rereading Behavior**

Considering that participants could make multiple requests for simplifications on the same portion of the text, we then looked at how many participants engaged in rereading (i.e. requesting a simplification, undoing it, requesting it again, and potentially repeating).

A total of 7 participants engaged in this rereading behavior at least once when using *lexical* simplification, with a total number of 13 instances of rereading on 10 unique words. These accounted for 5.2% of the overall number of simplification requests we observed for *lexical* simplification, and 15.8% of the occurrences of participants directly undoing *lexical* simplifications. This rereading behavior was observed on 9.4% of the unique words requested.

When using the setting for *lexical and syntactic* simplification, in turn, a total of 10 participants engaged in rereading at least once generating 60 occurrences on 24 unique sentences. In fact, the highest numbers of rereadings by a single participant on the same portion of text was 8, which occurred when using this setting. These requests for rereading accounted for 18.3% of the overall requests when using this setting, and 61.2% of participants’ undo requests changes when using this setting. Rereading took place on 28.5% of the sentences that received at least one request for *lexical + syntactic* simplification.

Lastly, 9 participants exhibited this behavior at least once when using *syntactic* simplification alone, with 50 occurrences in total. These accounted for 32% of the overall number of requests for syntactic simplifications and 65% of the instances of participants undoing requests for *syntactic* simplification alone. A total of 29% of the unique sentences that participants requested *syntactic* simplifications for were reread.
Switching Simplification Settings

Considering that participants were provided with the ability to switch between the 3 simplification settings provided by the tool, we also logged how many times participants switched between the options for simplification types. Overall, we recorded a total of 356 instances of participants switching simplification settings. Out of those 356, 49% of the instances consisted of switching to *lexical* simplification, while 29% of the time participants were switching to *lexical + syntactic* simplification, and 22% to *syntactic* alone. This yields an overall average of 13.2 changes per participant, across their reading tasks in which they engaged with the tool, which means that, on average, participants changed 6.5 times to lexical simplification (SD = 3.7, range = 0 to 19), 3.8 times to *lexical and syntactic* simplification (SD = 3.1, range = 0 to 9), and 2.8 times to *syntactic* simplification (SD = 2.5, range = 0 to 9).

12.5.3 Qualitative Results

The tool is beneficial, but there’s space for improvement

Overall, participants suggested that the tool would be beneficial, with most participants commenting on various positive aspects of the tool as a whole or a benefit they could obtain. Participants commented on how the tool “made reading easier” (P25), helped them “understand the whole content” (P7), helped them better understand “jargon” (P1) or “made scientific words clearer,” and “educated” them on “new words that mean the same thing” (P19). Many participants also indicated that they can see themselves using the tool in the future, such as P19 who said “I could definitely see myself using [the tool] in the future for higher level reading.” And some participants expressed that even though they did not feel a great need to use the tool, they could still see how it could still be beneficial in specific situations such as when one is “not in the mood to read the complex words” (P1). Participants also envisioned various reader groups the tool may benefit, including “DHH people especially” (P9), “people who are learning English as a second language” (P6), “high
school, middle school students or even children as they learn how to read” (P20), “people who have dyslexia or ADHD” (P22) or, more generally, people who “are not as fast readers or struggle with reading” (P5).

In spite of the benefits and preferences, however, participants were also cognizant of errors in the simplifications. More specifically, many participants pointed out some grammatical or semantic mismatches when using lexical simplification, which they described as potentially detrimental to their reading experience. For instance, P6 commented that “it took extra time, and I felt like an editor looking more at the grammar and the words rather than what was actually being discussed in the article.” P1, in turn, was concerned of potential changes in meaning as “authors tend to use words for a specific reason.” Other lexical errors that were pointed out by participants included times when the system simply providing an alternative spelling of a word as its simplification. P23, for example, commented “take the word ‘behavior’, I guess it replaced it with the English version ‘B-E-H-A-V-I-O-U-R’. Yes, when it changed to simplify that word it made me a little concerned because it was replaced with the same word just with a ‘u’ [...] I think it should really just simplify the word itself by simplifying the meaning of it”. P12 also brought up that “sometimes I would click the word and the suggestion would come up and I still wouldn’t know what the word meant”. Lastly, P7 also pointed out as an error that upon requesting a simplification for some words, they were simply removed: “for some of the words, when I clicked it to get the suggested word it just removed it, but I don’t understand why they would just remove it.”

When discussing the simplifications that involved syntactic rewritings, in turn, participants also complained that for many sentences, the changes applied were not enough. For example, P4 said “I couldn’t quite understand what this would be used for or how it would help. It seemed like it was just the same information and sometimes it just felt like it was too much.” More specifically, participants found that the restructuring of sentences (which was often referred to as “reordering of words,” noticing the focus on words) did not provide enough support. In the words of P14, for example “The syntactic was good, it just [...] didn’t do much other than switch some of the words around.” Closely related to this,
participants also commented that with these settings, it could be hard to track what had actually changed, which was described as potentially overwhelming: “The [lexical + syntactic] and syntactic were both overwhelming because it was hard to tell what information was being changed in comparison to the original text” (P16). Participants also found it hard to distinguish between the functions of syntactic simplifications alone as compared to lexical + syntactic simplifications. For instance, P21 commented that “I really didn’t see a huge difference between the both and the syntactic options,” which may also involve the visual presentation of the two. P15, for instance, said “For the syntactic option the way that it was set up was a little confusing because it looks the same as [the lexical + syntactic option].” This overall difficulty in identifying changes was identified as negative because of the additional time it takes to do so, often in the form of switching back and forth between the two versions of the text. For example, P6 said “having to switch between the two [versions] is too much of a process and makes it take longer to actually understand and learn the concept of what was read.”

There was an overall, but nuanced, preference for words

Overall, participants displayed a strong preference towards lexical simplification, including P7 who said “I liked the tool because it made it a lot easier to read, especially the Lexical option where it focused on one word that you might not know.” In fact, even participants’ comments regarding the simplifications that involved syntactic rewriting (i.e. lexical + syntactic and syntactic) tended to focus on words as the main unit of change. For instance, when discussing syntactic simplification, P22 said negatively “it wasn’t really able to help me understand any of the words better,” while P8 said positively “I liked being able to change the order of the words to make it more clear.” It also seemed as though participants who did not feel a great need to use the tool could still see the benefit of having lexical simplification as there will always be words one does not know, with jargon being an example brought up by many. For instance, P21 said “the lexical is nice because I can use it when I am reading these scientific articles that I might not know what the specific words are.”
However, participants expected syntactic rewritings to be beneficial in more specific settings. Namely, participants expected syntactic simplification to be more beneficial when reading about unfamiliar topics or for people who struggle with reading more generally. For example, P15 said “if I am just reading the article, I don’t think I would use the “both” option that much unless the topic is completely unclear to me based on the complexity of the article,” while P6 thought that “the syntactic was nice especially if you struggle with the language.” Thus, it seems as though the higher utility participants envisioned for lexical simplification may also stem from the fact that needing support with words is perceived as something more general that anyone might struggle with, whereas syntactic rewritings are more specific to certain tasks, topics or people.

**Participants wanted additional forms of support with words.** Participants’ focus on words also surfaced in participants’ requests for additional features. Many participants expressed a desire to have the ability to expand on a word if the synonym provided was not enough, with the ideas for expansion including definitions, additional thesaurus information such as antonyms, being able to request other synonyms that could be simpler than the one provided, or forms of visual support such as pictures or ASL translations. For instance, P20 suggested “possibly having a second word if the first one is still too complex - like a thesaurus, you could also include a word that means the opposite as well.” P15, in turn, suggested that it would be beneficial to add the option for “another word or even a picture or video of the sign for that word” as it might benefit “deaf people who rely on visual access.”

**The labels should be descriptive of the type of simplification**

Most participants agreed that the use of the labels “lexical” and “syntactic” for the respective simplification types would not be effective. Many participants, for instance, indicated that they understood only because there had been an explanation at the beginning of the experiment. Others expressed concern that even though they understood it, others with lower literacy may not understand what those labels mean. For example, P24 said “not
everyone has the same level of education and therefore they might not be able to understand what those labels mean”.

When discussing what the types of simplification should be labeled, participants indicated that they should be descriptive of their purpose. So, for lexical the suggestions tended to be literal and centered around the focus on words, with many participants suggesting just calling it “words”, while other ideas included “define/describe,” “synonyms,” “detail” and “alternative.” For syntactic, many participants suggested just using the word “sentence,” while other participants suggested labels such as “paraphrase,” “rephrase” or, focusing a bit more on the words, “reorder.” However, most participants were comfortable with using the label “both” for having lexical + syntactic simplification.

**Increase autonomy in requesting simplifications**

Some participants also indicated a desire to request any word, finding the fact that they had to choose from words with available simplifications too limiting. P15 summarized it as:

“The tool itself assumes what words that you may not know and does not let you choose which words you want to be defined. I would like to be able to choose which words I could have access to. It is a lot like when hearing children are able to read a book, ask their parents what the word means and then their parents can give them a definition. […] So, I was wondering who gets to decide which words are chosen to be highlighted and shown with the synonym.” – P15

In the context of using lexical + syntactic simplification, participants also expressed a desire to actually have control over the application of both types. In other words, participants wanted to be able to still request lexical simplifications for specific words after applying the simplifications with the lexical + syntactic setting, as opposed to it being an umbrella for lexical and syntactic rewritings applied simultaneously. P19, for instance, said “my suggestion would be to add to the [lexical + syntactic] option where you would still be able to see which words also have a synonym with the lexical tool.”
The user interface can impact the reading experience

Many participants commented on the importance of the UI for the experience. Participants found the pop-ups to be effective for lexical simplification, citing reasons such as the benefits of being able to compare the original word and its replacement. For instance, P13 said “I felt the pop-ups were good and you could see the new words there which was nice”. P6 also found this comparison beneficial to check the quality of the simplifications: “I liked the pop-ups because you are able to check the accuracy and tell if it was the right or wrong word that popped up.” P7 also thought they might be helpful to keep open for words that reappear across the text: “when there was a word that came up repetitively throughout the article, it was nice to have another word already there.” While some participants also indicated wanting to use pop-ups for all types of simplifications for the simplifications that involved syntactic rewritings, others were concerned about the space it would take like P16 who said “the pop ups were nice for the lexical, I don’t know if I would suggest it for the other two options” or P25, who said they “did have a few concerns with the functionality of it and how you may be limited with space.”

The highlights, in turn, were controversial. While many participants commented both positively and negatively on the colors of the highlights, most suggested this should be adjustable. This preference for colors also related to participants’ current use of highlights. P9, for example, enjoyed the neutrality of the gray highlights because it didn’t interfere with their color highlighter as they read. P17, in turn, suggested replacing the yellow as they use this color often to highlight text. However, participants also had conflicting comments regarding the effect of the highlights in their ability to focus on the words. P20, didn’t like the selective highlighting of words “because it made me want to skip the parts that had no highlight,” whereas P24 said “I just liked using the highlights to help me focus”. While some participants liked it because it allowed them to not skim over complex words, other participants found it distracting or overwhelming, regardless of the color. Participants’ concerns for distraction were especially noticeable in the context of simplifications that
involved syntactic rewritings. For instance, P24 commented that “the highlight being behind all of the text made it really hard to read”, while P6 thought that “it might help some people but for me it was more of a distraction.” So, participants suggested not only allowing the user to adjust color, but also the use of alternative markup (e.g. underline, italics or other font styling). P1, for instance, suggested having an “option to underline or not have the highlighted parts because that can be distracting,” whereas P10 said “I might want to make it so that you could italicize so that you can see it rather than everything being there and being so overwhelming.”

12.6 Discussion

Participants’ subjective judgements of the overall usefulness of the tool, as well as their open-ended comments at the end of their task-based reading activities, suggest that participants found utility in the tool and could envision scenarios in which it could be beneficial. However, we did not observe any statistically significant differences in the reading metrics or in participants’ judgements that focused on the texts between the conditions of reading with and without the tool. Our behavioral data shows that participants did not engage with the fully simplified versions of the articles as no participant requested all the possible simplifications for any text. In other words, our participants did not choose to apply every possible simplification that could have been provided. Thus, metrics that focus on the texts may not fully capture the differences in complexity in the text when participants are allowed to request simplifications on demand. Participants were also cognizant of grammatical errors that the tool may introduce, which may have played a role in participants’ judgements of the tool’s impact in their ability to understand the articles.

12.6.1 Comparing Among the Simplification Types

In terms of participants’ preferences for the various simplification types, we observed an overall trend towards a preference for lexical simplification. This was evidenced by the
statistically significant difference we observed in participants’ subjective judgements of the usefulness of each simplification type, as well as participants’ open ended comments about their experiences and preferences. This trend is also in line with our previous work in this dissertation where we had observed a focus on vocabulary as a main source of difficulty in our work presented in Chapter 4, as well as perceived benefits from lexical simplifications in Chapter 10.

However, there were additional insights obtained from participants’ open ended comments in this study, which suggest that this trend towards preferring lexical simplification should not be interpreted as participants only preferring word-level lexical simplifications, as there was a lot more nuance to their views on the various approaches. We observed that participants saw the applicability of lexical simplifications to be more general as the issue with words may not necessarily be complexity, but simply not knowing the words, and anyone might encounter words they do not know, with a common example being jargon or scientific words. This aligns with prior literacy research that had identified word recognition as a key source of difficulty for college-aged DHH readers [1]. However, participants could see the options that involved syntactic simplification being more beneficial for when the text itself is complex in specific contexts (e.g. reading unfamiliar topics) or for people who might need additional reading support. Furthermore, one of participants’ main complaints about the options that involved syntactic simplifications was that there were not enough changes, which we explore in more detail in the subsection below.

**Syntactic simplifications: “did not really change”**

Participants’ comments regarding how the approaches that involved syntactic simplifications did not provide enough changes could be interpreted at two levels: their content and visual presentation. First, focusing on the content, participants’ suggestions for how to label the syntactic simplification option included “sentences,” “rephrasing” or “paraphrasing.” Thus, it may be the case that, with participants’ focus on words, it was the restructuring a sentence’s syntax that was seen as not being enough. Instead, paraphrasing or rephrasing
of the actual content may be seen as more beneficial (which does not necessarily rule out
the incorporation of syntactic rewritings in the process). Thus, participants’ focus on words
may not mean an exclusive preference for word-level lexical simplifications, but may also
include sentence-level paraphrasing. As we approached this study from the perspective of
informing ATS research, it was important to represent the two broad categories outlined
in the field: lexical and syntactic simplification. In the automatic text simplification lit-
erature, paraphrasing is often categorized as a form of lexical simplification (e.g. [111]).
But, considering that we had observed vocabulary as a key source of difficulty discussed
by participants, we had decided to prioritize word-level lexical simplifications. However,
our insights obtained in this study suggest that, from our participants’ perspective, the
lexical-syntactic dichotomy is not as relevant as the word- vs. sentence-level distinction.
Thus, our findings suggest that future explorations of ATS-based reading assistance tools
with DHH adults should explore this word-sentence dichotomy by not only exploring syn-
tactic approaches at the sentence level, but also incorporating lexical approaches such as
paraphrasing (e.g. [79, 112]) or multi-word replacements (e.g. [97]), which may lead to
more changes in the text.

Then, there was also the visual presentation, where participants found that it was also
difficult to actually track the changes. Tracking changes has been explored in the context
of collaborative writing, by looking at various methods [126, 127] and their effects [15].
However, in that context, the main goals relate to social aspects such as reducing conflict
or showing engagement reviewing, to mention a few [15]. In our reading context, as we had
observed in Chapter 11, participants value maintaining their reading efficiency and fluency
instead. If we look at the percentage of participants’ repeated requests for simplifications
(i.e., the requests related to rereading) out of the total requests made, we observe that
they were higher for the simplification types that involved syntactic simplifications. Now,
while rereading generally been previously observed as a strategy used by DHH readers [10],
in this case the repeated requests to compare the two versions of the text may be due to
participants’ efforts to identify the changes in the text. In other words, when a syntactic
transformation occurs, it can be difficult for someone to spot what had specifically changed in the sentence, and the participants therefore had to flip the simplification on and off for that sentence. This thus highlights the importance of the visual presentation of the simplifications on participants’ reading experience.

12.6.2 User Interface Considerations

Our results highlighted both benefits and downsides to the use of user-interface elements to indicate complex or simplified texts. More specifically, participants found that highlights could become overwhelming and the duality of highlighting, especially with colors that are traditionally used to highlight important portions of texts, can have an impact on participants’ reading experience. Our findings align with prior work on the use of visual indications of confidence in automatic captioning, where participants were also generally receptive to the idea of visual indications of confidence, but they were concerned about their potential to distract [11]. This in contrast to participants’ perspectives on using visual decorations to highlight word importance in captions, where they were seen more positively overall [63]. Thus, the potentially conflicting values between maintaining reading efficiency (which may involve quickly tracking changes in the text) without introducing visual elements (that can distract from their reading) provides an interesting challenge in this space of ATS-based reading assistance tools. Furthermore, balancing these factors may become more challenging when exploring multi-word lexical simplification or paraphrasing, where the number of one-to-one connections between words may decrease.

It is important to note that various participants commented that they wanted greater autonomy in requesting which words or sentences to request simplification for (e.g. being able to request any word or sentence in a text, not only those for which the system indicated it could provide a simplification). This result aligns with our findings from Chapter 10 from the perspective of wanting greater autonomy. However, in that earlier study the notion of autonomy had been more focused on whether they had the choice to request simplifications or not. In this study, we observed that participants also wanted more autonomy in terms of
which portions of the text they can request those simplifications for. From a user-interface perspective, if participants are allowed this choice, then this would also reduce number of visual indications shown, as there would be no need to signify complex words or sentences. Reducing this visual clutter could be beneficial.

When using the setting for lexical + syntactic simplification in our prototype, users have no control over which lexical simplifications are applied (i.e. any word identified as complex in the sentence is replaced after the syntactic rewriting). However, participants also indicated a desire greater autonomy in requesting lexical simplifications for any word after applying syntactic simplifications. Thus, if we approach this from the word-sentence dichotomy, it looks as though the ability to request word-level and sentence-level simplifications should a user-interface mode, where participants can request word-level lexical simplifications even after sentence-level simplifications have been applied. Thus, from a linguistic perspective, sentence-level simplifications can involve both lexical (paraphrasing) and syntactic rewritings. But at the user-interface level, being able to apply request word-level and sentence-level simplifications could be the default mode where participants are allowed to request individual words to be simplified regardless of whether sentence-level simplifications have been applied.

Finally, in our study, we observed a significant difference in participants’ reading speed when comparing their reading tasks in which they used the tool with the tasks in which they did not. This is not an unexpected finding given that the interactions with the tool naturally take additional time (e.g. to switch settings, or request and undo simplifications). While we employed an exploration stage to mitigate this, it is important to consider that this was the first time many of our participants interacted with these tools. Thus, the novelty effect may have motivated additional time in exploring the use of the interface. Furthermore, in light of our findings in Chapters 4 and 11 where participants discussed the importance of saving time when reading, this finding also highlights the importance of the user interface itself for participants’ reading experiences. Thus, there may be a tension between participants’ values of autonomy and saving time which may result in conflicting
preferences for various design choices, as systems that require greater autonomy may require additional interaction time. Considering this tension, and the fact that some design choices may also increase rereading behaviors – which some participants also quantified in terms of the extra time those behaviors took – it is important to distinguish how much time may be added by the additional interactions required to provide users with greater autonomy as compared to the time added based on the time it takes to identify the changes that have been made to make sure the choices support participants’ reading efficiency.

**Evaluating the User Interface**

As we identify participants’ values in the context of interacting with an ATS-based reading assistance tool, our results suggest that considering the linguistic aspects of the resulting text is insufficient. Characteristics of participants’ interactions with the user interface of ATS tools are also important to consider. While our methodological research in Chapters 7 and 8 had identified methods for evaluating the texts themselves, our findings from this study also highlight the importance of evaluating the user interface of the tools, including various factors: the time spent interacting with the tool, the number of repeated requests participants make, and the level of distraction that the tools may introduce, etc. In addition, potential obstruction of visual reading strategies by the tool, such as pointer-assisted reading [71], may influence the effectiveness of the user interface of ATS-based reading assistance tools’ user interfaces. Additional methodological research into such evaluations is needed to shed light into how to best measure these tools effectively.

**12.6.3 Additional Recommendations**

Finally, it is important to consider participants’ recommendations for settings that should be made adjustable, as well as their requests for additional features or support. Participants’ comments suggest that if visual indications such as highlights will be used, then users should be allowed to adjust their colors, but also alternative renderings of the text such as
alternative fonts [123] or font decorations (e.g. boldface, italics).

Participants suggestions’ for additional forms of support such as dictionary definitions, or visual illustrations including ASL definitions, also align with prior work which has identified benefits from providing those to DHH readers approaches [51]. However, that study found that providing those approaches alone may not be sufficient, so it is important to explore the implications of using them in conjunction with simplifications, especially from the perspective of the user interface and their impact of participants’ reading experiences.

12.7 Conclusion

In this study, we explored how DHH adults may benefit from and interact with a fully functional prototype of an ATS-based reading assistance tool capable of providing various types of simplification, and their perspectives on their experiences using the tools. Participants’ subjective judgements indicated that they found the tool useful overall, while comparisons between the simplification types revealed a preference towards word-level simplification. Through participants’ open-ended comments, we observed that participants found the tool beneficial overall, but were concerned about potential errors introduced by the tool. Furthermore, we found that participants’ preference towards lexical simplification may come from the fact that it’s a more generalizable form of reading support as anyone may encounter words that they may not know.

Participants’ preference may not be exclusive to word-level lexical simplifications, but may stem from a slightly different dichotomy from the lexical vs. syntactic perspective typically used in the ATS field. From participants’ perspectives, this dichotomy seemed to be more of word-level vs. sentence-level distinction, where syntactic rewritings may not be enough at the sentence-level and thus there may be space for paraphrasing or multi-word lexical simplifications.

Lastly, our study identified that participants also wanted ways to quickly identify the changes between the original and simplified versions, and thus more work is needed on the
design of the user interface for sentence-level simplifications. Our insights suggest potential ways of measuring these alternative user interfaces, including the time added by interacting with the tool, which was significantly impacted in our study. Other potential measurements may include the number of repeated requests for simplifications participants engage in as those may signify rereading for comparison, as well as the level of distraction or interference with existing reading behaviors that may be introduced.
Epilogue to Part III

In this part, we presented work focused on the design of ATS-based reading assistance tools. We first investigated the perceived benefits of providing lexical simplification to DHH adults, as well as the influence of the level of autonomy provided to users on their subjective preferences about lexical-simplification tools. Then, we conducted an exploration of various design parameters, in addition to those related to autonomy, where we investigated the preferences of DHH adults for setting those different parameters when employing different types of simplifications. Finally, we conducted a final evaluation of the different approaches of ATS among DHH adults. More specifically, in Part III of this dissertation, we investigated the following research questions:

**RQ3.1: Does providing a lexical-simplification reading-assistance provide benefits to DHH adults?** This question was explored in Chapter 10, where our study revealed that users perceived certain designs incorporating lexical simplification as beneficial in making texts easier to understand.

**RQ3.2: Are DHH users’ subjective preferences about lexical-simplification tools affected by whether they are provided with greater autonomy (i.e. in requesting simplifications on-demand or seeing what words have been replaced)?** This question was also explored in Chapter 10, where we found that providing a sense of autonomy influences DHH users’ acceptance of such reading-assistance tools. Specifically, having investigated a variety of
designs that vary in the degree of autonomy they provide (i.e. control of what words are replaced and visibility of past replacements), we found that DHH adults indicate they are more likely to use systems with greater autonomy.

R3.3. When employing ATS-based reading assistance tools that provide syntactic or lexical simplifications, what are the preferences of DHH users for the various parameters involved in the design of these tools? This question was explored in Chapter 11, where we conducted an exploration of the design space with the goal of identifying parameters to employ in an evaluation presented in Chapter 12. In addition to identifying participants’ preferred settings, we also identified participants’ rationale for their choices, which included values such as saving time, maintaining reading fluency and having control over the reading experience.

RQ3.4. When providing an ATS system that provides both lexical and syntactic simplification features, as well as hybrid approaches that combine lexical and syntactic simplification to DHH users, how do users interact with and benefit from such system, and how do their preferences vary for the various simplification approaches? This question was summarized in Chapter 12, where we conducted a study in which people engaged with a tool capable of providing various simplification types using the designs identified in Chapter 11. Overall, we found that participants found the tool useful, but they were mindful of errors in the simplifications. We also observed a strong preference towards lexical simplification, and the potential to explore sentence-level simplifications that involve paraphrasing or multi-word lexical simplifications. We also identified that participants want to track changes efficiently when applying sentence-level simplifications, and that the time added by interacting with the tool, the number of rereadings required to track changes and the level of distraction or interference from the tool are
important to consider in the design of the user interfaces for these tools.
Chapter 13

Conclusion

This dissertation has consisted of research investigating the use of ATS to provide reading assistance to DHH adults. To do this, we explored the needs, interests and requirements of DHH adults with experience in computing for ATS-based reading assistance tools, including their perspectives on the social accessibility of those tools. Then, two methodological studies provided guidance for the evaluation of the complexity and fluency of the output of ATS tools among DHH adults at different literacy levels. Finally, we explored the design of the user interface of ATS-based reading assistance tools by exploring the benefits obtained from lexical simplification and designs that provide the user with greater autonomy, as well as by examining the preferences of DHH adults for the setting of the various design parameters of the user interface for ATS-based reading assistance tools and evaluating a tool that employed the preferences identified.

13.1 Contributions

There are several contributions from each part of this work. Beginning with Part I, the contributions include empirical findings on the requirements of DHH individuals in the computing field, which motivate further technical and design work on ATS-based reading
assistance tools for this group. Our empirical findings suggest that our participants frequently read on electronic devices to learn about new topics for their work, and they had strong interest in these tools. Part I also provided a prioritized list of: a) the most frequent workarounds DHH computing workers currently use to understand complicated text (with the two most frequent approaches being analogous to typical ATS approaches), b) reading purposes for which these individuals would be interested in using ATS, and c) their most frequent purposes for reading text. In Part I, we also presented empirical findings that identified social accessibility issues, along with their design implications, in the context of DHH adults using ATS-based reading assistance tools in a workplace environment. These issues related to a perceived relationship between reading skill and intelligence, which may come into conflict with how participants want to be seen in a professional context if they are seen using tools that may imply a lack of reading skill. Finally, Part I also highlighted the implications of our findings for the design of ATS-based reading assistance tools, which included identifying design variables to consider related to autonomy, users’ concerns about ATS accuracy, and potential design directions that may help mitigate social accessibility issues that may arise with the ways in which these tools are often designed, when used in a public setting such as the workplace.

The contributions of Part II include methodological guidance for evaluating two key characteristics of the output of ATS: complexity and fluency. Our methodological guidance includes which metrics are effective for evaluating each of these characteristics with DHH adults at different literacy levels, as well as which of these metrics require carefully reporting participants’ literacy levels because of literacy biases observed with some metrics. Considering that different user groups may face different challenges when it comes to text complexity [49, 92], our methodological framework can be used by other researchers studying the use of ATS for other user groups to investigate the effectiveness of these metrics for evaluating text complexity with those user groups.

Finally, the contributions of Part III, which focused on the design of ATS-based reading assistance tools for DHH adults, included identifying that DHH users perceived a benefit
CHAPTER 13. CONCLUSION

from tools that provide lexical simplification on demand; prior research had not investigated lexical simplification tools among DHH users. Furthermore, in Part III we also found that DHH users preferred prototypes that provided greater levels of autonomy for ATS-assisted reading. We also found that DHH users could articulate the benefits of autonomy in the interface and comment on the trade-offs among designs in this space. Then, we also analyzed various design parameters typically involved in the design of the user interface of ATS-based reading assistance tools, establishing a framework which can be used by researchers to explore this design space with various user groups. We specifically explored the design space of these tools with DHH adults, identifying their preferences of DHH adults for the design parameters identified when employing the different types of ATS. We also identified factors that affect participants’ preferences, which related to themes and values that were revealed across several of our studies, including saving time, maintaining reading fluency and various perspectives on user autonomy. Finally, Chapter 12 conducted the first evaluation of an ATS-based reading assistance tool with DHH adults capable of providing lexical and syntactic rewritings on demand using the realistic output of state-of-the-art ATS tools. Our empirical findings suggest that while participants found the tool to be useful overall, they were cognizant of errors in the systems and we identified an overall preference towards lexical simplification. Our findings also reveal, however, that there may be space to explore additional forms of support both at the word and the sentence level. Lastly, our evaluation highlighted how design choices to address participants’ values of saving time, maintaining fluency and having greater autonomy may come into conflict, thus highlighting the importance of the user interface on participants’ reading experience.

As a final contribution of this dissertation, the following section provides a detailed analysis of the limitations, highlighting how both our limitations and contributions also open up several avenues for future work.
13.2 Analysis of Limitations and Recommendations for Future Work

The limitations of this dissertation, along with our contributions, highlight several avenues for future work. In the next few sections, we explore some of these avenues, which include: exploring specific subgroups of DHH readers or larger sample sizes that may support the analysis of how demographic factors may affect our findings, as well how our findings may generalize to other user groups or text domains; studies investigating the long-term effects of the use of ATS-based reading assistance tools in realistic settings, as well as additional methods for evaluation; methodological work on how to evaluate other characteristics of simplified texts, and impact of errors in readers’ experience; additional work on the user interface, and other forms of support; and better understanding what can be learned for the context of ATS-based reading assistance tools from the visual reading strategies and sources of text difficulty for DHH adults. The following subsections provide more details about these limitations and directions for future work.

13.2.1 Groups, Demographic Factors and Sample Sizes

In our studies, we recruited DHH participants with a diversity in terms of their demographic characteristics, including their gender, age, occupation, hearing identity, language preference, and literacy skill or level, level of technological self and literary self-efficacy, or their level of comfort using assistive technology. Thus, our findings may not necessarily generalize to DHH adults with different distributions of these demographic factors.

Furthermore, in some of our studies (e.g., in Chapter 4 and Chapter 5), we observed that participants’ preferences or inclinations towards using these tools may vary based on some of these demographic factors. Thus, future work could conduct research with a larger set of DHH participants to support building regression models to understand those potential relationships. Similarly, when we explored participants’ interests and needs in Chapter 4, we had recruited participants with experience in the computing field. Future
work could also explore whether these are reflective of the general DHH population in the U.S., or how they may vary based on participants’ occupations or other demographic factors. These demographic factors, and their intersectionality, may not only influence their preferences towards the technologies, but also may lead to unique characteristics in their reading skills. Thus, future work could also explore this relationship between the intersectionality of various demographic factors and reading skill, and how that relationship may affect readers’ preferences towards ATS technologies.

If participants’ preferences, as well as their reading skills needs and interests are indeed modulated by demographic factors, then future work may also benefit from explorations of these technologies with more specific sub-groups of DHH adults. For instance, when looking at the demographic characteristic of communication preferences, it may be interesting to conduct studies focused specifically on DHH adults who are ASL users. This may also open up alternative methods for evaluating the texts or the tools (which are further discussed below), by allowing the introduction of ASL-based metrics such as comprehension questions presented in ASL. Another characteristic that may be interesting to recruit by is participants’ literacy levels, as those may affect their perceptions of needing the tool. Thus, either recruiting participants with specific literacy levels as a focus, or recruiting larger groups of participants to be able to compare preferences, interests and needs across groups of participants with varying literacy levels. Future work can also explore how these demographic factors may affect participants’ reading behaviors, which may also influence their interactions with ATS-based reading assistance tools.

When it comes to our methodological contributions, we were able to group participants into two groups based on their literacy levels, and observed that participants’ literacy levels did affect the effectiveness of the metrics evaluated. Thus, future work could recruit a greater number of participants, or use a more granular method of measuring literacy, to investigate whether further subdivisions reveals other patterns in the effectiveness of those metrics. These analysis may also benefit from explorations of other demographic factors to understand if those may also modulate the effectiveness of the metrics.
CHAPTER 13. CONCLUSION

Relatively small sample sizes

Amid the participants’ demographics issue describe above, an adjacent limitation is also our sample sizes for our studies. Our relatively small sample sizes did not support the analysis of the relationships between responses and demographic factors. Larger sample sizes may enable us to, for instance, further explore potential relationships between communication preferences and interests in the tool, or the relationship between age and level of comfort using assistive technologies with concerns about social accessibility issues.

This limitation also surfaces in our methodological work presented in Part II, as our sample sizes did not allow us to guarantee that with a larger sample size, and thus more statistical power, significant differences would not emerge for metrics that we did not observe significant differences in the studies presented in Chapter 7 and 8. However, since evaluations of ATS typically involve fewer participants than the number included in our current methodological study, our contribution is still useful to researchers as metrics that did not reveal significant differences in our study would be unlikely to reveal significant differences in evaluations with an equal or smaller sample size. As mentioned above, however, future researchers could benefit from employing participants across broader ranges of literacy levels.

Other people who might benefit from ATS

In this dissertation, we focused exclusively on people who are DHH adults. However, as our participants themselves commented in Chapters 5 and 12, it is also important to explore these technologies with other user groups. In addition to participants’ comments suggesting this, prior work has explored the potential of ATS-based reading assistance tools to benefit various user groups including people with dyslexia (e.g. [99]), aphasia [34], second-language learners [9], and even DHH children [58].

However, there has been little prior work on gathering requirements or assessing interest among these other user groups, nor has there been prior methodological or design work
to understand how to best design and evaluate these technologies for those user groups. Thus, future requirement-gathering work, similar to our work presented in Part I of this dissertation, could be conducted with these other groups to understand how their needs and interests may vary from those of people who are DHH. Future methodological work, as presented in Part II of this dissertation, could focus on exploring the efficacy of the metrics explored in our study with these other user groups. It is also our hope that our design work presented in Part III, including the identification of the various parameters of the design space of ATS-based reading assistance tools, may be used, to identify how their preferences for the various design parameters, and their values that affect those preferences, may differ from those of DHH readers.

Future Evaluations with Target-Reader Groups

Finally, it is our hope that future work will include DHH users, and other target reader groups, in actual user evaluations of different aspects of the output of ATS technologies, including the fluency of the output texts. And, if our methodological work was replicated with other target-reader groups, that they also be included in those evaluations to ensure that the results are representative to the fluency perceptions of target-reader groups.

13.2.2 Other Topics or Complexity Levels

Another limitation in this dissertation is that our studies mostly employed stimuli that focused on science news. As texts in different domains may present specific challenges to readers, participants’ preferences may vary when reading texts from different genres or domains. Thus, future work could explore how our findings may generalize to DHH adults or other user groups, or it could explore reading texts from other genres or domains. Future work could investigate whether the genres or domains of text affect the effectiveness of the metrics identified in Part II for evaluating various aspects of the output of ATS technologies. Furthermore, considering participants’ comments in Chapter 12 regarding
how syntactic simplification may be more applicable in specific contexts, future work could explore whether the preferences of DHH adults towards the various simplification types vary based on the genre or domain of the text.

Furthermore, our study included texts within specific ranges of complexity levels, which means that our findings may not necessarily generalize to other ranges of complexity levels. For instance, the stimuli in the high-complexity conditions in our methodological work in Chapter 7 may not have been challenging enough for participants with higher literacy levels. Future work could determine whether some of the metrics analyzed in our study would be effective for evaluating texts with higher complexity among participants with higher literacy. In turn, our simplified conditions in Chapter 12 may not have been simple enough to observe a measurable impact on participants reading performance, especially considering that participants never saw the fully simplified texts as they were provided the autonomy to request which portions of the text were simplified. Thus, future work could explore how our findings may differ when working with original texts at other complexity levels, as well as simplified texts that vary in their complexity level, too.

13.2.3 Longitudinal and In Situ Studies

A key limitation of our work was also our focus on lab-based studies and the use of surveys, for instance, for participants to self-report their reading habits. Future work could explore the use of diary or observational studies to better understand participants’ existing reading habits, especially considering that reading is a complex social phenomenon that may look different, depending on the context [68]. Furthermore, our studies in Part III had participants engage in brief reading tasks that may not be representative to the tasks they engage in their day-to-day life.

Our lab-based sessions also made it difficult to assess the long-term impact of the use of an ATS-based reading assistance tool, or the potential benefits and downsides of such long-term use. Thus, future work could employ longitudinal, or in situ observational studies, to
better understand how DHH adults would engage with ATS-based reading tools in various environments, as well as to understand the impact of the use of these technologies in their text comprehension skills and reading experiences.

13.2.4 Additional Methodological and Evaluative Work

In this dissertation, we focused on how to evaluate two aspects of the potential output of an ATS system: the level of complexity of the output and the fluency of automatically simplified texts, which may be damaged as a result of grammatical errors being introduced. However, semantic errors may also be introduced in the simplification process [110]. Those types of semantic, or meaning-preservation errors are typically evaluated by asking expert readers to examine the original and simplified texts side-by-side. However, future work could explore whether this type of evaluation is possible among DHH readers, or whether some of the metrics employed in our studies (e.g., comprehension questions) can effectively distinguish the loss of information caused by those errors. Future work should focus on how to holistically measure the overall quality of the output of simplification.

13.2.5 The Impact of Errors

In Chapter 12, we conducted the first exploration of DHH adults’ reading experiences when engaging with the realistic output of ATS technologies in a reading assistance tool. And we observed that participants were sensitive to errors in the simplifications. However, we did not measure the potential impact of those errors on participants’ reading experiences and preferences towards the tool. Future work could focus on this impact specifically. And, considering participants’ comments in Chapter 12 regarding how the technology needs to improve, future evaluative work can also explore whether there are thresholds of quality that participants may find sufficient (or, inversely, thresholds for the levels of errors that participants may find tolerable) to enable the use of these imperfect automatic technologies.
13.2.6 Additional Measurements (e.g. Eye-Tracking)

In the studies presented in Part II, we only evaluated one type of comprehension questions, namely, multiple-choice questions. Thus, our findings may not generalize to other types of comprehension questions, e.g., cloze tests or summarizing tasks. Future work could also incorporate comprehension questions recorded in ASL as a possible metric for use among participants who are ASL signers.

It was our original intention to also explore the effectiveness of measures such as those obtained through eye-tracking (e.g. fixation duration) in our methodological work. However, we were prevented from using eye-tracking because we had to conduct our study remotely to abide by social-distancing restrictions due to COVID-19. Future work can thus investigate the effectiveness of eye-tracking for distinguishing text complexity levels among DHH readers. Considering participants’ mouse-tracing behavior mentioned in Chapter 11, future work can explore the effectiveness of other measurements which have been found to be effective at measuring complexity with other reader groups, such as the level pointer-assisted reading [71] or scrolling behaviors [49], when evaluating complexity with DHH adults.

13.2.7 User Interface

Our work provides several contributions towards the design of the user interface of ATS-based reading assistance technologies, with a focus on the preferences of DHH readers. These include more generally applicable such as outlining the various design parameters involved in the tool to support explorations of this design space. However, the goal of our exploratory studies presented in Chapters 10 and 11 were narrow: we wanted to identify reasonable choices to use in subsequent studies exploring other aspects of the tools. With that narrow view, we thus recruited small number of participants which did not support an analysis of all possible combinations of the design parameters to identify statistically significant differences in their preferences, or the optimal choices for this reader group.
Thus, future work could explore these design parameters with larger groups of readers to identify these optimal settings.

Furthermore, three is an additional design parameter that was not explored in this dissertation, which is the platform used to allow the users to interact with ATS tools. This design parameter represents whether the simplification service is provided to users through a dedicated application (e.g. [17, 129]), browser extensions (e.g. [14, 27, 101] and our work in Part III), or a website with text boxes to submit the text itself or URLs containing the text (e.g. [34, 86, 105]). Future work could explore how variations of the design parameters explored in this dissertation may be employed to attend to readers’ values in, and determine how much our findings generalize to those platforms.

The recent emergence of chat-based interfaces powered by generative AI technologies such as ChatGPT\(^1\), which are capable of providing automatic text simplifications also brings exciting opportunities for future work in the context of ATS. More specifically, additional work is needed to explore how our findings may generalize to reading texts simplified using a chat-based interface. Considering participants’ desire to efficiently compare the original text to the replacements in our study, the natural separation that comes from the sequential nature of chat-based interfaces may introduce additional challenges in supporting readers’ fluency and efficiency.

Our work also focused on user interfaces appropriate for mouse-based laptop or desktop computers. However, with the increasing popularity of touchscreen interfaces for reading \([25]\), many times on smaller devices such as tablets or smartphones, future work should explore how our findings may generalize to those interfaces. We speculate that while some findings, such as users’ values for efficiency and autonomy may translate, some design preferences may have been affected by the mouse interactions and thus may not generalize to touchscreen devices (e.g. participants preferring manual duration to avoid mouse interference). Furthermore, it would be important to explore whether ATS can be beneficial beyond the constraints of digital environments by looking at its use for printed or written

\(^1\)https://openai.com/blog/chatgpt
texts, for example.

Based on our findings from our study, future work should also look closer into how to allow participants to efficiently track changes when using sentence-level simplifications, which may include highlights or other forms of text decoration. Future work should also look into the relationship between the color used for the decorations and participants’ preferences, or identify whether colors should indeed be adjustable. As participants brought up in many of our studies, future work should also ensure that these visual text decorations remain accessible to Deaf-Blind readers.

Another contribution of our work included identifying participants’ values when using these tools (e.g. time, reading fluency and autonomy). Thus, future work should explore how various designs may approach further supporting these values, which at times may become conflicting (e.g. a design that provides greater autonomy may require more time from the reader). Participatory design approaches can be employed to explore design ideas with intended target readers. Future work should also explore how to balance moving towards universal design while also attending to potentially differing user needs.

13.2.8 Additional and Adjacent Simplification Capabilities

While in this dissertation we conducted explorations of the design space focused on several design parameters of ATS-based reading assistance tools, there were additional design parameters that we did not explore, including some that participants expressed desire to incorporate into the tool. Once such parameter is the number of simplifications provided to the user. Future work should explore how providing more than one alternative for a portion of a text may affect participants’ preferences towards the various design parameters explored in this work as a form of additional support.

Furthermore, participants’ comments in Chapter 12 also indicated that they would be interested in additional forms of reading support. As we only provided word-level lexical simplifications, but identified a different dichotomy from DHH participants’ perspectives of
words vs. sentences, future work should explore paraphrasing approaches at the sentence level to better understand how the preferences of DHH readers may vary when employing those. Furthermore, several participants across our studies also suggested the exploration of other forms of support such as dictionaries, thesaurus, pictures, videos or ASL signs for specific words. While prior work has examined some of these approaches with DHH readers [51], future work should explore how these approaches may work in conjunction with ATS.

Lastly, future work should explore how our findings may generalize to other adjacent technologies when provided to DHH readers or other user groups. For instance, considering that ATS can be considered a mono-lingual translation (i.e., a complex text translated to a simpler text in the same language), some of our work on the user interface of ATS-based reading tools may be applicable to the context of providing translations to bilingual readers.

13.2.9 Visual Reading Strategies and Understanding Text Complexity

While we had a strong body of prior research to rely on to understand the cognitive and metacognitive reading strategies employed by DHH readers [10], as well as information regarding the visual perception of these readers in reading contexts, there has been less work on the visual reading strategies they may employ to support comprehension, such as the pointed-assisted reading behavior participants referred to in Chapter 12. Thus, future work could focus on understanding these visual reading strategies to ensure that the user interfaces of ATS-based reading assistance tools do not interfere with such strategies or how they can be leveraged in the design of these tools.

Prior research has investigated sources of difficulty for DHH readers, identifying, for example, word recognition as an important source of difficulty [1]. However, little research has looked into what characteristics of a text makes it difficult for DHH readers. Considering that prior work has identified that what makes text complex varies depending on the reader group [92, 99], it is important to conduct future research to investigate what makes text complex for DHH readers so that ATS strategies can be adapted to match those linguistic
needs. As part of this project, in work not presented in this dissertation, we also contributed a dataset of word-complexity judgements from DHH adults, identifying that the way various characteristics of words affected text complexity may differ from the way they affect text complexity for second-language learners [3]. However, future work should investigate what characteristics of those words particularly affect text complexity for these reader groups. Lastly, as self-judgements of text complexity may rely on meta-cognitive skills of literacy, it would also be important to understand how much of the preference observed in Chapter 12 towards word-level lexical simplifications may also stem from greater difficulty in assessing the complexity of sentences and complex syntactic structures.

13.3 Concluding Thoughts

Considering the diversity in literacy skill among DHH adults, as well as their under-representation in fields like computing and IT [90], it is our hope that our research focused on the use of ATS-based reading assistance tools may lead to better reading support for DHH adults, especially those who may need to engage in self-directed learning tasks for their professional advancement.

Our approach to this technology from the perspectives of human-computer interaction and computer accessibility aims to inform researchers working on the underlying technologies from a natural-language processing perspective. More specifically, our work provides further motivation for improving the underlying ATS technologies in support of this user group, as well as guidance on how to evaluate their technologies. Beyond motivating technical advancements, our work also highlights the importance of human-centered design when employing AI-based technologies to support people with disabilities, as we identify there are several values that ATS-based reading assistance tools need to consider to effectively support the reading tasks of DHH adults.

Finally, we hope that our human-centered work on ATS-based reading assistance tools also motivates other researchers in the field who are working with other user groups to also
CHAPTER 13. CONCLUSION

conduct similar work into the design and evaluation of these technologies. Ultimately, the goal is to empower the target users of these technologies and increase their participation in the process of creating these technologies.
Bibliography


[37] Lisa Elliot, Michael Stinson, James Mallory, Donna Easton, and Matt Huenerfauth. Deaf and hard of hearing individuals’ perceptions of communication with hearing


[63] Sushant Kafle, Peter Yeung, and Matt Huenerfauth. Evaluating the benefit of highlighting key words in captions for people who are deaf or hard of hearing. In Proceedings of the 21st International ACM SIGACCESS Conference on Computers and


[71] Ilan Kirsh and Mike Joy. Exploring pointer assisted reading (par): Using mouse movements to analyze web users’ reading behaviors and patterns. In *HCI International*


[78] John L Luckner and C Michele Handley. A summary of the reading comprehension


[118] M. Cecil Smith. Reading habits and attitudes of adults at different levels of education


[125] Dawn Walton, Georgianna Borgna, Marc Marschark, Kathryn Crowe, and Jessica Trussell. I am not unskilled and unaware: deaf and hearing learners’ self-assessments


Appendices
Appendix A

Publications

The research presented in this dissertation has led to five technical papers presented at both the ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), as well as the ACM CHI Conference on Human Factors in Computing Systems is the premier international conference of Human-Computer Interaction. We also published a workshop paper at the Workshop on Text Simplification, Accessibility, and Readability. These publications include:

**CHI’20:** The work presented at the CHI’20 conference was summarized in Chapter 10, focusing on the benefits of providing DHH adults with lexical simplification technologies, as well as the benefits obtained from providing these users with designs that provided more autonomy [4].

**ASSETS’20:** Our study focused on the needs, interests and requirements of DHH adults for ATS-based reading assistance tools, as summarized in Chapter 4 was also presented at the ASSETS’20 conference [2]. In that paper, we presented the results from our online survey and follow-up interviews investigating their current reading experiences, as well as their interests in ATS-based reading assistance tools for supporting their professional learning tasks.
**CHI’21:** Finally, the work summarized in Chapter 7, which focused on comparing metrics for evaluating the complexity of simplified texts among DHH adults at different literacy levels, was also presented at the CHI’21 conference [5].

**TACCESS (Volume 15, Issue 2):** An extended version of our ASSETS’20 paper was published in the TACCESS journal. This extended version of our paper includes the work presented in Chapter 5 on the perspectives of DHH adults with experience in computing on the social accessibility of ATS-based reading assistance tools.

**CHI’22:** The work presented in Chapter 8, which evaluated metrics for measuring the fluency of automatically simplified texts among DHH adults with varying literacy levels, was presented at the CHI’22 conference [6].

**TSAR’22:** We also collected a dataset of word-complexity judgements from DHH adults to provide to the text simplification community as training data for lexical simplification systems [3].
Appendix B

Video Demonstration

The video demonstration used during the survey and interview studies for Chapters 4 and 5 can be found at the following URL: https://drive.google.com/file/d/1lfkGiIE8sCB0D13FXWwo80EYoYpvWcAwL/view
Appendix C

Survey and Follow-Up Interview Materials

C.1 Survey Questionnaire

C.1.1 Introduction

In this survey, we will be asking you questions about:

- How you learn about new topics for your job.

- What you do to understand complicated text online. Your suggestions about technology that could help simplify reading text online.

We have found the survey generally requires approximately 10 to 12 minutes. At the end of the survey, you will have the opportunity to enter your e-mail for a raffle to win a $100 gift card.

C.1.2 Questions

1. In general, how often do you read?
□ Rarely (less than once a month)
□ Monthly (one to three times a month)
□ Weekly (once a week)
□ Often (two to four times a week)
□ Daily (five or more times a week)

2. On a day that you read, how much time do you spend reading text on a screen (e.g. computers, laptops, phones, tablets, etc).

   ◦ None at all
   ◦ 1-4 minutes
   ◦ 5-9 minutes
   ◦ 10-15 minutes
   ◦ 15-29 minutes
   ◦ 30-60 minutes
   ◦ Over 60 minutes (please specify)
   ◦ This doesn’t apply to me

3. On a day that you read, how much time do you spend reading text that is not on a screen (e.g. books, magazines, newspapers, print-outs).

   ◦ None at all
   ◦ 1-4 minutes
   ◦ 5-9 minutes
   ◦ 10-15 minutes
   ◦ 15-29 minutes
   ◦ 30-60 minutes
APPENDIX C. SURVEY AND FOLLOW-UP INTERVIEW MATERIALS

○ Over 60 minutes (please specify)
○ This doesn’t apply to me

4. For what purposes do you read? (select all that apply)

□ Work (e.g. technical text, specific terminology and jargon)
□ Academic (e.g. research papers, scientific articles, class, exams, textbooks)
□ Medical (e.g. health insurance, diagnosis)
□ Legal (e.g. terms of service, contracts)
□ Personal communication (e.g. e-mail, text messages, social networks)
□ Visual media (e.g. movies, tv shows)
□ Personal reading (e.g. books)
□ Recreation (e.g. restaurant menus)
□ News (e.g. newspapers, magazines)
□ Other (please specify) ____________________________

5. How often do you read to learn about technical topics for your work (e.g. information about new technologies, new software, programming information, etc.)?

○ Rarely (less than once a month)
○ Monthly (one to three times a month)
○ Weekly (once a week)
○ Often (two to four times a week)
○ Daily (five or more times a week)

6. When you read to learn about technical topics for your work, which websites do you use? ____________________________

7. How often do you watch videos to learn about technical topics for your work (e.g. information about new technologies, new software, programming information, etc...
○ Rarely (less than once a month)
○ Monthly (one to three times a month)
○ Weekly (once a week)
○ Often (two to four times a week)
○ Daily (five or more times a week)

8. When you watch videos to learn about technical topics for your work, which websites do you use? _______________________

9. When you learn about technical topics for your work, what are some situations in which you would prefer:
   • Watching videos _______________________
   • Reading _______________________

10. How often do you encounter text that is complicated?
    ○ Never
    ○ Seldom
    ○ Neither Seldom not Often
    ○ Often
    ○ Very Often

11. How do you try to understand text that is complicated? (select all that apply)
    □ I use a dictionary
    □ I look for a translation to ASL
    □ I ask coworkers for help
    □ I ask a supervisor for help
    □ I look for other websites talking about the same topic
APPENDIX C. SURVEY AND FOLLOW-UP INTERVIEW MATERIALS

□ Other (please specify): __________________________

□ This doesn’t apply to me

12. Please rate how much you agree with the following statement: I’m happy with the techniques I currently use to understand complicated text.

   ○ Strongly agree
   ○ Somewhat agree
   ○ Neither agree nor disagree
   ○ Somewhat disagree
   ○ Strongly disagree

13. When trying to understand text that is complicated, please indicate how much you agree with the following statements.
14. Please indicate how much you agree with the following statements

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Somewhat Agree</th>
<th>Neither Agree nor Disagree</th>
<th>Somewhat Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>It is important for doing my job well</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
</tr>
<tr>
<td>It is important for doing my job quickly</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
</tr>
<tr>
<td>It is important for enjoying my job</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
</tr>
<tr>
<td>It is important for doing my job independently</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
</tr>
<tr>
<td>It is important for being respected by colleagues</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
</tr>
<tr>
<td>It is important for being able to find information about my work</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
</tr>
<tr>
<td>It is important for learning more vocabulary and technical words for my work</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
<td>◯</td>
</tr>
<tr>
<td></td>
<td>Strongly Agree</td>
<td>Somewhat Agree</td>
<td>Neither Agree nor Disagree</td>
<td>Somewhat Disagree</td>
</tr>
<tr>
<td>-----------------------------------------------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>-----------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>I feel that I am responsible for my own learning at work</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can set my own goals for learning at work</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I have control over what I learn at work</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I have a role to play in my own learning at work</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I make decisions about how my learning progresses at work</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

15. Please watch the following video before continuing with the survey. [Video link can be found in Appendix B]
16. Please choose your interest in a tool like the one in the video for each activity below.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Not Interested</th>
<th>Slightly Interested</th>
<th>Somewhat Interested</th>
<th>Very Interested</th>
<th>Extremely Interested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work (e.g. technical text, specific terminology and jargon)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Academic (e.g. research papers, scientific articles, class, exams, textbooks)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Medical (e.g. health insurance, diagnosis)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Legal (e.g. terms of service, contracts)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Personal communication (e.g. e-mail, text messages, social networks)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Visual media (e.g. movies, tv shows)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Personal reading (e.g. e-books)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Recreation (e.g. restaurant menus)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>News (e.g. newspapers, magazines)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Other (please specify)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>
17. Please indicate how much you agree with the following statement. I would be interested in a tool that helps me to:

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Agree</th>
<th>Somewhat Agree</th>
<th>Neither Agree nor Disagree</th>
<th>Somewhat Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learn new vocabulary.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Understand text by making it simpler.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Read better by making complicated text easier to read</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Get my job done faster by reducing the time it takes to read complicated text.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Get my job done well by helping me understand text that is complicated.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Enjoy my job more by helping me understand text that is complicated.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>
18. Please indicate how much you agree with the following statement. I would be interested in a tool that helps me to:

<table>
<thead>
<tr>
<th>I would be upset if the tool replaced text before I got to see it.</th>
<th>Strongly Agree</th>
<th>Somewhat Agree</th>
<th>Neither Agree nor Disagree</th>
<th>Somewhat Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
</table>

| I would be upset if the tool replaced text without asking me. | 0 | 0 | 0 | 0 | 0 |

| I would be embarrassed if my colleagues noticed I was using a tool like this. | 0 | 0 | 0 | 0 | 0 |

C.1.3 Information About Yourself

19. Age: ___

20. Gender

- Male
- Female
- Other _________
- Prefer not to answer

21. Do you identify as:

- deaf
- Deaf
- Hard of Hearing
22. Highest degree obtained
   - High School
   - Associate’s
   - Bachelor’s
   - Master’s
   - Doctoral
   - N/A
   - Other (please specify) __________

23. Are you a current university student?
   - Yes
   - No

24. Within computing, what type of work have you done? (select all that apply)
   - [ ] Computer Network Architect
   - [ ] Computer Programmer
   - [ ] Computer Systems Analyst
   - [ ] Database Administrators
   - [ ] Database Architects
   - [ ] Information Security Analyst
   - [ ] Network Support Specialist
   - [ ] Research Scientist
   - [ ] Software Developer
   - [ ] Software Quality Assurance Analysts and Tester
□ User Support Specialist (e.g. Help Desk Technician, IT Support Specialist)
□ Web Designer
□ Web Developer
□ Other (please specify) __________

25. How many years have you worked in the Computing or Information Technology field?

__________

26. Was your most recent work experience part of a co-op/internship?

○ Yes
○ No

27. Overall, I prefer to use:

□ Spoken language only
□ Mostly spoken language, with a little sign language
□ About half spoken language, half sign language
□ Mostly sign language, with a little spoken language
□ Sign language only
□ Other (please specify)

28. Please rate your skill in producing ASL

○ No skill
○ Some skill
○ Average skill
○ Above average skill
○ Excellent skill
29. Please rate your skill in understanding ASL
   - No skill
   - Some skill
   - Average skill
   - Above average skill
   - Excellent skill

30. How comfortable are you writing in English?
   - Not comfortable at all
   - Not comfortable
   - Neutral
   - Comfortable
   - Very comfortable

31. How comfortable are you reading English text?
   - Not comfortable at all
   - Not comfortable
   - Neutral
   - Comfortable
   - Very comfortable

32. If you’d like to be included for the raffle, please type your e-mail address: _________

33. Is it okay if we contact you for a follow-up interview? (The interview would be a 70-minute appointment over video call, and you would be compensated with $40 for being in it)
   - Yes
   - No
C.2 Follow-Up Interviews Questionnaire

C.2.1 Warm-up

1. When you think about reading, what’s the first thing that comes to mind?

C.2.2 Reading

2. How often do you read?

3. What are some topics you usually read about?

4. Thinking about the activity of reading, what do you enjoy about reading?

5. What are some things you do not enjoy about reading?

6. What are some of the purposes you read for?

7. What are the reasons that you read for those purposes?

C.2.3 Learning for Work

8. Do you ever have to learn about new topics by yourself for your job, including academic work?

9. If yes to previous, what are some ways in which you learn about these new topics?

10. When you have to learn about new topics for your job by yourself, do you prefer to watch videos about it or to read?

11. Under what conditions do you prefer reading over watching videos to learn about new topics for work?

12. Under what conditions do you prefer watching videos over reading to learn about new topics for work?
C.2.4 Facing Complicated Text

13. Some people can read very well, while others struggle with reading. Why is this, do you think?

14. From what you’ve seen, what’s the difference between people who read well and people who struggle with reading?

15. What about you, how would you describe your level of reading? Why?

16. That was talking about people. Now, talking about text: there are texts that are complicated and texts that are easy to read. Why is this, do you think?

17. When you read, do you ever encounter text that is too complicated?

18. What is your reaction when that happens?

19. What do you do to try to understand text that is complicated?

C.2.5 Reactions to Tool

Introduce video: [Video link included in Appendix B]

20. What do you think about the tool shown in the video?

21. Let’s think about pros and cons. Under what conditions do you think this would be useful for you?

22. What are some situations that you think this would not be useful for you?

23. What do you think would be the benefit of a tool like this?

24. What are some negative impacts that you think a tool like this could have?

25. Would you consider using this for your job? Why?
26. What are some things you would not use this for?

27. Consider we told you that this system is not 100% accurate every time, how would that affect your perception of it?

28. If you were using a tool like this and a co-worker saw you using it, how do you think you would feel?

29. What would you think if you saw a co-worker using a tool like this?

30. How would you describe the tool to that co-worker?
Appendix D

Social Accessibility Interview Materials

D.1 Questionnaire

D.1.1 Warm-up

1. What type of work do you do / have you done?

2. How long have you done this work for?

3. Do you like reading? Why or why not?

4. What type of things do you like to read?

5. How much reading does your work require? Do you usually read for work?

6. How often would you say you use dictionaries or other tools to help you in reading?

D.1.2 Work

Let’s talk a little bit more about work.
7. What are the 3 most important things to you at work?  
   a. Imagine that you are putting in your best work at your job, what does that look like?  
      What are you doing?  
      i. How are you doing that?  
   b. Imagine the ideal workplace for you. What would that workplace look like for you?  

8. Think about your workplace environment, where you work with different colleagues.  
   What do you want your colleagues to think about you in this professional environment?  
   a. Would this change if that colleague is your boss? How so?  
   b. Would this change if that colleague is someone you manage? How so?  

9. What are some situations in which you would ask others for help at work?  

10. How do you feel about asking others for help at work?  

D.1.3 Reading / Assistive Technologies  

Now, let’s talk a little bit about reading and assistive technologies that may help you with reading.  

11. Many Deaf or Hard-of-hearing people can read English very well, while others may have difficulty when reading. If a hearing person asked you about this, how would you explain it to them?  

12. How would you personally feel about using assistive technologies to access written information while reading (for example: dictionaries, summarization tools, or tools specifically designed for making text easier to read)?  

D.1.4 ATS Tools  

Now, some people are working on tools created specifically for taking complicated text and making it easier to read. Let me show you a video of what this can look like. Pay
attention to what it does, what it looks like and how it works. Show video demo [Video link included in Appendix B].

Personal

13. What would you think about a tool like that?

14. Some people call that a “text simplification tool.” What do you personally think about calling a tool like that a “text simplification tool?”

15. What would you say if someone recommended using a tool like the one shown in the video?

16. This question could vary depending on the response to previews questions:

(a) If they’re okay with it: Some people that we have talked to have said that they would find it offensive if someone suggested using a text simplification for reading English. Why do you think they would find it offensive? Is it because the other person is suggesting it? Is it because of what the tool does? Is it because the tool is called a “text simplification tool”? OR

(b) If they would find it offensive: So you think it would be offensive if someone suggested using a “simplifier.” I’d like to understand what part of it would be offensive. Is it because the other person is suggesting it? Is it because of what the tool does? Is it because the tool is called a “text simplification tool”?

17. Are there other ways that you could imagine calling this tool in a way that could be less offensive?

Others

Now, let’s imagine using a tool like the one in the video when you’re reading specifically at work.
18. What do you think that colleague would think if they saw you using a tool like the one shown in the video? How would you feel about them thinking that?

- Would you think differently if it was your boss? How so?
- Would you think differently if it was someone you manage? How so?
- **IF THERE'S TIME:** How would you feel differently if you were using it around family or friends?

19. This question could vary depending on the response to the previous questions:

- **If they're okay with it:** Some people that we have talked to have said that they would be embarrassed if their colleagues saw them using a tool like this. Why do you think they would be embarrassed? Is it because of what the others would think? Is it because of what the tool does? Is it because of how it looks? **OR**

- **If they would be embarrassed:** So you think it would be embarrassing if a colleague saw you using a tool like this I'd like to understand what part of it would be offensive: Is it because of what the others would think? Is it because of what the tool does? Is it because of how it looks?

20. How could it look different to make it less embarrassing?

21. Do you think people's decisions to use tools that simplify text at work or school would be influenced by what their colleagues think? Why or why not?
Appendix E

Articles for Methodological Studies

This appendix contains the links to each of the Newsela articles that were used in our study presented in Chapter 7 as reading stimuli in each complexity condition, identified by the codenames we created for each one of them based on their main topic. The same articles, in their High (Original) and Medium complexity levels were used in the procedure for generating our stimuli for our study presented in Chapter 8.

Bubonic Plage:

- High Complexity (Original):
  https://newsela.com/read/plague-prairie-dogs/id/56211/

- Medium Complexity:
  https://newsela.com/read/plague-prairie-dogs/id/56214/

- Low Complexity:
  https://newsela.com/read/plague-prairie-dogs/id/56212/

Cryodrakon:

- High Complexity (Original):
  https://newsela.com/read/frozen-dragon-of-north/id/57267/
• Medium Complexity:
  https://newsela.com/read/frozen-dragon-of-north/id/57271/

• Low Complexity:
  https://newsela.com/read/frozen-dragon-of-north/id/57268/

Garfield:

• High Complexity (Original):

• Medium Complexity:
  https://newsela.com/read/garfield-beach-phone-plastic-problem/id/50717/

• Low Complexity:
  https://newsela.com/read/garfield-beach-phone-plastic-problem/id/50714/

Lizard:

• High Complexity (Original):
  https://newsela.com/read/extinct-australian-lizard/id/53264/

• Medium Complexity:
  https://newsela.com/read/extinct-australian-lizard/id/54103/

• Low Complexity:
  https://newsela.com/read/extinct-australian-lizard/id/54102/

Salmon Cannon:

• High Complexity (Original):
  https://newsela.com/read/salmon-cannon/id/56046/

• Medium Complexity:
  https://newsela.com/read/salmon-cannon/id/56047/
• Low Complexity:
  https://newsela.com/read/salmon-cannon/id/56048/

Smartphone Typing:

• High Complexity (Original):
  https://newsela.com/read/smartphones-typing-texts/id/2000000267/

• Medium Complexity:
  https://newsela.com/read/smartphones-typing-texts/id/2000000269/

• Low Complexity:
  https://newsela.com/read/smartphones-typing-texts/id/2000000271/
Appendix F

Articles for Chapters 11 & 12

This appendix contains the links to each of the ScienceDaily articles that were used in our studies presented in Chapters 11 and 12 as reading stimuli. The first six articles were used in Chapter 11. In Chapter 12, the first article was used to allow participants to familiarize themselves with the tool, while the last four were used for the actual reading tasks.

1. Creating the human-robotic dream team:
   https://www.sciencedaily.com/releases/2021/12/211214104202.htm

2. Fast changes between the solar seasons resolved by new sun clock:
   https://www.sciencedaily.com/releases/2021/08/210817131420.htm

3. Fully 3D-printed, flexible OLED display:
   https://www.sciencedaily.com/releases/2022/01/220107164610.htm

4. How robots learn to hike:
   https://www.sciencedaily.com/releases/2022/01/220121124853.htm

5. New electronic paper displays brilliant colors:
   https://www.sciencedaily.com/releases/2021/07/210712092227.htm
6. New software predicts the movements of large land animals:

7. Nonverbal social interactions – even with unfriendly avatars – boost cooperation:

8. Scientists uncover how the shape of melting ice depends on water temperature:
   https://www.sciencedaily.com/releases/2022/01/220131083848.htm

9. Spiders’ web secrets unraveled:
   https://www.sciencedaily.com/releases/2021/11/211101105356.htm

10. Using ergonomics to reduce pain from technology use:
    https://www.sciencedaily.com/releases/2021/12/211217123806.htm