Gloss Positioning for a Gaze Aware L2 Reading Aid

by

Zixin Zhao

A thesis submitted to the
School of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

Faculty of Science
University of Ontario Institute of Technology (Ontario Tech University)
Oshawa, Ontario, Canada
August, 2023

Thesis Examination Information

Submitted by: Zixin Zhao

Master of Science in **Computer Science**

Thesis Title: Gloss Positioning for a Gaze Aware L2 Reading Aid

An oral defense of this thesis took place on August 4, 2023 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee: Dr. Pooria Madani

Research Supervisor: Dr. Christopher Collins

Examining Committee Member: Dr. Alvaro Quevedo

Thesis Examiner: Dr. Pejman Mirza-Babaei

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Abstract

An important aspect of language learning is reading texts written in the learner's second language (L2). However, reading can be a daunting task for language learners, as text may contain a large amount of unknown vocabulary. We developed a gaze-informed application that has the ability to augment articles with information about the text during reading, dynamically providing overlaid captions – or glosses – interlinearly, in the right margin, and a combination of both glosses. To test the usability of each type of gloss, we ran participant studies (N=19) with French learners where marginal glosses led to an increase in vocabulary matching scores (*p<.01*). We found that gloss positioning preferences mainly depend on L2 proficiency, with intermediate learners preferring marginal glosses and beginner learners preferring interlinear and combination of both glosses. Overall, participants all found that glossed text was helpful, but personal goals and L2 proficiencies affected gloss preferences.

Keywords: human-computer interaction; language learning; eye tracking

Author's Declaration

I, Zixin Zhao, hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Ontario Institute of Technology (Ontario Tech University) to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Ontario Institute of Technology (Ontario Tech University) to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

The research work in this thesis that was performed in compliance with the regulations of Research Ethics Board under REB Certificate number 16293.

____Zixin Zhao

Statement of Contributions

I hereby certify that I am the sole author of this thesis and that no part of this thesis has yet been published or submitted for publication as of August 23, 2023. I have used standard referencing practices to acknowledge ideas, research techniques, or other materials that belong to others. Furthermore, I hereby certify that I am the sole source of the creative works and/or inventive knowledge described in this thesis.

Acknowledgements

First, I would like to thank my supervisor Dr. Christopher Collins for the numerous advice given throughout my entire M.Sc. process. His knowledge and enthusiasm for innovation pushes me to stay curious and pursue further studies. During my studies, he has consistently allowed me to work independently at my own pace, offering support to guide me towards my goals, which I highly appreciate.

Thank you to my fellow lab mates at vialab who provided me with support throughout this past two years. Especially by giving me invaluable feedback about my projects, keeping me on track with my work, and everything else in between.

Finally, I'd like to thank my friends, partner, and sister for your constant support towards my academic pursuits. Thank you for being there for my last minute decisions and pushing me to be the best version of myself. Also, for listening to my rants and letting me daze off on the camera during video calls.

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

API	Application Programming Interface
CALL	Computer Assisted Language Learning
CEFR	Common European Framework of Reference
CLB	Canadian Language Benchmark
FABRA	French Aggregator-Based Readability Assessment
FS	Full System with interlinear and marginal gloss
HCI	Human Computer Interaction
IG	Interlinear Gloss
L1	First Language
L2	Second Language
LL	Language Learning
MG	Marginal Gloss
NCLC	Niveaux de Compétence Linguistique Canadiens
NH	No Help
NLP	Natural Language Processing
POS	Parts-of-Speech
RBF	Radial Basis Function
RNN	Recurrent Neural Network
SLA	Second Language Acquisition
SVM	Support Vector Machines
TOEIC	Test of English for International Communication
TTR	Type-Token Ratio

Glossary

gloss	help text which contain information on words via definitions or synonyms
lemma	dictionary or root form of a word
lexical unit	a single word, part of a word, or chain of words (catena) that forms the basic elements of language's vocabulary

1 Introduction

Second language acquisition (SLA) differs from children learning their first language (L1). By definition, a second language (L2) learner is a learner who already has formed metalinguistic awareness of how languages work and can engage in problem-solving tasks. So, in situations where a learner is given a sentence with an unfamiliar word or lexical unit, if they had prior knowledge of what the text is about or knows a language within the same language family, they could make guesses at the meaning of the word, despite never having seen that word before. They may also apply grammar rules in one's first language when speaking their second language. Some researchers have suggested that cognitive maturity and metalinguistic awareness can interfere with the learning process of older learners. For example, the English verb *"demand"* and the French word *"demander"* both share a common Latin root, however in contrast to its English meaning, in French *"demander"* means to *"request"* or *"ask a question"* [57].

The path toward gaining a more extensive vocabulary in one's second language is arduous and requires learners to interact with new content continuously. The content can be in the form of movies, music, people, books, and news. Prior works by Krashen [53] suggest that reading for pleasure is one of the best sources of vocabulary growth. However, reading in one's L2 can be daunting for language learners, especially if the text involves a large amount of vocabulary beyond the reader's scope of knowledge. Therefore, to assist learners with vocabulary acquisition and reading comprehension, past work has explored augmenting the text with glosses, which are help texts which contain information on words via definitions or synonyms [81]. Traditional glosses are limited by the boundaries of static paper and most commonly appear at the end of the text as glossaries or as separate paper handouts; however, in comparison, the digital text allows for more dynamic interactions and text augmentations. Digital texts that can be augmented with glosses can span from online articles to video subtitles, and glosses can come in images placed in the sidebar or pop-up windows. In many existing systems, the glosses can be interactive and triggered by a mouse click or static and shown at all times on the screen. Such technology exists as web extensions on personal computers as well as e-readers and tablets.

The acquisition of a new language serves various purposes, ranging from practical reasons such as moving to a new country to personal enrichment [35]. Regardless of the initial motivation, language learners should acknowledge that there is an inseparable relationship between language and culture. Culture gives languages context and gives words broader meanings. Successful language learners recognize that language is not just a tool for communication but also a reflection of a community's values, beliefs, and customs. Thus, learners need to interact with native materials–which are materials authored by and intended for native L2 speakers– as it provides them with a cultural context of the norms and conventions [89]. Interacting with native content also allows the learner to find the content they are interested in, which can motivate them to continue learning despite encountering learning plateaus. Enjoyable reading experiences intrinsically motivate learners and lead to higher reading comprehension [60]. To understand native content, however, learners need to be able to read in their L2.

One of the reasons why learners do not interact directly with content in their L2 is due to a lack of vocabulary knowledge [53]. When confronted with a large block of text, learners may lose motivation to read it even if most of the vocabulary within the text is below their L2 level [54]. Additionally, manually selecting words and phrases within the text to look up in a dictionary on the side is more work and may add to the already high cognitive load learners are experiencing while reading [24]. Therefore, we need a flexible system that implicitly addresses learner needs and shows help when learners make explicit commands. Moreover, there needs to be a more thorough exploration of where to position glosses to not hinder the reading process and aid the learner. This project explores how gaze can be leveraged in an attentive application that takes implicit and explicit input from users to provide help while reading in one's L2.

1.1 Scope and purpose

Reading is a two-step process which requires two sets of skills, the first being **lexical competency** and the second being **phrase decoding** [80]. The first step involving lexical competency deals with understanding the meaning of individual words or phrases and relies heavily on vocabulary knowledge in L2. The second step is one where learners use prior knowledge to understand the meaning of the text and is dependent on the learner's ability to derive sentences from the lexical units and interpret the meaning of the text. There are various methods in which technology has been used to aid the reading process. Tack et al. asserts that there are three main ways technology is used to make reading for L2 learners more comprehensible:

input selection, input modification, and input enhancement [80]. Input selection is when tools filter content that the reader can access and only allow the learner to read texts suitable for their language level. Input modification generally involves translating or simplifying text at word or sentence levels. Lastly, input enhancement refers to applications that help users without modifying the content of the texts.

In this thesis, we focus on input enhancement reading assistance for beginner to intermediate adult learners of French. The choice of building a system to support French learners is due to Canada being an English and French bilingual country, where French courses are offered at every grade level in English-speaking provinces. One of the pervasive issues for French language instructors in primary and secondary education in Canada teaching core French is maintaining motivation among students [5]. As a result, adults who have gone through the Canadian schooling system have taken French courses to a certain level, most adults forget the French they are taught in primary and secondary school [18]. This decreased French knowledge, or language attrition, is often attributed to a lack of practice and interaction with L2 materials [77]. Lightbrown et al. support this fact and state that to maintain basic L2 knowledge, learners must continue learning and interacting with materials in their L2 [57]. Therefore, creating technology that encourages and enables learners to access enjoyable reading material is necessary.

One of the ways to encourage people to continue learning their L2 is to assist people while interacting with materials in their L2 by providing help for words and phrases in the text [80]. This thesis explores which gaze control method people prefer when using a gaze-enabled reading aid and where to place the gloss to increase reading comprehension. We focus our efforts on web-based text enhancement applications, however, the results can be generalized to be used with tablet and mobile device usage. Much of the existing research in second language acquisition (SLA) and computer-assisted language learning (CALL) focuses on the vocabulary gained from using glosses. However, there is a lack of research on the overall comprehension and preference for the presentation of glosses. Additionally, there has not been a recent study on the usability of gaze as a method to control gloss display. This leads to the following research questions, which this thesis addresses:

RQ1: What resources are available for language learning in French, and how can they be leveraged to create informal, beyond-the-classroom, learning materials?

RQ2: Does the placement of computerized glosses affect perceived and measured reading comprehension for L2 learners?

RQ3: What gaze intervention methods are most useful in attentive reading support applications for language learners?

1.2 Overview of the thesis

To address the research questions, we designed a gaze-enabled reading aid with multiple gloss options and tested its functionalities in a participant study. The following two chapters provide background information for this thesis and provide an overview of works in SLA, L2 vocabulary acquisition, and eye tracking within the context of language learning. Due to the use of a non-English language, the selection of off-the-shelf models and APIs was limited, so in Chapter 4, we detail the methods and tools used to annotate the French text and extract complex phrases. Then, in Chapter 5, the system design is described, including how gaze inaccuracy was handled. From pilot studies conducted using system prototypes, we designed a participant study to evaluate the system. In Chapter 6, the design of the participant study is outlined along with its results. Following the results, the next chapter delves into an analysis of evaluation results, how they tie with the research questions of this thesis and possible future research directions. Finally, the last chapter concludes the thesis by overviewing the lessons learned.

2 Second Language Acquisition

Language studies or language education is said to be the "biggest and best known" area of studies within applied linguistics [51]. Second language (L2) refers to the second language learned after one's mother tongue, but it can also refer to any subsequent language learned after childhood. Becoming fluent in one's L2 requires not only vocabulary and grammar knowledge but also other metalinguistic knowledge about the L2 and cultural knowledge of how speakers use the language. The learning process for each individual is different, as there exist individual differences which affect how a person learns. For example, their language learning aptitude, learning style, personality, attitude, and motivation.

2.1 Learning plateaus

Past researchers have framed the process of SLA using different perspectives like behaviourist, innatist, cognitive, and sociocultural [57]. Furthermore, beyond research explaining how people learn languages, there are works looking at how to quantify learner knowledge into levels and how to move from one level to another. Generally, it is accepted that learners are categorized into three levels: beginner, intermediate, and advanced. Moving from one level to the next comes with difficulties, and learners encounter various problems as they advance through their learning journey. For example, learners often experience **learning plateaus**, where they do not feel like they are making progress with their learning. Richards [73] suggests that the perceived plateau is due to mainly five reasons

1. *There is a gap between receptive and productive competence.*
2. *Fluency may have progressed at the expense of complexity.*
3. *Learners have a limited vocabulary range.*
4. *Language production may be adequate but often lacks the characteristics of natural speech.*
5. *There are persistent, fossilized language errors.*

From a language instruction perspective, it is impossible to address all the reasons for the intermediate plateau using a single instruction technique. Individual preferences of learners also limit the types of content people consume and the learning methods used. Formal classroom instruction often will try to broaden the vocabulary knowledge of learners, however, formal instruction alone does not provide enough practice and exposure to the L2. Thus, formal instruction alone is insufficient for overcoming plateaus, so learners need to seek out informal learning methods, such as reading, to supplement their learning [73]. Reading extensively outside of formal language instruction has also been shown to be helpful in increasing learner vocabulary. Studies have shown that an increase in vocabulary is directly linked to an increase in fluency [54, 57, 73, 87], but increased exposure to one's L2 also leads to increased fluency [52]. Additionally, interacting with various types of texts, like news articles and blog posts, increases the learner's vocabulary range and helps connect the gap between knowledge learned in the classroom with native speech patterns [73]. However, learners may not be incentivized to interact with content that they are not interested in. When given reading assignments, learners prefer materials that personalize to their language proficiency and cater to their interests [62]. In summary, extensive reading has been shown to increase vocabulary in one's L2, which could help learners become more fluent. Also, to increase one's motivation to read, we need to provide aid to texts that learners are interested in reading.

2.2 Reading in L2

Reading is a complex skill that requires not only processing and understanding a text's vocabulary but also semantics. As previously mentioned, reading is helpful in increasing fluency in one's L2 and a prominent figure in language learning, Stephen D. Krashen, stated that "[f]ree voluntary reading may be the most powerful tool we have in language education" [54]. Light reading can bridge weaker readers who may feel intimidated by long texts to read more frequently. Therefore having short reading material can help beginners become more comfortable with the act of reading in L2. Reading informal texts written by native speakers can also help learners learn words used in day-to-day settings as 5000 of the most commonly used words make up 95% of conversations and casual reading content (e.g., news articles, blog posts, etc.).

2.2.1 Text modification

Modifying or augmenting text with additional information can be helpful for individuals with poor reading abilities. Prior research has explored how educational tools can support adults with low literacy skills, and they found that providing summarizations, simplifications, elaborations, or explanations can help comprehension [30, 40]. Researchers in this field have found that participants preferred elaborated text with multimedia help, including images and audio. In SLA, studies have examined the effect of text complexity on language learners, including comparing learner attitudes towards varying difficulty, simplified texts, elaborated text, and translated text.

Yano et al. conducted a study with Japanese college students learning English using two types of text modification: simplified and elaborated. They found that although students who read the simplified text performed higher than the group with elaborated text, the results were not significant (*t=1.26, p>.05*). Still, simplification was shown to be better when questions asked for explicit information about the text, while elaboration improved performance on inference type questions [87]. Tweissi investigated the effect of simplification level on learner performance: syntactically simplified, lexically simplified, semi-simplified, and fully simplified [82]. He found that all participants with modified text performed better. However, participants who interacted with the lexically simplified text outperformed all other groups and the lowest performing group within the modified text was the participants with fully simplified text. He concluded that too much simplification can reduce reading comprehension.

Chiang, building from Krashen's Input Hypothesis, which states that "sufficient exposure to comprehensible input is needed for SLA", examined the effect of simplified versus elaborated text on reading comprehension for language learners. They conducted a study to examine learner attitudes towards reading in their L2 at two difficulty levels: one above their current level and one below their current level. Participants given texts below their knowledge level reported more positive attitudes towards reading, while those who read higher level text remained unchanged in their motivation [24]. Other works suggest that increasing vocabulary range is hindered when learners only interact with content at their level or lower [52, 53, 54]. Those works found that the highest learning gain occurs when learners read content slightly higher than their current level and have help available to them. Therefore, although it motivates readers to read simplified texts, it would be more beneficial for learners to read the text slightly higher but have simplifications or help available.

2.2.2 Motivating learners to read

Previously, we've seen how learners generally prefer content catered to them, as it gives them more motivation to read for pleasure. Lungu et al. found in a longitudinal study that providing learners with personalized articles from the internet and vocabulary practice increased the usage of learning applications outside the classroom [62]. Moreover, from their interviews with language instructors after the study, instructors argue that reading aid applications should be used with students with some basic knowledge of their L2 and the aid provided by the system does not need to be perfect for the students to use it. The instructors mentioned that translations provided by the system do not need to be entirely correct because incorrect translations may prompt the learner to do more research on the word themselves actively. The importance of interest is further emphasized by Harackiewicz et al., who argue that promoting interest creates more engagement, motivation, and improves the learning experience for students [42]. Technology has also been used in language learning to enhance the learning experience and attempt to address stagnation in learning by presenting teaching material in different ways. CALL systems allow learners to move along the curriculum at their own pace, and the most popular device used for CALL systems are computers [36].

Curating personalized learning materials for students is a complicated process to do manually, thus, many reading applications work by modifying pre-existing text on the web to use as learning material. Works like Anita [69] and CASSA [72] all leverage online text as L2 learning resources and provide users with simplifications and definitions of words. The next section explores the types of text modification made to texts and how the modification can aid L2 vocabulary acquisition.

2.3 L2 vocabulary acquisition

In relation to reading, the Input Hypothesis suggests that individuals gradually learn vocabulary as they are repeatedly exposed to them. Although Krashen's Input Hypothesis holds to a certain extent, it is based on studies which have methodological problems and ignores factors such as motivation and additional learning outside reading [26]. Extensive reading and reading for pleasure can aid in L2 vocabulary acquisition, but the actual learning gain from purely reading, without external help, is low. There are also various other methods to learning L2 vocabulary, such as rote memorization, keyword technique, and semantic elaboration techniques [76]. The most studied vocabulary method is rote memorization, which involves memorizing the L2 word and the L1 translation. Although it does not require deep cognitive processing, it is the most straightforward and has been proven effective in writing tasks and helpful for intermediate learners [76].

2.3.1 Influence of L1 use on L2 learning

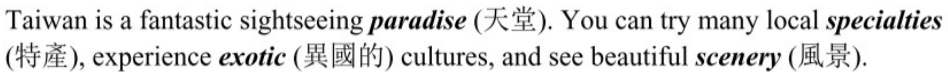
Despite the parallels between vocabulary acquisition in children and adults learning a second language, there is a difference because as children acquire new vocabulary, they also learn about the world around them [57]. Learners learning a second language will often draw upon linguistic and conceptual knowledge from their L1, and they may utilize incorrect word forms or grammar structures originating from their L1. When learners make errors in their L2 due to L1 knowledge, that is called negative language transfer.

Shimabukuro et al. created a visualization tool which shows cross-linguistic errors made by learners of each language using a hierarchical matrix [79]. Their visualization shows that learners make more syntactic and grammar errors dependent on their L1 than lexical errors. Additionally, some research suggests that reliance on their L1 as a bridge between their cognitive process and L2 may hinder their learning progress. However, there are also works which show that the errors made by learners are often unrelated to vocabulary but more due to incorrect collocations [85].

The influence of L1 use during L2 vocabulary learning for adult learners is explored in more detail by Liu in a study with first-year undergraduate students learning English. Liu argues that even in classrooms where instructors shun L1 usage, L1 is inevitably part of an adult learner's learning process since thoughts and connections about the new vocabulary will inevitably connect with their prior knowledge and assumptions about the world [58]. Therefore, providing translation equivalents benefits adult learners, as they are an easy and efficient way of depicting the core meaning of a word.

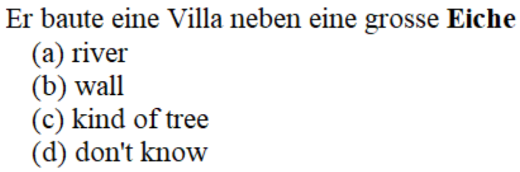
2.3.2 Glossing

Through studies conducted in reading and language learning, it is clear that providing translations or help to learners during reading is valuable. Gloss is a term commonly used in language learning to refer to a brief notation of the meaning of a word or wording in a text. Glossed reading has been shown to increase even delayed posttest scores significantly ($p<.001$) compared to non-glossed reading [86]. Before the popularization of computers, glosses were often placed at the end of texts as glossaries and additional documentation to refer to. Taylor argued for a shift of glosses to be digitalized due to speed and ease of use [81] since online dictionaries did not require users to spend a lot of time looking up words. In addition to exploring the learning gain due to different types of glosses, the effect of gloss placement was examined by prior works.



Taiwan is a fantastic sightseeing ***paradise*** (天堂). You can try many local ***specialties*** (特產), experience ***exotic*** (異國的) cultures, and see beautiful ***scenery*** (風景).

Figure 2.1: In-text gloss example used by Cheng and Good [23].



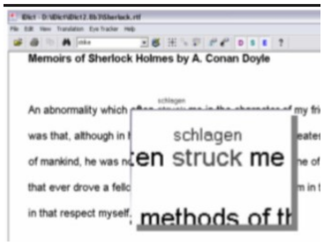
Er baute eine Villa neben eine grosse **Eiche**
(a) river
(b) wall
(c) kind of tree
(d) don't know

Figure 2.2: Multiple choice gloss example used by Rott et al. [75].

Types of glosses

Prior studies have tested various glosses, and the two main categories glosses fall into are either non-interactive or interactive. Non-interactive glosses include glossaries (shown at the end of the text, see Figure 2.4b), marginal (shown in the margins, see Figure 2.4c), interlinear (shown between lines of text, see Figure 2.3), in-text (shown next to the word, see Figure 2.1). These types of glosses existed before the existence of CALL systems, as they can be shown on paper. Interactive glosses require learners to do an action to activate them. An example of this is hyperlinked glosses (see Figure 2.4a) and computerized multiple-choice glosses (see Figure 2.2). Hyperlinked glosses generally appear when users click on the word, and multiple-choice glosses appear in the margin with multiple choices for the definition or translation of the word to prompt a more profound cognitive process about the word. Yanagisawa et al. conducted a meta-analysis looking at all previous studies conducted comparing different types of glosses and found that although all glosses improved word recognition after reading compared to non-glossed text reading (*p<.037*), there are performance differences for each type of gloss [86]. Multiple choice glosses, hyperlinked, marginal, and interlinear glosses led to significantly more significant learning gain in immediate posttest compared to no-gloss conditions (p<.05) and with delayed posttest multiple-choice, hyperlinked, and marginal reached significant improvements [86].

In addition to the different presentation of the glosses, there is also the consideration of which language the gloss should be shown in, i.e., the learner's L1 or L2. Taylor presented previous studies where L2 help is rarely used when L1 help exists. In their meta-analysis, he found that L1 glosses are the most effective mean to assist learners in accessing the most significant amount of text [81]. Although, the downsides to using L1 glosses include a loss of context, pragmatics, and confusion when languages do not have an equivalent phrase.


iDict - D:\iDict\iDict2.0b7\Sherlock.rtf
File Edit View Translation Eye Tracker Help
D S E ?

Memoirs of Sherlock Holmes by A. Conan Doyle

schlagen

An abnormality which my fri
was that, although in t eate
of mankind, he was n ne of
that ever drove a fello m in t
in that respect myself.

schlagen
ten struck me

methods of th

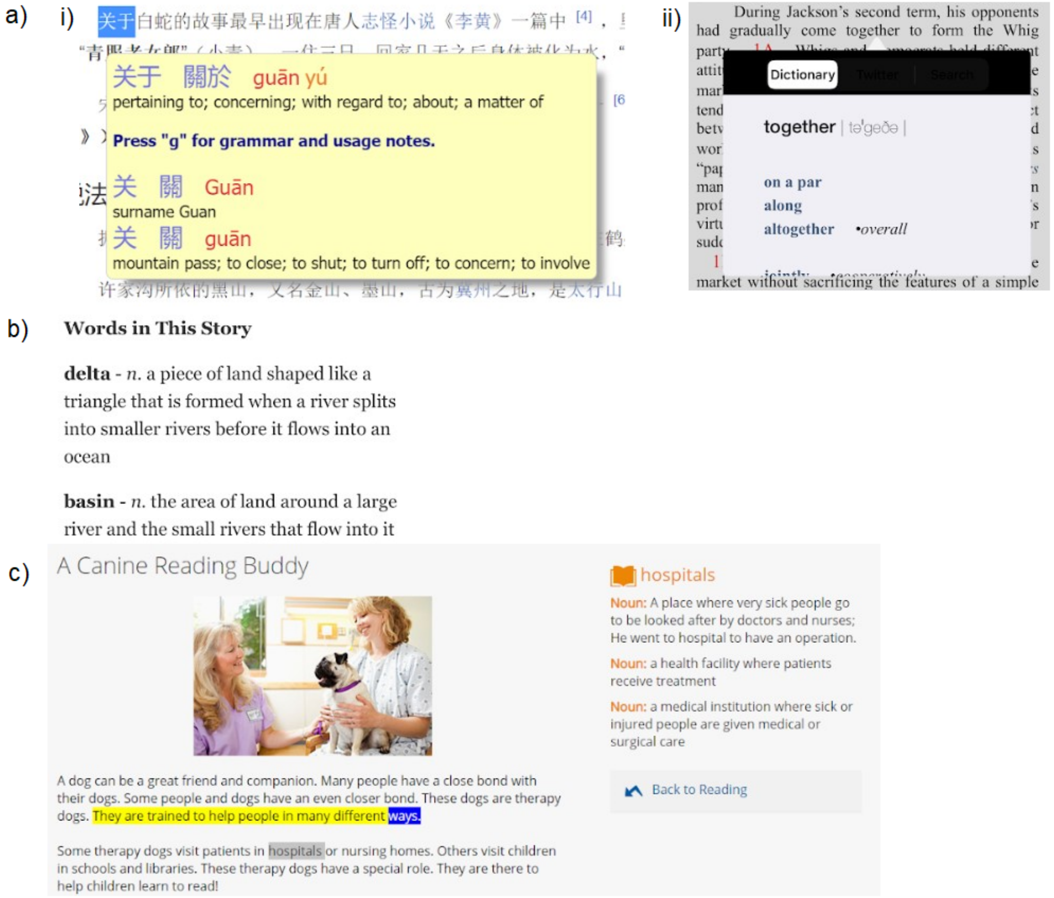
Figure 2.3: Interlinear gloss from iDict [46]. (Used with permission)

Despite the disadvantages of L1 glosses, they are helpful for personal learning, Taylor argues [81]. Cheng and Good conducted a longitudinal study over two weeks that supported L1 glosses by showing non-interactive glosses to students during the reading of English texts and learner performance on delayed tests. They found that L1 gloss with L2 examples and L1 in-text gloss resulted in the highest prolonged retention [23]. Additionally, the learner's language proficiency influenced gloss effects, with participants with higher proficiency performing significantly better on vocabulary recall tests [23].

Computerized glosses

Interactive glosses can be easily implemented on computers, as digital texts can be augmented as one reads. Glosses shown on the computer are often referred to as hypertext glosses, hypertext annotations, hypermedia annotations, or computerized glosses. Many extensions embedded within web browsers and mobile applications offer glosses for words present in the text. Words with glosses are often bolded, highlighted, or italicized since studies show that readers retain vocabulary that is emphasized better [2]. However, some study results suggest that invisible annotations, or not marking words that have annotations, allow users more freedom to consult meanings for specific vocabulary [2]. Overall, there has not been a consensus on what is the preferred method of displaying vocabulary with glosses or any standardization for best practice. For example Figure 2.4 shows systems that mainly invisible annotations, while Figure 2.5 shows Lexi [13] where glosses are highlighted in green.

The type of help provided for texts also varies from system to system. Works like Anita [69], developed by Paetzold et al., provide text simplification and enhancement, including definitions, synonyms, translations, and images to aid learners in reading in English. Glosses can also be used outside the SLA context. For instance, CASSA [72] by Rello et al., a Chrome extension that shows synonyms and definitions, was built to help people with dyslexia and has the potential to be used for language learners. Rello et al. found that the extension was helpful for those with dyslexia. However, their study size was small (n=5), and they did not



关于白蛇的故事最早出现在唐人志怪小说《李黄》一篇中 [4]，

关于 關於 guān yú
pertaining to; concerning; with regard to; about; a matter of

Press "g" for grammar and usage notes.

关 關 Guān
surname Guan
关 關 guān
mountain pass; to close; to shut; to turn off; to concern; to involve

许家沟所依的黑山，又名金山、墨山，古为冀州之地，是太行山

During Jackson's second term, his opponents had gradually come together to form the Whig

together | təˈgeðə |

on a par
along
altogether •*overall*

market without sacrificing the features of a simple

Words in This Story

delta - *n.* a piece of land shaped like a triangle that is formed when a river splits into smaller rivers before it flows into an ocean

basin - *n.* the area of land around a large river and the small rivers that flow into it

A Canine Reading Buddy

A dog can be a great friend and companion. Many people have a close bond with their dogs. Some people and dogs have an even closer bond. These dogs are therapy dogs. They are trained to help people in many different ways.

Some therapy dogs visit patients in hospitals or nursing homes. Others visit children in schools and libraries. These therapy dogs have a special role. They are there to help children learn to read!

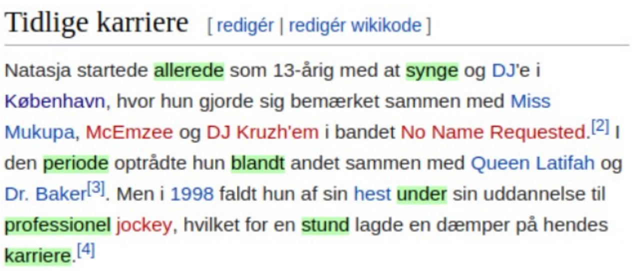
hospitals

Noun: A place where very sick people go to be looked after by doctors and nurses; He went to hospital to have an operation.

Noun: a health facility where patients receive treatment

Noun: a medical institution where sick or injured people are given medical or surgical care

Figure 2.4: Three main types of computerized gloss positioning. (a) **In-text pop-up**, where (i) is from a Chrome Extension called *Zhongwen: Chinese-English Dictionary* and (ii) is from a mobile application *LSATMax*. (b) shows a **glossary** from *TOEFL IBT* and (c) shows **marginal gloss** from *FluencyTutor*. (Fair use)



Tidlige karriere [redigér | redigér wikikode]

Natasja startede allerede som 13-årig med at synge og DJ'e i København, hvor hun gjorde sig bemærket sammen med Miss Mukupa, McEmzee og DJ Kruzh'em i bandet No Name Requested.[2] I den periode optrådte hun blandt andet sammen med Queen Latifah og Dr. Baker[3]. Men i 1998 faldt hun af sin hest under sin uddannelse til professionel jockey, hvilket for en stund lagde en dæmper på hendes karriere.[4]

Figure 2.5: A screenshot from *Lexi*, a mouse click-controlled reading aid created by Bingel et al. uses highlighting to show words that have glosses [69]. (Used with permission)

report in-depth findings. Robertson creates visual dividers for Finnish expressions. Finnish is an agglutinative language, meaning multi-word expressions usually form sentences, thus it is helpful to separate the different parts of an expression to highlight the different parts and distinguish the headword from the rest. Text 2.0 by Biedert et al., mentioned in more detail later, uses a similar technique and breaks up compound words into their morphemes, prefixes, and suffixes [12].

Additionally, studies have also been conducted to examine the optimal position of computerized glosses and found that most students prefer help to be displayed close to the word that is being looked up [1, 2, 21, 22], examples of each type of gloss can be seen in Figure 2.4. AbuSeileek first conducted a study with English L2 learners. He separated participants into five groups: no gloss, right-hand side margin gloss, bottom margin gloss, gloss pop-up on the right-hand side, and glossary. From this study, he found right side margin and bottom margin groups both outperformed the other groups [2]. Following this, AbuSeileek conducted another participant study with five new groups, this time comparing in-text gloss, bottom margin, right margin, a pop-up window in the margin, and no gloss. The best performing group was the in-text gloss, followed by the right margin and pop-up window, which follows his initial closeness theory [1]. Chen and Yen also investigated gloss positions for English L2 learners by comparing in-text, pop-up, and glossary, where the pop-up would appear near the vocabulary in question. Their participant study separated users into four groups and ran over six weeks, which involved a two-hour session per week. In their findings, the pop-up format received the highest reading comprehension scores, followed by glossary conditions, no annotations, and in-text [22]. Although prior studies show the benefit of each gloss position quantitatively through comprehension scores and vocabulary posttest, their studies were conducted using between-group comparisons, meaning each condition group could not test other systems and thus cannot compare the performance of other gloss positions. Furthermore, there lacks a quantitative study of gloss position based on participant preferences.

3 Eye Tracking for Language Learning

Eye tracking is a popular research tool in medicine and psychology, as it allows researchers to study human gaze behaviour and patterns. Data collected from eye movement can be categorized into two types of movements: fixations and saccades. Fixations are pauses, usually ranging from 100 ms to 600 ms, in the eye movement on a specific area of the visual field and occur in the eye's foveal area. Saccades are rapid eye movement from one fixation point to another and are used to piece together visual scenes. Currently available commercially available eye tracking technology typically only looks at foveal area and ignores the parafoveal and peripheral area [45]. Although the foveal area accounts for visual acuity, it only makes up less than 8% of the visual field [10]. Eye movements have been studied in various different contexts, including predicting native language using gaze movement [11, 55], predicting reading comprehension from gaze behaviour [3], and differentiating learner proficiency using word recognition tasks [14]. Much of the current research on gaze movement related to language learning is done with learners with English as their L2, and there are studies conducted to show multilingual gaze movement [55]; however, they are in the minority.

3.1 Gaze during reading

Within literature published using eye tracking in the linguistic field, the main areas of study include auditory processing (listening), visual word processing (reading), and simultaneous auditory and visual processing [27]. In this thesis, we focus on the visual word processing aspect of eye tracking. Reading is a task that involves extracting and processing visual information to decode what is contained within the text, and the foundation of skilled reading involves quick and efficient processing. During reading, the window of perceived text is asymmetrical around a fixation point and generally extends 3–4 characters to the left and 14–15 characters to the right for left-to-right reading [10]. When reading text in L1, people will often not fixate on words that are 2–3 letters long and generally, complex and less frequently seen words are fixated longer than simple and commonly seen words [71]. One of the pioneer psycholinguists who worked extensively with eye movement during reading was Keith

Rayner. Rayner found that experienced readers do not look at every word in the text, and the three significant factors that affect reading include frequency, length, and predictability [25]. Frequency refers to readers spending less time processing words that occur more frequently in the language. The length of words affects the processing time, so longer words typically required a longer processing time. Finally, more predictable words are skipped more, fixated less, and refixated during regressions less often. Regressions refer to backtracking during reading.

Since gaze can be used to estimate the skill level of a reader [14, 66], pairing it with knowledge of gaze movement during reading can show us information about text difficulty, if we have prior knowledge of the reader's skill level [37]. Additionally, due to personal differences during reading, eye movements can be used to create personalized reading help tools. For instance, Jiang et al. explored how to use gaze data and facial expressions captured during webcams to create personalized summaries of texts for readers [49]. Other than using webcams and stationary eye trackers, some research investigated the use of wearable eye-tracking devices during reading [16].

3.2 Gaze in language learning contexts

Gaze features used in language learning contexts are similar to those used in reading since detecting difficulty during reading can apply to individuals with low literacy skills and language learners. Moreover, eye tracking provides a "rich moment-to-moment data source" [27], which is helpful for SLA research. Within language learning, eye tracking data has been used to examine vocabulary processing and learning, listening comprehension, syntactic processing, written text production, reading comprehension, text-based computer-mediated communication, oral communication, and data validation [56]. For example, some prior research has examined the performance of modifying text presentation to reflect the reader's cognitive state [48] and the variation between how learners look at video captions based on their L2 [84]. In this thesis, we focus on past works on L2 proficiency, detecting unknown words, and applications that provide real-time help for learners.

3.2.1 Estimating L2 proficiency

Multiple prior studies have focused on estimating leaner L2 proficiency using gaze movement. Most of the studies conducted used stationary eye trackers mounted at the bottom of a monitor [6, 16, 50, 65, 88], while one gathered eye tracking data using wearable devices [7]. All the studies for estimating proficiency involve post-study analysis of obtained gaze data, and later sections will cover real-time gaze-responsive applications. To estimate L2 proficiency, Bulling et al. found that fixation and blink duration are effective indicators for L2 proficiency, specifically, lower proficient readers have longer average fixation time than proficient readers [16]. Martínez-Gómez and Aizawa worked with SVM (support vector machines) and random forest-based models to categorize participants into low and high level understanding [65]. From their experiment, they found the most influential features in predicting English skill and level of understanding include reading time, saccade median length, fixation mean acceleration, and fixation average.

Similarly, Karolus et al. ran a study with English L2 and German L1 participants and found that lower proficient readers have longer average fixation than proficient readers [50]. Yoshimura et al. used gaze features to categorize learners' English skills into discrete levels. Using the participant's TOEIC scores as their L2 level, they experimented with estimating the participant's L2 skills using an SVM classifier trained using total fixation duration and total saccade velocity. TOEIC, or Test of English for International Communication, is an international standardized test of English language proficiency for non-native speakers. It is intentionally designed to measure the everyday English skills of people working in an international environment. They found that the model could accurately categorize learners into three categories (low, middle, high) with up to 90.9% accuracy, and the accuracy falls to below 60% with any additional categories [88].

Following this, Augereau et al. attempted to predict the learners' TOEIC score–which could range from 10 to 990–based on eye movement during reading, with a fixed position tracker. They found that total reading time, saccade velocity, saccade count, blink count, scanpath length, and maximum saccade velocity were the most important features when used in a machine learning model to predict TOEIC scores within 14.4 points from the true value on average [6]. Augereau et al. also used mobile eye trackers to predict TOEIC scores, and the main features used were average fixation duration, total number of fixations, sum of duration, average saccade length, average saccade speed, and average saccade time [7]. They were able to estimate TOEIC scores to within 36.3 points of the true value on average.

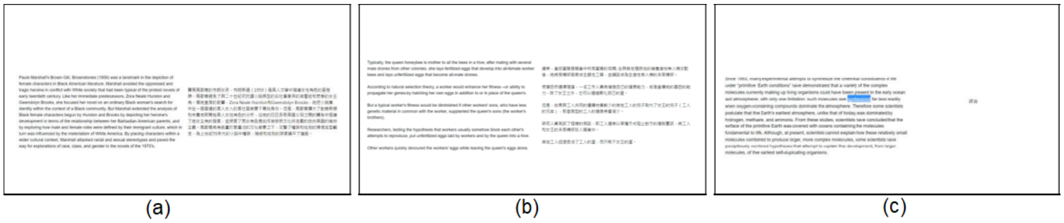
3.2.2 Detecting unknown words

Detecting when readers are experiencing difficulty while reading using machine learning has been a topic of research for some more recent publications. For example, Gedeon et al. investigated using neural networks and fuzzy output error to predict reading comprehension scores from gaze data [34]. Reading comprehension of the entire article sometimes first lies in understanding individual vocabulary in the text. As prior research in SLA has shown, vocabulary knowledge is one of the most important aspects of L2 fluency [54, 57, 73, 87] and eye movement can give insight into cognitive processing, gaze can be used to detect when a learner is having difficulty, and interventions could be made. Garain et al. attempted to categorize eye movement features that can be used to detect reader-specific difficult words using machine learning-based classifiers like SVM, RNN (recurrent neural network), and random forest [33]. To improve eye-tracking accuracy, they added a line id to each line of text to improve vertical eye-tracking error. Similarly, Hiraoka et al. used SVM with RBF (radial basis function) kernel with gaze features as input, but their results did not lead to large improvements in accuracy or F-measures [43]. Gaze features that they found had a high correlation with word-level difficulty are word rarity, word length, max gaze duration, first gaze duration, and total gaze duration. Some researchers have noted the lack of accessibility of commercial eye trackers and proposed utilizing webcams as a source of gaze data [28]. Ding et al. used a combination of gaze data extracted from webcam footage and linguistic information about the text to extract unknown words using an NLP model. Their model uses multiple LSTM layers [78], RoBERTa [59], and large amounts of data passed into a binary classifier to determine whether the word is known or unknown, making it unsuitable for real-time use.

3.2.3 Gaze-enabled applications

Gaze-enabled applications refer to applications which use real-time gaze movement either to explicitly control functions in the application or implicitly inform the system about user behaviour. An example of an implicit gaze-enabled system is the recommender system created by Augereau et al., which follows their other work surrounding unknown word detection for English L2 learners. Their system detects when users come in contact with an unknown word and stores it within a dictionary, which is then used to find other articles with words they have struggled with [8]. Their work uses the pedometer metaphor and applies it to text, aiming to help users keep track of words and the number of words they have read. Other works have also explored implementing glosses with text in multimedia. Fujii et al. utilized movie subtitles and eye tracking to determine the English level of a learner. Their application, SubMe, can estimate their comprehension of the movie and then use the estimated skill level to implement pauses during video watching to provide a gloss of some words on top of the subtitles.

Gaze-enabled applications that are relevant to this thesis and utilize explicit commands augment text while reading based on gaze information [12, 44, 46, 68]. One of the first gaze-enabled systems used for reading in English is iDict, created by Hyrskykari et al. [47] which



Paule Marshall's Brown Girl, Brownstones (1959) was a landmark in the depiction of female characters in Black American literature. Marshall avoided the oppressed and tragic heroine in conflict with White society that had been typical of the protest novels of early twentieth century. Like her immediate predecessors, Zora Neale Hurston and Gwendolyn Brooks, she focused her novel on an ordinary Black woman's search for identity within the context of a Black community. But Marshall extended the analysis of Black female characters begun by Hurston and Brooks by depicting her heroine's development in terms of the relationship between her Barbadian American parents, and by exploring how male and female roles were defined by their immigrant culture, which in turn was influenced by the materialism of White America. By placing characters within a wider cultural context, Marshall attacked racial and sexual stereotypes and paved the way for explorations of race, class, and gender in the novels of the 1970's.

[illegible]

Typically, the queen honeybee is mother to all the bees in a hive; after mating with several male drones from other colonies, she lays fertilized eggs that develop into all-female worker bees and lays unfertilized eggs that become all-male drones.

[illegible]

According to natural selection theory, a worker would enhance her fitness --or ability to propagate her genes-by hatching her own eggs in addition to or in place of the queen's.

[illegible]

But a typical worker's fitness would be diminished if other workers' sons, who have less genetic material in common with the worker, supplanted the queen's sons (the worker's brothers).

[illegible]

Researchers, testing the hypothesis that workers usually somehow block each other's attempts to reproduce, put unfertilized eggs laid by workers and by the queen into a hive.

[illegible]

Other workers quickly devoured the workers' eggs while leaving the queen's eggs alone.

[illegible]

Since 1953, many experimental attempts to synthesize the chemical constituents of life under "primitive Earth conditions" have demonstrated that a variety of the complex molecules currently making up living organisms could have been present in the early ocean and atmosphere, with only one limitation: such molecules are synthesized far less readily when oxygen-containing compounds dominate the atmosphere. Therefore some scientists postulate that the Earth's earliest atmosphere, unlike that of today, was dominated by hydrogen, methane, and ammonia. From these studies, scientists have concluded that the surface of the primitive Earth was covered with oceans containing the molecules fundamental to life. Although, at present, scientists cannot explain how these relatively small molecules combined to produce larger, more complex molecules, some scientists have precipitously ventured hypotheses that attempt to explain the development, from larger molecules, of the earliest self-duplicating organisms.

[illegible]

Figure 3.1: Interface used by Ho et al. to examine the difference in learner performance due to margin showing (a) paragraph-level translations, (b) sentence-level translations, and (c) word-level translations [44]. (Used with permission)

provided L1 glosses in various languages to words while reading. Hyrskykari ran exploratory studies comparing the performance of mouse click-based systems and gaze duration based for triggering glosses in the text. She found that similar amounts of words were correctly triggered (∼86%) in the gaze and mouse-only conditions, with less than 1% of gloss triggered being false positives and the only system which led to a lower percentage of true positive triggers was the combined version which used both mouse and gaze for triggering [46]. L1 glosses were shown interlinearly above the word, and glosses containing translations with definitions were shown on the right side margin. Text 2.0 [12] by Biedert took inspiration from Hyrskykari and developed an application that provides gaze-activated interlinear L1 glosses, bookmarking, word breakdown (similar to NiinMikäOli?! [74]), smart footnotes for explanations, and a quick skim feature.

Other researchers who worked with text augmentation focused on more specific functionalities. Ho et al. compared the use of word level, sentence level, and paragraph level translations in the right margin triggered by gaze movement, the application interface can be seen in Figure 3.1. Due to gaze accuracy constraints, they did not implement word-level translations during their study. They found that there were no statistically significant differences between each level of help, however, sentence level mapping led to higher learner-perceived performance and understanding [44]. Okoso annotated difficult words in the text and tracked the reading progress using the gaze movement of learners, their choice of tracking reading progress at paragraph and sentence level was due to inaccuracy of eye tracking [68]. Their reading aid focused on providing feedback on reading progress and showing how well a learner has read by colouring each paragraph in the text with a different colour, darker colours meant slow reading speed and light colours meant faster reading speed.

4 Text Processing

Most of the existing research in NLP centers around English, and amongst those on language learning, many focus on learning English as an L2. There does exist literature on learning other languages like French, Spanish, and Portuguese; however, research in language learning research in for non-English accounts for less than 30% of published literature [80]. As such, there is a lack of choices within readily available resources that can be used for the purposes of this thesis. In this section, we describe the resources used to gather and annotate French text necessary to test a reading aid application. In this chapter, we address research question 1 (**RQ1**) and show the different types of resources available and how to use them to analyze French text easily for language learning applications. We tested available resources that can easily integrate with app development can help future researchers build a French text processing pipeline to annotate online text automatically and estimate its complexity in real time. A large source of online reading material are in the form of news articles, we looked for texts that were recently written and contain jargon commonly seen on popular news sites. In this thesis, we compiled open-source French articles to annotate and evaluated them using available French text complexity models. Furthermore, we show how we annotated the text for our purpose of creating a reading aid that provides glosses for words and phrases.

Proficient user	C2	Mastery
	C1	Advanced
Independent user	B2	Vantage
	B1	Threshold
Basic user	A2	Waystage
	A1	Breakthrough

Table 4.1: CEFR levels to categorize learner's L2 proficiency.

CEFR	**NCLC**
C2	12
C1	11
	10*
B2	9
	8
B1	7
	6
	5
A2	4
	3
A1	2
pre-A1	1

Table 4.2: Comparison between the NCLC and CEFR levels.

4.1 French resources

In much of the literature surrounding language learning, there has been a lack of standards used to measure the level of learners. With ESL, there are well-known and commonly used English as an L2 standardized tests like the TOEFL that can provide a generalized level for the learner, however with other languages, although standardized tests exist, because the level needed to achieve sufficient mastery of the L2 varies, the research conducted is hard to generalize across languages. Within Canada, there exists a French proficiency scale called niveaux de compétence linguistique canadiens (NCLC) or Canadian Language Benchmark (CLB) in English. This scale has a total of 12 distinct levels separated into three stages beginner (1-4), intermediate (5-8), and advanced (9-12) [19].

Many of the recently published papers use the Common European Framework of Reference (CEFR) to measure the language level across different languages [50, 56, 80, 83, 86]. Therefore, to align this thesis with other commonly referenced work within language learning, we will be measuring learner proficiency in alignment with the CEFR. The CEFR level separates L2 learners into 6 levels, A1 to C2, which can be regrouped into three broader levels: basic user, independent user, and proficient user. Each level is further separated into two, see Table 4.1, with a level given to 4 different dimensions: listening, reading, speaking, and writing [29]. Refer to Table 4.2, for a comparison between CEFR and NCLC levels.

To obtain appropriate texts for our application, we combed through the websites provided, various language learning applications (both web-based and mobile), available corpora, and

*For NCLC level 10, proficiency at this level can be considered partway between CEFR B2 and C1.

websites with open-source text to find appropriate text. After obtaining the text, we worked on finding appropriate datasets containing French words with their corresponding CEFR levels. The final part of the necessary data needed to complete text processing was an API that could give us the definition, examples, synonyms, and translations of the words we want to create glosses for. Additionally, our application also wants to provide glosses for phrases and idioms as they should be treated as one unit, therefore, we had to decide on the appropriate translation API.

4.1.1 Search for appropriate text

With any language instruction, one must be aware of language variation between different countries and communities who speak the language. We are working with French learners within the Canadian context, so we hoped to obtain materials written for Canadian French. We started by contacting the Centre for Canadian Language Benchmarks and other organizations that provide Canadian French education to new adult immigrants in Canada and Canadian French learners. Here is a summary of the received recommendation from the Centre for Canadian Language Benchmarks for online and textbook resources:

- **Tutela**[1] *(Canadian French)* web resource funded by the government of Canada with vocabulary practice, lesson outlines, and practice worksheet sorted by NCLC/CLB levels for teaching adult newcomers focusing on teaching English and French as a second language to professionals across Canada.

- **Par Ici**[2] *(Canadian French)* exercise book and instructional guidebook developed by Quebec teachers for self-paced learners containing listening, speaking, reading, and writing exercises.

- **Le Point du FLE**[3] *(France French)* website with CEFR level categorized practices with links to listening comprehension and reading exercises using texts in poems, song lyrics, and literature.

- **Coerll by University of Texas**[4] *(Various Languages)* open educational resources and practices for various languages, including French. However, the site has not been maintained and many links for resources lead to unavailable sites.

[1] https://www.tutela.ca/PublicHomePage

[2] https://methode-parici.com/

[3] https://www.lepointdufle.net

[4] https://www.coerll.utexas.edu/coerll

Article No.	Topic (FR)	Topic (EN)	Word Count
1	Pas de l'ours	Grizzly Bear (dance)	543
2	Carnaval de Québec	Quebec Winter Carnival	545
3	Courer indien	Indian runner duck	542
4	Mont Saint-Michel	Mont Sait-Michel	532
5	Dada	Dada	536
6	La Dernière Licorne	The Last Unicorn (film)	538
7	Lionel Messi	Lionel Messi	542
8	Crise économique asiatique	Asian financial crisis	539
9	Théodore Ier Lascaris	Theodore I Laskaris	541
10	Plante carnivore	Carnivorous plants	534
11	Maewo	Maewo	545
12	Chatbot	Chatbot	554

Table 4.3: Articles chosen from French Wikipedia and their total word count.

Although these resources are useful, most of the texts available on the sites were unavailable, too short, dated, or copyrighted. We looked to popular online news sources in Quebec like Montreal Gazette, Le Journal de Montréal, Le Journal de Québec, La Presse, and Le Devoir for articles to use. Due to legal reason, we chose to pursue other sources that were explicitly open source. Newsela[5], could have been another potential source for news articles labelled with corresponding CEFR levels, however, they only have corpora containing English and Spanish texts. Project Gutenberg[6] was considered, but the texts available in the corpus consist of works with expired copyrights and do not suit the purposes of this project. In the end, we opted to use French Wikipedia and chose 15 articles ranging from various topics as texts used to test the reading aid application we built. The texts we chose were also shortened to be similar length and take about 5-10 minutes to read. A word count of the twelve chosen articles and their topic can be seen in Table 4.3. From now on, articles and texts are used interchangeably to refer to Wikipedia articles.

4.1.2 French corpus with CEFR levels

From prior research in complex word identification (CWI) we know that lexical complexity is subjective and from language learning research [38], there are many factors which can affect a learner's vocabulary knowledge. Although we do not aim to summarize the experiences of every learner, we hoped to better determine a lexical unit's complexity by using cor-

[5] https://newsela.com/

[6] https://www.gutenberg.org/browse/languages/fr

pora annotated by non-native speakers or corpora derived from French language learning material. Since we hoped to provide glosses for whole lexical units, including both phrases and named entities, we hoped to find a dataset that annotated the complexity of words and phrases. However, among the available datasets, none fit our needs perfectly, so we had to rely on a conglomeration of different tools to provide the needed translation and complexity evaluation.

We opted to use FLELex [32], where each entry in the FLELex dataset contains a lemma (i.e., word), POS-tag, frequency in CEFR level, and total frequency in the source text. The frequency in the CEFR level is approximated using French as a foreign language teaching material. Pintard and François describe methods to combine expert knowledge with word frequency information to infer the CEFR levels of words for French learners using different machine learning algorithms like tree classification, SVM, and boosting [70]. The version we use is the FLELex with CRF Tagger parts of speech, which contains 17,871 entries and supports multi-word expressions. This version provides us with the best approximation of complexity for words and short phrases, however, it lacks support for longer phrases. From the text, more than 85.3% of the extracted and annotated words were singular words, while the remaining 14.7% were phrases containing one or more words. We will describe how we handle the complexity of longer phrases and named entities in future sections.

4.1.3 French text parsing

POS cataloguing, or POS tagging, is an important part of the NLP pipeline for text processing. POS tagging involves allocating the POS tag to every word in a piece of text. There are various methods for determining the POS tag of a word in a sentence. We tested two methods for POS tagging in Python: `nltk` with StandfordPOSTagger[7] and `spaCy` with a Lemmatizer trained using `fr_core_news_lg`[8]. Through testing, we found that the tagging abilities of both models performed similarly, but `spaCy` integrated with the entire text annotation process better, so it was selected for use.

4.1.4 French dictionary API

The adoption of reading through screen interfaces allows for learners to interact with the text in additional ways, like looking up the meaning of a word using online material and resources. Within language learning, online dictionaries can not only provide the definition of

[7] https://nlp.stanford.edu/software/tagger.shtml#About

[8] https://spacy.io/models/fr

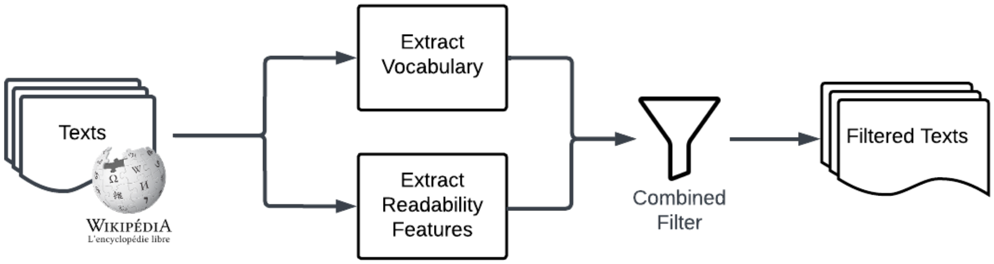


Figure 4.1: Pipeline used to process the text to filter out articles that contain words that overlap with other articles and are too complex or too simple according to readability scores.

the word in one's L2 but also translations of the word in L1. Many different online dictionaries offer access to its large source of data, and many at a cost. For example, Le Robert[9] which Google uses as a primary source for definitions for French words can be accessed through a monthly subscription. There exist large open source dictionaries like Wikitionnaire[10] which provide definitions and have API support which allows users to easily extract information from the site itself. For French, there are over 1.9 million entries, however, although most entries containing more than one definition the quality of definitions vary. Additionally, it does not provide translations for the words or phrases looked up, so an additional translation API, like Google's `googletrans` API, needs to be incorporated. Another option is using APIs like Lexicala[11], built by K Dictionaries, a multilingual dictionary that provides definitions, synonyms, examples, and vocabulary translations. One limitation of using Lexicala and similar APIs is the lack of support for multi-word and named entities, so an additional translation API must be applied for missing lexical units.

4.2 Processing the text

To select and process the text into a format that can be used in the reading aid, we had to first manually shorten the text to the appropriate length so that the content of the text remains cohesive. Following, we preprocessed the text so that words, phrases, and named entities can be extracted. The words and phrases were extracted to append other necessary information to form a gloss, i.e. definitions and translations. In order to properly assess the differences in performance of gloss positions, we need to have text that is similarly levelled yet

[9] https://dictionnaire.lerobert.com/
[10] https://fr.wiktionary.org/wiki/Wiktionnaire
[11] https://lexicala.com/k-dictionaries/

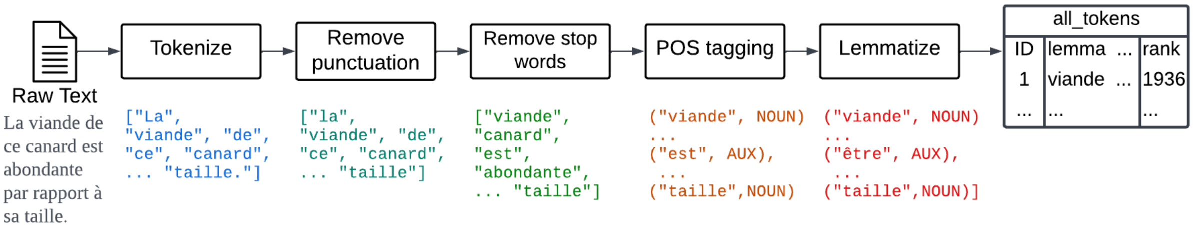

Raw Text
La viande de ce canard est abondante par rapport à sa taille.
Tokenize
["La", "viande", "de", "ce", "canard", ... "taille."]
Remove punctuation
["la", "viande", "de", "ce", "canard", ... "taille"]
Remove stop words
["viande", "canard", "est", "abondante", ... "taille"]
POS tagging
("viande", NOUN) ... ("est", AUX), ... ("taille",NOUN)
Lemmatize
("viande", NOUN) ... ("être", AUX), ... ("taille",NOUN)]
all_tokens
ID
lemma
...
rank
1
viande
...
1936
...
...
...

Figure 4.2: Text preprocessing steps taken to extract words and formulate a dataset containing all the words, POS tags, and word frequency ranking of single word vocabulary in the text.

contains different content. Therefore, after the extraction of words and phrases, we compared vocabulary range, level, readability, and other features of the text to filter out the texts that differed from others. The overall pipeline used to process the text for our application is shown in Figure 4.1 and details of each step are described in the following section.

4.2.1 Preprocessing

Preprocessing text is an important part of an NLP pipeline, algorithms, and large machine-learning models need text to be parsed and labelled before being used as inputs. Similarly, for a reading aid application, the text needs to be preprocessed to extract and label the lexical units we want to provide glosses for. Text preprocessing techniques vary depending on the types of text that need to be processed, for example parsing through comments on social media will be different from parsing through conference proceedings. There are commonalities between all preprocessing techniques, including the previously mentioned POS tagging, and we will detail how we preprocessed the text in this section. The preprocessing process we follow can be seen in Figure 4.2.

In this project, we used a preprocessing step to extract singular words from the text to evaluate and compare the various articles, phrases and named entity extraction followed a different pipeline which will be covered in depth in later sections. The first step to preprocess the text is to tokenize it, this involves splitting the words in the text by space and creating a vector of tokens containing singular or hyphenated words. The tokens were then set to lowercase, and punctuation was removed, including the typical English ones provided by `string.punctuation` in Python and additional ones present in the French text we utilized «»–…""-.. Following the punctuation removal, stop words were removed using spacy's built-in French list, `fr_stop`. We also appended the following words to the list for removal of non-essential words: *-vous, -là, -t, xix, xv, xvii, xx, xxi, vi, s, xviii, xi, xiii*. The tokens left

word	tag	freq_a1	freq_a2	freq_b1	freq_b2	freq_c1	freq_c2	freq_total	rank
ainsi	ADV	42.53	249.17	543.04	436.88	524.97	1173.33	400.78	210

Table 4.4: Sample entry from FLELex CRF Tagger corpus with ranking appended.

after removing the stop words were then tagged with their POS tags and lemmatized, both processed using spacy's built-in tagger and lemmatizer.

The last step to extract singular vocabulary from each article text was to annotate each word with its word frequency rank. This was accomplished by first sorting the FLELex dataset by descending total word frequency (freq_total) and appending an additional column to contain a frequency rank (rank), a sample of the appended corpus can be seen in Table 4.4. Each entry contains a word, POS tag associated with it, its frequency across different CEFR levelled texts, and its total frequency in French text overall.

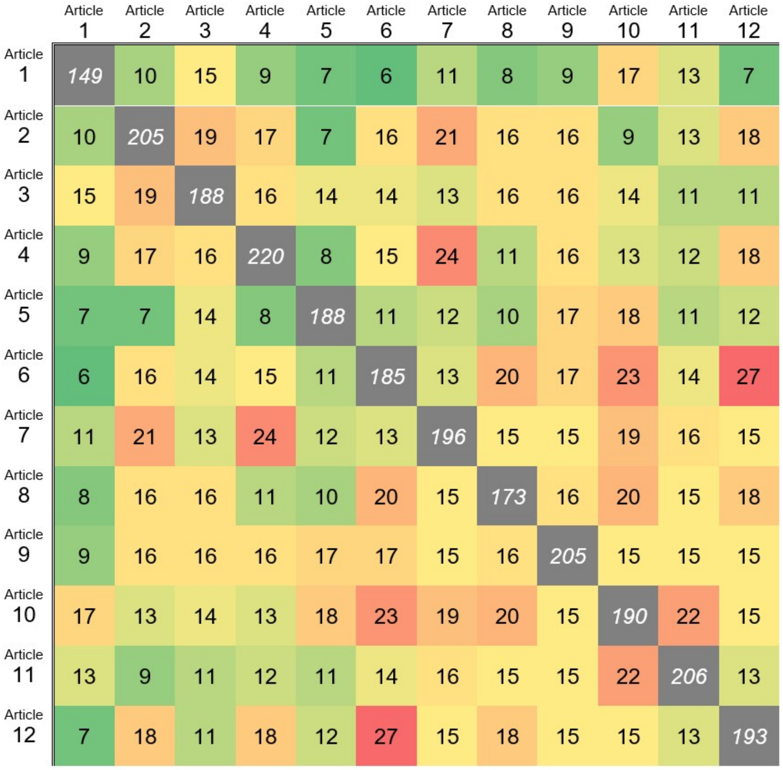
4.2.2 Filtering the texts

Several features of the texts were considered when filtering out articles that would be inappropriate for our application testing. The basic word, or lexical, level features that we considered included vocabulary overlap with other texts and the ratio of vocabulary CEFR levels in each text. For readability, we followed the suggestion derived from Wilkens et al. and evaluated the text using FABRA, a French Aggregator-Based Readability Assessment toolkit [83]. FABRA evaluates text complexity over several dimensions: length, syntactic, and lexical-based features.

Lexical differences between the texts

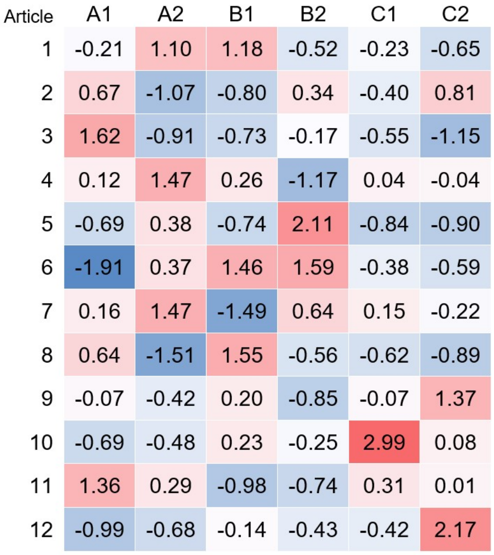
Using the words extracted from the preprocessing step, we were able to further filter and separate words that would be useful to evaluate the complexity of each article. Since the goal of extracting words is to compare the overlap between different articles, we filtered out words which had a frequency ranking of less than 200. The same step was repeated for each article, resulting in 12 lists of words, one for each text. Next, each list of words was compared with the other articles and the overlap matrix, which can be seen in Figure 4.3. The number at the intersection refers to the total number of words which appear in both texts.

Using the CERF level prediction model built by Pintard and François, we approximated the number of words in each text belonging to the six different CEFR levels. The values were made into ratios by dividing the total number in each level by the total number of words in



	Article 1	Article 2	Article 3	Article 4	Article 5	Article 6	Article 7	Article 8	Article 9	Article 10	Article 11	Article 12
Article 1	*149*	10	15	9	7	6	11	8	9	17	13	7
Article 2	10	*205*	19	17	7	16	21	16	16	9	13	18
Article 3	15	19	*188*	16	14	14	13	16	16	14	11	11
Article 4	9	17	16	*220*	8	15	24	11	16	13	12	18
Article 5	7	7	14	8	*188*	11	12	10	17	18	11	12
Article 6	6	16	14	15	11	*185*	13	20	17	23	14	27
Article 7	11	21	13	24	12	13	*196*	15	15	19	16	15
Article 8	8	16	16	11	10	20	15	*173*	16	20	15	18
Article 9	9	16	16	16	17	17	15	16	*205*	15	15	15
Article 10	17	13	14	13	18	23	19	20	15	*190*	22	15
Article 11	13	9	11	12	11	14	16	15	15	22	*206*	13
Article 12	7	18	11	18	12	27	15	18	15	15	13	*193*

Figure 4.3: Matrix showing the number of overlapping words appearing in two texts compared to other shortened article texts.



Article	A1	A2	B1	B2	C1	C2
1	-0.21	1.10	1.18	-0.52	-0.23	-0.65
2	0.67	-1.07	-0.80	0.34	-0.40	0.81
3	1.62	-0.91	-0.73	-0.17	-0.55	-1.15
4	0.12	1.47	0.26	-1.17	0.04	-0.04
5	-0.69	0.38	-0.74	2.11	-0.84	-0.90
6	-1.91	0.37	1.46	1.59	-0.38	-0.59
7	0.16	1.47	-1.49	0.64	0.15	-0.22
8	0.64	-1.51	1.55	-0.56	-0.62	-0.89
9	-0.07	-0.42	0.20	-0.85	-0.07	1.37
10	-0.69	-0.48	0.23	-0.25	2.99	0.08
11	1.36	0.29	-0.98	-0.74	0.31	0.01
12	-0.99	-0.68	-0.14	-0.43	-0.42	2.17

Figure 4.4: Matrix with z-scores coloured using a diverging colour scale to show the differences in the average number of words belonging to a CEFR level.

Variable	Category	Description
LENsntWRD	Length	Number of tokens per sentence, including punctuation
SYNposDET	Syntactic	Number of different POS types in the text, following universal guidelines
SYNdevNPHRS	Syntactic	Number of constituents
SYNdevHGT	Syntactic	Deepness of sentences in the text
SYNdevSUB	Syntactic	Number of words directly subordinate in the dependency tree
LEXdvrWLC	Lexical	CTTR of all words in the text in lemma form
LEXdvrVSU	Lexical	UberIndex of verbs in the text in word form

Table 4.5: Linguistic features used for determining text readability for L2 French learners.

the text to obtain a ratio of each level over an article. For example, the equations used to calculate the ratio of words at the A1 level are as follows,

$$\hat{n}_{i,A1} = \frac{n_{i,A1}}{n_{i,unknown} + \sum_{j=A1}^{C2} n_{i,j}} \tag{4.1}$$

where *i* refers to the article number and $n_{i,unknown}$ refers to the number of words not present in the FLELex dictionary. Z-scores for each ratio were then calculated to find the divergence from the average calculated using all the texts and calculated using the equation

$$z_{i,level} = \frac{\hat{n}_{i,level} - \mu_{level}}{\delta_{level}} \tag{4.2}$$

where μ_{level} is the average of the normalized ratio of each CEFR level per article and δ_{level} is the standard deviation for each CEFR level. Figure 4.4 shows a colour-coded matrix containing the z-scores of the ratios of words belonging to a CEFR level for each article. One thing to note for the CEFR levels assumed in this step is that the accuracy of an automatic machine learning-based estimation is between 50% to 70% when compared to expert labels [70]. Thus, further testing must be done on the test to ensure that the texts are similar to readability.

Text complexity

In the previous section, we looked at word overlap and the overall ratio of complex words to not complex words, however, these measures were to compare the chosen texts within the smaller sample of text that we collected. In addition to basic lexical features like word overlap and word complexity, there are various different features that affect the readability

i	1	2	3	4	5	6	7	8	9	10	11	12
d_i	1.277	0.957	0.989	1.069	0.611	1.170	1.159	1.645	1.285	1.753	1.280	2.236

Table 4.6: Euclidean norm of complexity variables for each article.

or complexity of text. Generally, the complexity of a text is measured by its syntactic and lexical complexity. Syntactic complexity refers to the difficulty of sentences in the text and can involve counting the number of verbs per sentence or the number of complex noun phrases. Lexical complexity refers to the variety of words occurring in the text and is commonly measured using the type-token ratio (TTR). TTR is the total number of unique words (types) divided by the total number of words (tokens) in a given segment of language, as the TTR ratio approaches 1 the lexical variety of the text increases.

FABRA is a toolkit that outputs several readability features that can help assess the complexity of a piece of text, similar to automatic English readability metrics evaluators like Coh-Metrix [39]. The features that were evaluated to be the most important for determining the complexity of text for French L2 learners are listed in Table 4.5 along with a short description for each feature. The length-based features which were seen to affect text complexity for language learners include the length of each sentence. For syntactic complexity variables, the number of different parts of speech present, the number of constituents present, and the number of subordinate clauses. A constituent is a word or a group of words that function as a single unit within a hierarchical structure. As for lexical complexity variables, two different varieties of TTR were considered: CTTR [20] for all words in the text and UberIndex [4]for verbs.

To account for all the features and compare them across multiple variables. Similarity measures for texts generally compare two texts' semantic similarity, however, in our case, we want to compare the similarity of complexity features which are represented as numerical values. Thus, to compare multivariate data across multiple texts we scaled each complexity value, and used them to calculate the Euclidean norm for each article. The formula used was as follows,

$$|d_i\|_2 = \sqrt{\sum\left(\frac{x - x_{min}}{x_{max} - x_{min}}\right)^2} \tag{4.3}$$

where d_i is the Euclidean norm for article *i*, x is the value obtained from FABRA for the variables listed in Table 4.5, and x_{min} and x_{max} is the minimum and maximum value across all articles for that specific variable. The results of the Euclidean norm calculation can be seen in Table 4.6.

Article	Topic
A	Carnaval de Québec
B	Lionel Messi
C	Courer indien
D	Crise économique asiatique
E	Théodore Ier Lascaris
F	Maewo
G	Pas de l'ours
H	Mont Saint-Michel

Table 4.7: Articles used to test the application.

Results after filtering

Articles 6, 10, and 12 were set aside due to their high overlap in vocabulary with other articles, as seen in Figure 4.3. The overlap was calculated by taking the sum of all the overlap values and filtering out ones with the highest values. Following, looking at the distribution of word complexity by CEFR level, we noticed that article 6 had a lower ratio of A1 word, or lexicon, compared to the other articles and higher intermediate (B) level words. Additionally, article 10 has a much higher ratio of C1 vocabulary compared to the other texts, same with article 12 in the C2 level. The final decision for which articles to exclude was made after considering the linguistic features for language learner text readability. We found that the Euclidean norm of article 5 (0.611), article 10 (*1.753*), and article 12 (*2.236*) diverged from the average of 1.286 more compared to other articles. Therefore, using a combination of lexical and syntactic features to compare the texts, we filtered out articles 5, 6, 10, and 12. For better integration with the reading application, we renamed the text files from numerical to alphabetical, the final articles and their names can be seen in Table 4.7. Extra articles were kept in case participants have prior knowledge of the topics covered in texts shown to them.

4.2.3 Word and phrase extraction

After determining which articles were to be used to test the application, we set forward to extract the lexical units that will be annotated with glosses. From the preprocessing step, we were able to extract singular words within each text. We considered filtering extracted words to remove cognates. Cognates are words that have the same linguistic root as another, for example, Spanish *"atencion"* and English *"attention"*. However, there also exist false cognates, like *"demander"* mentioned in Chapter 1, which are words that are spelled similarly in two languages but have different meanings. For example, *"location"* means rental in French and *"location"* means place in English. Therefore, due to the potential error which can arise from users being unaware of false cognates, we decided to not filter out cognates in our system.

Oftentimes, a singular word does not represent what the author of the text intended, for example, the word *"pomme de terre"* in French means *"potato"* in English. Thus, for proper annotation, it was necessary to look at surrounding words to extract phrases from the text. This was done by first using `spacy`'s built-in noun phrase extraction tool that outputs noun phrases as "chunks" using the command `document.noun_chunks` where `document` is your text passed through an NLP model. Afterwards, a researcher with French knowledge manually sorted through extracted phrases and excluded the phrases that were simply a determiner with a noun. Similar steps were taken to extract named entities from the text and had a researcher sort through the extracted list to remove entries as needed. This step was necessary as some named entities have direct English equivalents while some do not have translations, for instance, *"États-Unis"* meaning *"United States"* as compared to *"Célia"* which is a name of a person and has no equivalent translation. After extracting the words, phrases, and named entities, we were left with three lists of words and phrases for each article, which will be used in the text annotation step.

4.2.4 Annotating the text

The final annotation of the text to prepare it for the reading application was completed through four steps, first, the noun phrases extracted using a machine learning model were annotated, second, the named entities, third, the extracted words, and lastly, the phrases were manually marked. The entire process along with a sample sentence to show the text marked at each step is shown in Table 4.8.

For our purposes, we enclosed the annotations between `&&` symbols, the ampersand symbol was chosen because it does not appear in the texts that we chose. We first started the annotation process by using a Python script which would run through each article three times; first to annotate the noun phrases, then the named entities, and finally the singular words. To avoid double annotations of words and words in phrases, we first annotated all the noun phrases obtained through the list generated by `spacy`. Following, we annotated all the named entities that had direct translations and then the extracted words that were not part of a phrase. Due to the lack of accurate phrase detection in French and resources to build and run large complex models, we conducted manual annotations of additional phrases using the output from the Python script.

Annotation	Method	Example
spacy phrases	Automatic	Le port du rouge, les chansons carnavalesques, la ceinture fléchée, et Bonhomme qui met dans l’ambiance sont autant de traditions qui remontent aux origines du Carnaval de Québec.
Named entities	Automatic	Le port du rouge, les chansons carnavalesques, la ceinture fléchée, et Bonhomme qui met dans l’ambiance sont autant de traditions qui remontent aux origines du Carnaval de Québec.
Words	Automatic	Le port du rouge, les chansons carnavalesques, la ceinture fléchée, et Bonhomme qui met dans l’ambiance sont autant de traditions qui remontent aux origines du Carnaval de Québec.
Phrases	Manual	Le port du rouge, les chansons carnavalesques, la ceinture fléchée, et Bonhomme qui met dans l’ambiance sont autant de traditions qui remontent aux origines du Carnaval de Québec.

Table 4.8: Steps taken to annotate the text with a sample sentence.

4.2.5 Creating a dictionary for the glosses

The annotation of the text allowed us to obtain a text where phrases, named entities, and words were enclosed by && symbols. Due to manual annotations, we were unable to completely automate the process of creating a dictionary database from the text, therefore we created a separate Python script to extract the phrases, named entities, and words from the annotated documents and create a dictionary entry for each. We used Lexicala to populate the translation, definition, and example for words that exist within their database, a sample can be seen here:

```
"syndiqués": {
"word": "syndiqué",
"fr": "syndiqué",
"en": ["join a union"],
"definition": "(inf. syndiquer) adhérer à un syndicat de
travailleurs",
"ex": "Elle s'est syndiquée à son arrivée dans l'entreprise.",
"level_freq": 3000,
"level_cefr": "b2"
}
```

To determine the word frequency rank we used the rank of the word in FLELex. For the CEFR level estimation, we approximated the level by annotating the word with the level where it has the maximum occurrence frequency. For entries that are not in the Lexicala database, like named entities and phrases, we used the Google Translate API to find the translation of it. The frequency rank was approximated by using an average of from all the words present in the phrase, and the CEFR level was determined by the level of the highest levelled singular word in the phrase. Lexical units were extracted from each text and combined into a single dictionary database containing 1624 entries.

4.3 Summary

Due to the lack of readily available resources in French compared to English, we needed to find and process the text used for the reading application built. The main requirements that we wanted from the text were that it must be French, cover various topics containing minimally overlapping vocabulary, have a similar difficulty, and preferably be written in Canadian French. We reached out to French learning centres in Canada and the CLB staff to request recommendations. Many of the recommended resources only provided curriculum guidelines and copyrighted texts, therefore were not suitable for our purposes. In the end, we used shortened open-sourced text from French Wikipedia from 12 different topics for a total of 12 articles. For the sake of comparison during participant testing, we needed to evaluate the texts and filter out articles that either contain too much similar vocabulary as others or diverge from the average text complexity. We needed the texts to have a similar number of complex words at each CEFR level and a similar readability rating. The overall process of the filtering can be seen in Figure 4.1. The text similarity and complexity were evaluated on a lexical and syntactic scale. To evaluate lexical similarity, we first looked at the number of overlapping words between each article, then looked at the number of words belonging to each CEFR level, and then the TTR values, obtained using FABRA [83]. Secondly, we compared the readability of each article based on syntactic features like the number of POS tags per sentence and the number of constituents. Finally, after obtaining the final seven articles to be used in the participant study, as seen in Table 4.7, we annotated each one with words that needed glosses and gathered the data for gloss information into a large dictionary. The process of annotation can be seen at a high level in Table 4.8. The dictionary was stored as a JSON file and populated using a combination of Lexicala and Google Translate API.

5 System Design

This chapter goes through the design process for building a reading aid in French utilizing the processed and annotated text from Chapter 4. First, the various components of the system are described, including the back-end server and the front-end interface with the different systems built for each gloss type. Following, we highlight the different gaze correction methods employed by the system to improve the accuracy of gaze tracking including filtering methods, variable offset based on screen position, and manual adjustments options available during the study. To build the system, we used an iterative design process where we first built the system and then conducted pilot studies to test the system and receive feedback, then based on the feedback implement any necessary changes. When referring to lexical units which have glosses in this chapter, word and phrase will be used interchangeably to refer to singular words and multi-word phrases.

5.1 Application components

The reading aid system is made up of a front end and a back-end, however, most of the processing is done in the front end. A high-level diagram of the system can be seen in Figure 5.1. The back-end first obtains the gaze position, as pixel coordinate, from the Tobii eye tracker and outputs the gaze data onto a console. The front-end interface is made up of an article text display area alongside interlinear glosses and a right-side margin area that is used to display margin glosses when available. Prior research has shown that interlinear and marginal glosses are effective glosses for L2 learners [1, 2, 22, 86]. The front end first fetches the gaze data asynchronously from a renderer process created by the back-end to the main process for post-processing. Details on each post-processing step is presented later in this chapter. The gaze point from the back-end is then transferred into the main app processing to increase gaze position accuracy. The system first detects if a fixation or saccade has occurred, and if a regression with large enough vertical shift has taken place, then the system will check if a new line event has taken place. Otherwise, it will determine whether the system needs to trigger a marginal gloss or an interlinear gloss. Marginal glosses are displayed after the

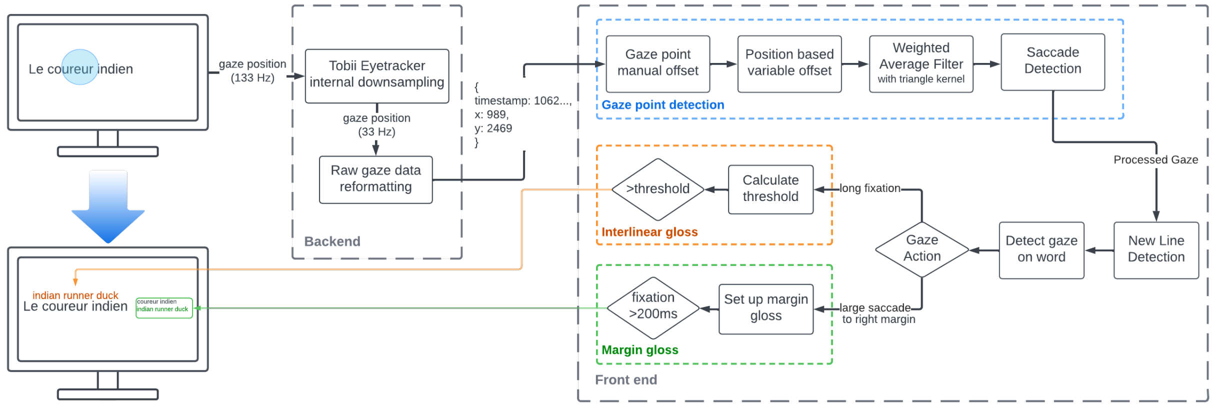


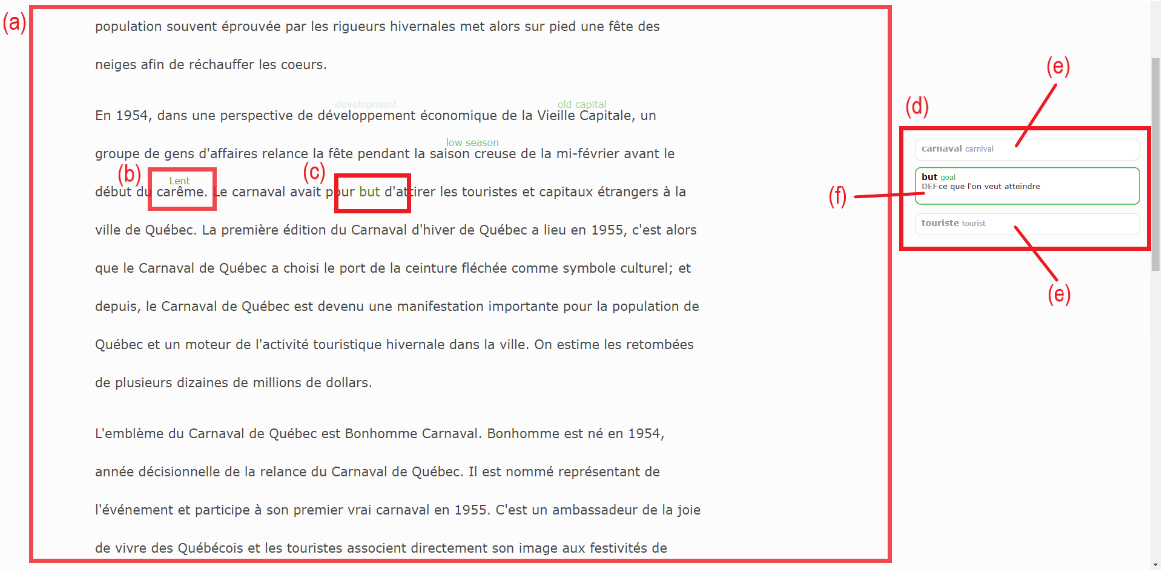
Figure 5.1: System diagram showing how eye tracking data is processed where the gaze is shown in the monitor on top as a translucent blue circle.

gaze point shifts into the right margin, and interlinear glosses are displayed after the fixation duration exceeds some threshold value.

5.1.1 Back-end

The back-end of the reading application is made up of C# scripts which extract data from the internal Tobii processor. Most eye tracking studies are conducted using eye trackers that are able to sample at 60 Hz or higher [45]. During testing, we utilized commercially available Tobii eye trackers 4C and 5. The Tobii 4C collects samples at 90 Hz and allows users to access the interlaced gaze data at 90 Hz. In comparison, the Tobii 5 samples at 133 Hz, however, due to internal downsampling, the non-interlaced gaze data that is accessible is at 33 Hz meaning the time interval between each gaze sample is 30.3 ms. The raw gaze position containing the x-coordinate, y-coordinate, and attention boolean is extracted from the Tobii internal processor, then the timestamp is appended before being printed onto a console. The attention boolean is `TRUE` when the gaze position is on the screen that the eye tracker is set up on, and `FALSE` when it isn't. A sample data packet that is sent to the console is shown here:

```
{ "timestamp":6930675.1383, "attention":"TRUE", "x":890, "y":2969 }
```



population souvent éprouvée par les rigueurs hivernales met alors sur pied une fête des neiges afin de réchauffer les coeurs.

En 1954, dans une perspective de développement économique de la Vieille Capitale, un groupe de gens d'affaires relance la fête pendant la saison creuse de la mi-février avant le début du carême. Le carnaval avait pour but d'attirer les touristes et capitaux étrangers à la ville de Québec. La première édition du Carnaval d'hiver de Québec a lieu en 1955, c'est alors que le Carnaval de Québec a choisi le port de la ceinture fléchée comme symbole culturel; et depuis, le Carnaval de Québec est devenu une manifestation importante pour la population de Québec et un moteur de l'activité touristique hivernale dans la ville. On estime les retombées de plusieurs dizaines de millions de dollars.

L'emblème du Carnaval de Québec est Bonhomme Carnaval. Bonhomme est né en 1954, année décisionnelle de la relance du Carnaval de Québec. Il est nommé représentant de l'événement et participe à son premier vrai carnaval en 1955. C'est un ambassadeur de la joie de vivre des Québécois et les touristes associent directement son image aux festivités de

carnaval carnival

but goal
DEF ce que l'on veut atteindre

touriste tourist

Figure 5.2: Front-end interface for the system showing (a) the **main text area** with (b) **interlinear glosses** shown on top of a previously glossed word and (c) the highlighted **word** *but* which has (d) **marginal gloss area** shown on the right side. (e) The **previous word** and **next word** which appears in the dictionary are shown above and below (f) the **margin gloss** highlighted is for which is coloured green in the text.

5.1.2 Front-end

The front-end of the reading aid application was built using Electron[1], which allows you to build desktop applications using HTML, CSS, and JavaScript. The main interface is made up of several components and is divided into two main areas: left-side text area and right-side margins. In the main text area, the article text is displayed with line masks for every line, and interlinear glosses appear on top of the word or phrase after some threshold time. When the gaze is on the main text, the margin glosses are invisible, only when the gaze shifts to the right side margin are the glosses displayed. Marginal glosses displayed include two greyed-out ones which are for surrounding lexical units and one for the most recently looked at phrase which is emphasized with a coloured border. When the user shifts their gaze from the margin back into the main text area, the phrase with the emphasized margin gloss is coloured green. A labelled screenshot of the interface can be seen in Figure 5.2. Words or phrases that have glosses are marked by `<span>` tags in HTML and labelled with a numerical ID for easy identification during post-processing.

[1] https://www.electronjs.org/

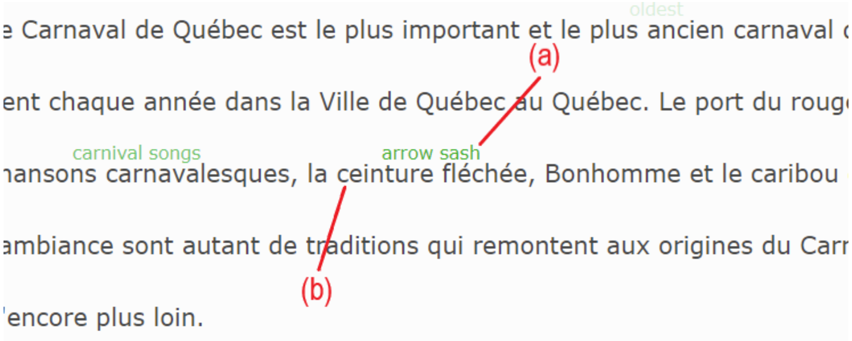
Website	Type	Text width (px)	Font size (px)	Ratio
Medium	Blog	650	14	46.4
Substack	Blog	720	14	51.4
BBC	News	713	16	44.6
CBC	News	720	16	45.0
New York Times	News	660	15	44.0
Reuters	News	700	13	53.8

Table 5.1: Websites and their text width to font size ratio.

Text style and colour

The colour of the background and text play an important role during reading on computer screens [41]. Prior studies have shown that dark characters on a light background are superior to light characters on a dark background when the refresh rate is fairly high, like on modern monitors. Bauer and Cavonius found that participants scored 26% higher when answering questions about texts when the text was presented with dark characters on a light background [9]. We initially set our background to white (`#FFFFFF`) and text to black (`#000000`) for maximum contrast. However, after running a few pilot studies, we received feedback saying that reading and answering questions with a screen with such large contrast caused eye strain. Therefore, in the final system, we changed the background colour to warmer off-white (`#FCFCFC`) and the text to dark grey (`#474747`). The colour of the marginal gloss border and interlinear gloss were set to a blue colour at the start. During a pilot study, it was mentioned that the contrast between the blue and black was too distracting, so the colour of the gloss were changed to a more muted green.

For font, we chose to use a sans-serif font, Verdana, as it was specifically created for computer screens and has high readability for long text. Our font size was set to 21 points, similar to Kuperman et al. [55]. To allow for spacing between lines to display interlinear glosses the line height was initially set to 1.6, suggested by Hyrskykari [46], as opposed to a spacing of 1.5 which is standard in eye tracking [55]. However, during testing, we found the spacing to be too small and set the spacing to 3 to increase the accuracy of the line detection algorithm. The suggestion by Hyrskykari was made based on smaller screen sizes and a smaller font size (18 points). The main text width was determined after examining text width to font size ratio of blog sites and news sites containing large blocks of text. A summary can be seen in Table 5.1. The width is set to 1000 px, thus the width and font size ratio is 47.6, which falls within the average of news and blog sites used as reference.



e Carnaval de Québec est le plus important et le plus ancien carnaval d

ent chaque année dans la Ville de Québec au Québec. Le port du roug

ıansons carnavalesques, la ceinture fléchée, Bonhomme et le caribou

ambiance sont autant de traditions qui remontent aux origines du Carr

'encore plus loin.

Figure 5.3: Interlinear gloss showing (a) the L1 gloss for (b) the L2 phrase in the text.

Interlinear gloss

Rote memorization involving mapping words from one's L2 to L1 has shown to be more useful for intermediate learners [76], who are part of the target audience for this application. Additionally, L1 glosses have been shown by many prior works to be more useful to learners than L2 glosses [23, 57, 58, 86]. Interlinear L1 glosses for words are shown on top of the word using a smaller green font, shown in Figure 5.3. Glosses were created using a superscript element in HTML, <sup>, and labelled with an ID matching the word glossed. The interlinear gloss fades as the user triggers more interlinear glosses, and the maximum number of interlinear glosses shown at a time is 4. When the maximum number of gloss is shown on screen, the earliest triggered one disappears. The gloss on the word is renewed when the user's gaze fixates on the word for a duration longer than a calculated threshold again.

For our system, we intended interlinear glosses to be activated using implicit gaze intervention techniques, where the system detects automatically when the user is having difficulty with a word and showing gloss. Prior studies have shown that gaze can be used to determine whether the reader is having difficulties or not by examining the fixation duration, with higher fixation duration correlating to lower proficiency in L2 [6, 50, 88]. Therefore, we chose to use dwell time to implicitly inform the system when the user is having difficulties. We tested using CEFR levels and word frequency to calculate the threshold, however, found that the estimation of CEFR level is not precise [70]. The word frequencies were obtained from FLELex, as detailed in Chapter 4. The equation, adapted from Hyrskykari [46], used to calculate the base threshold time to trigger intelinear gloss for word or phrase *i* to be displayed based on the frequency of the word is as follows:

$$th_{i,base} = th_H + (freq(i) - w_L)\left(\frac{th_H - th_L}{w_L - w_H}\right) \tag{5.1}$$

where th_H and th_L are the upper and lower boundaries for trigger fixation duration in ms, w_H and w_L are the upper and lower boundaries for word frequency in FLELex. The lower and upper boundaries for threshold time were initially set to 350 ms and 700 ms, respectively, after pilot testing the system multiple times. The boundaries for fixation duration and word frequency are calculated using previous word fixation duration and word frequency of other triggered glosses, for additional personalization of threshold timing based on user reading habits. The values are updated for every gloss trigger using Algorithm 1 and 2, both assuming that the distribution of duration and word frequency should be a Gaussian normal curve. In Algorithm 1, the input variable includes the word (*w*), gaze duration (*gaze_dur*), previous lower and upper thresholds (th_L, th_w), and an array containing duration per characters of previously glossed words (`prev_durations`). To re-calculate the threshold values, the duration per character for the most recently looked at word ($char\_dur$), the mean and median of all previous duration per character with *char_dur* are calculated. Considering the error margin for the eye tracker due to sampling error is 30.3 ms, if the mean (*mu*) falls within one standard deviation (*sigma*) on both sides–or 70 ms away from the median– then we shift the lower and upper threshold accordingly. We also take into account the threshold falling to extreme values by setting limit values of the lower and upper threshold to 200 ms and 400 ms, accordingly. The value 200 ms was chosen as the lower threshold because it is the average fixation duration and 400 ms is double that value try to ensure that the fixation is intentional. In addition, an upper boundary for the longest duration is set to 2000 ms. A similar approach is taken to calculate the boundaries for word frequency, except instead of per character duration, we calculate the mean, median, and standard deviation of word frequencies. The input variables into Algorithm 2 are the frequency ranking of the word ($w.freq$) and a list of frequency rankings of previously glossed words (`prev_freq`).

Additionally, two of the most important features used to determine a word's complexity are the word length and frequency in the language [46]. Therefore, the threshold for the duration of fixation on a word to determine if the word is difficult to the user should be a function of frequency and length. The previous equation takes into consideration the frequency, so we introduced an additional length based term to increase the threshold time for longer phrases since people tend to fixate longer on longer phrases,

$$th_i = th_{i,base} + 5 \times (length(i) - 8.25) \tag{5.2}$$

where the length counts how many characters are in the word or phrase and 8.25 (σ=2.86) is the average length of French words [63]. We are introducing an additional or subtracting

5 ms per character above or below the average from the calculated base threshold. This was necessary as we found during testing that long phrases were triggered almost instantly when fixated upon. Finally, after all calculations, if a word's frequency falls below th_{low} and above th_{high} then the threshold is set to the closest boundary.

Algorithm 1 Gaze fixation duration boundaries.

```
function GETGAZEDURATIONBOUNDARY($wh, gaze\_dur, th_L, th_H$, prev_durations)
    $th_{L,min} \leftarrow 200$                          ▷ minimum time to trigger is 200 ms
    $th_{H,min} \leftarrow 400$
    $th_{H,max} \leftarrow 2000$                         ▷ maximum time to trigger is 2000 ms
    $skewed \leftarrow$ True
    while $skewed$ and $th_L < th_H$ do
        $char\_dur \leftarrow$ length$(w)/gaze\_dur$
        $mu \leftarrow$ math.avg$([char\_dur, prev\_durations])$
        $median \leftarrow$ math.median$([char\_dur, prev\_durations])$
        $sigma \leftarrow$ math.std$([char\_dur, prev\_durations])$
        if $mu > median + sigma + 35 * 2$ then
            $th_L \leftarrow th_L + 10$
            $th_H \leftarrow th_H + 5$
        else if $mu < median - sigma - 35 * 2$ then
            $th_L \leftarrow th_L - 5$
            $th_H \leftarrow th_H - 10$
        else
            $skewed \leftarrow$ False
        end if
        if $th_L < th_{L,min}$ then
            $th_L \leftarrow th_{L,min}$
        end if
        if $th_H < th_{L,min}$ then
            $th_H \leftarrow th_{H,min}$
        end if
        if $th_H > th_{H,min}$ then
            $th_H \leftarrow th_{H,max}$
        end if
    end while
    return $[th_L, th_H]$
end function
```

Algorithm 2 Frequency boundaries of seen glosses.

```
function GETFREQBOUNDARY($w.freq$, prev_freq)
    $w_L \leftarrow 300$                                  ▷ lower frequency bound from FLELex
    $w_H \leftarrow 13000$                                ▷ upper frequency bound from FLELex
    $skewed \leftarrow$ True
    while $skewed$ do
        $mu \leftarrow$ math.avg($[w.freq, prev\_freq]$)
        $median \leftarrow$ math.median($[w.freq, prev\_freq]$)
        $sigma \leftarrow$ math.std($[w.freq, prev\_freq]$)
        if $mu > median + sigma$ then
            $w_L \leftarrow w_L + 20$
            $w_H \leftarrow w_H + 10$
        else if $mu < median - sigma$ then
            $w_L \leftarrow w_L - 10$
            $w_H \leftarrow w_H + 20$
        else
            $skewed \leftarrow$ False
        end if
    end while
    return $[w_L, w_H]$
end function
```

Marginal gloss

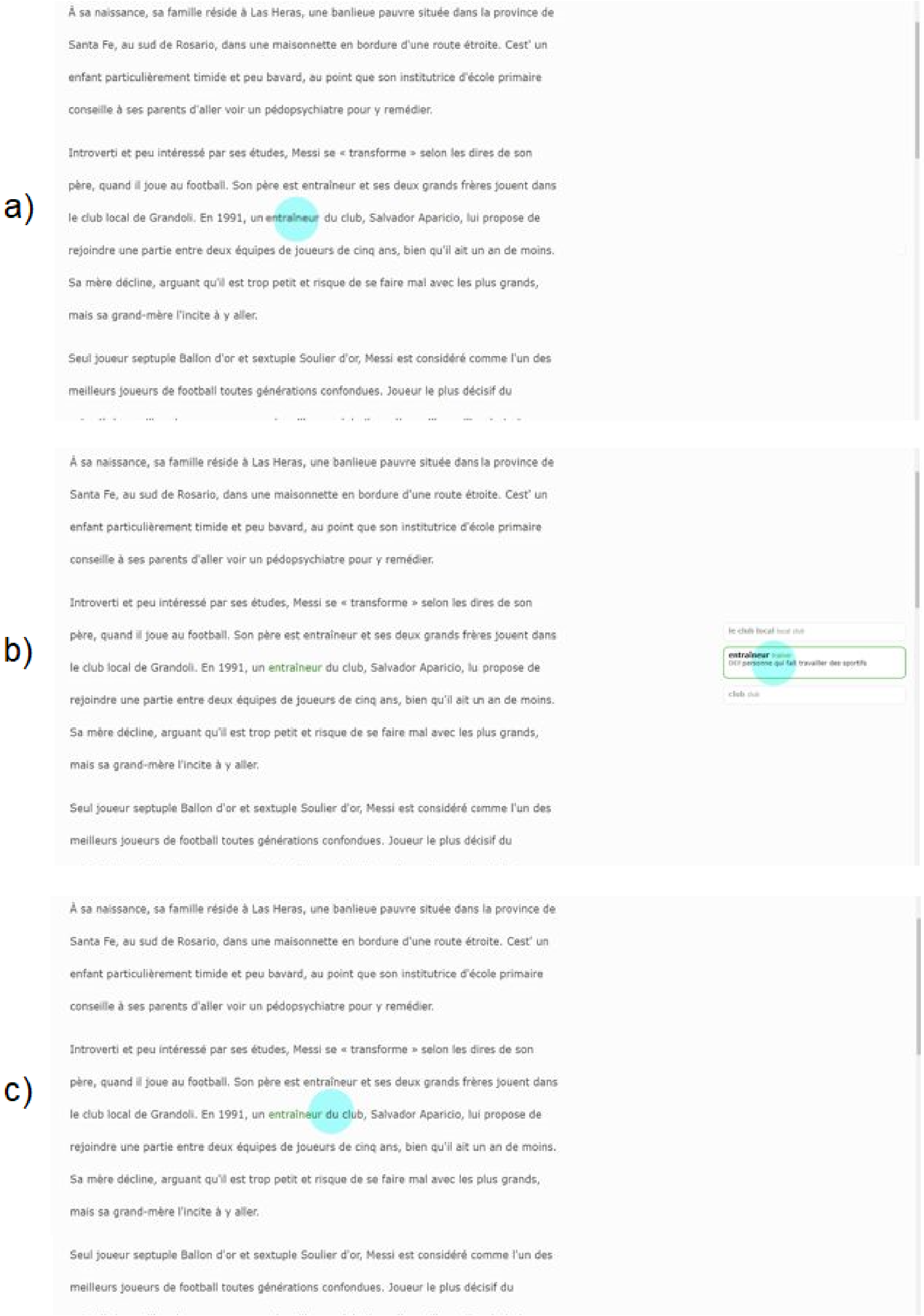
We created marginal gloss as a way to test a explicitly gaze controlled reading assistance technique. Marginal gloss were created in the application using D3.js[2], which is a JavaScript library used to produce dynamic and interactive data visualizations in web browsers. The right-hand margin was chosen after consulting prior research which showed that for reading in languages written from left to right, placing additional help on the right causes less disturbance to readers [1, 2, 86]. The marginal gloss is created after the user looks at the word or phrase in the text for at least 200 ms, which is the average fixation duration while reading. To show the marginal gloss, the user needs to shift their gaze to the right margin, thus explicitly telling the system that they want help for a specific word. The marginal gloss for the previous word and the word following the most recently looked at will be displayed within greyed out boxes on top and below the most currently looked at word's marginal gloss. The display of previous and following words, gives users a preview of the next word they might want help for and reviews recent content. Additionally, to help users keep track of their place in the text when they shift their gaze away from the right margin, we colour the most recently glossed word in green, shown in Figure 5.4. The colour is faded after another fixation event occurs.

Initially, we considered including synonyms, translations, definitions, and example usage as part of the marginal gloss. However, due to a lack of resources providing synonyms and example usage for a sufficient percentage of the dictionary words, we decided to simply include translations for all words and definitions for singular words. The final design of the marginal gloss can be seen in Figure 5.5. The highlighted gloss contains the word most currently gazed at in the text, a L1 translation, and a definition in L2. The combination of L1 translation and L2 definition has been shown to lead to increased learning gain when using glosses [86].

5.2 Gaze correction

This application requires gaze position to be accurate and precise on the word level since we are providing glosses for words in text while reading. The accuracy of eye trackers is dependent on their sampling frequency and internal processing. Other than reducing causes for optical artifacts (e.g. glares from glasses, reflections, shadows, physical obstructions) during usage, post-processing steps can be taken after obtaining the raw gaze data to improve

[2] https://d3js.org/



À sa naissance, sa famille réside à Las Heras, une banlieue pauvre située dans la province de Santa Fe, au sud de Rosario, dans une maisonnette en bordure d'une route étroite. Cest' un enfant particulièrement timide et peu bavard, au point que son institutrice d'école primaire conseille à ses parents d'aller voir un pédopsychiatre pour y remédier.

Introverti et peu intéressé par ses études, Messi se « transforme » selon les dires de son père, quand il joue au football. Son père est entraîneur et ses deux grands frères jouent dans le club local de Grandoli. En 1991, un entraîneur du club, Salvador Aparicio, lui propose de rejoindre une partie entre deux équipes de joueurs de cinq ans, bien qu'il ait un an de moins. Sa mère décline, arguant qu'il est trop petit et risque de se faire mal avec les plus grands, mais sa grand-mère l'incite à y aller.

Seul joueur septuple Ballon d'or et sextuple Soulier d'or, Messi est considéré comme l'un des meilleurs joueurs de football toutes générations confondues. Joueur le plus décisif du

À sa naissance, sa famille réside à Las Heras, une banlieue pauvre située dans la province de Santa Fe, au sud de Rosario, dans une maisonnette en bordure d'une route étroite. Cest' un enfant particulièrement timide et peu bavard, au point que son institutrice d'école primaire conseille à ses parents d'aller voir un pédopsychiatre pour y remédier.

Introverti et peu intéressé par ses études, Messi se « transforme » selon les dires de son père, quand il joue au football. Son père est entraîneur et ses deux grands frères jouent dans le club local de Grandoli. En 1991, un entraîneur du club, Salvador Aparicio, lui propose de rejoindre une partie entre deux équipes de joueurs de cinq ans, bien qu'il ait un an de moins. Sa mère décline, arguant qu'il est trop petit et risque de se faire mal avec les plus grands, mais sa grand-mère l'incite à y aller.

Seul joueur septuple Ballon d'or et sextuple Soulier d'or, Messi est considéré comme l'un des meilleurs joueurs de football toutes générations confondues. Joueur le plus décisif du

le club local local club

entraîneur trainer
DEF personne qui fait travailler des sportifs

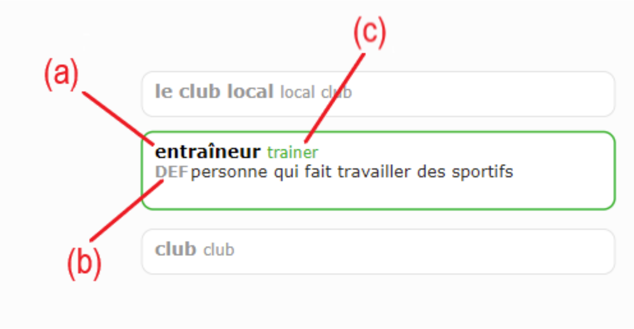
club club

À sa naissance, sa famille réside à Las Heras, une banlieue pauvre située dans la province de Santa Fe, au sud de Rosario, dans une maisonnette en bordure d'une route étroite. Cest' un enfant particulièrement timide et peu bavard, au point que son institutrice d'école primaire conseille à ses parents d'aller voir un pédopsychiatre pour y remédier.

Introverti et peu intéressé par ses études, Messi se « transforme » selon les dires de son père, quand il joue au football. Son père est entraîneur et ses deux grands frères jouent dans le club local de Grandoli. En 1991, un entraîneur du club, Salvador Aparicio, lui propose de rejoindre une partie entre deux équipes de joueurs de cinq ans, bien qu'il ait un an de moins. Sa mère décline, arguant qu'il est trop petit et risque de se faire mal avec les plus grands, mais sa grand-mère l'incite à y aller.

Seul joueur septuple Ballon d'or et sextuple Soulier d'or, Messi est considéré comme l'un des meilleurs joueurs de football toutes générations confondues. Joueur le plus décisif du

Figure 5.4: Margin glosses are triggered when the gaze is shifted into the right margin. For these three images the gaze is shown in blue, (a) shows the gaze focused on *entraîneur*, (b) shows the margin gloss activated when the gaze is shifted to the margin and the word highlighted in green, and (c) shows the word glossed highlighted in green and the gaze back to the text.



le club local local club

entraîneur trainer
DEF personne qui fait travailler des sportifs

club club

Figure 5.5: Marginal gloss showing (a) the L2 word that the gloss is provided for, (b) the L1 translation, and (c) the definition of the word in L2.

the accuracy including introducing an offset and reducing oculomotor noise [45]. As can be seen in the system diagram, Figure 5.1, after the gaze is extracted from the back-end, it goes through a gaze point detection process which involves first applying manual offsets, variable offset, and gaze filtering. Additionally, we keep track of which line the user is on to ensure that we are able to quickly detect which word they are fixated on, and when the algorithm fails we provide methods for the reader to re-calibrate.

5.2.1 Offset adjustments

Introducing offset adjustments during post-processing steps can increase the accuracy of gaze position. Hyrskykari found that eye trackers are inherently less precise in the vertical direction when reading because the target (line-height) is oftentimes smaller than the width of the target based on word length [47]. Therefore, in our system, we first have users make manual adjustments at the start of the application. The user is shown a calibration screen with arrow buttons and text, allowing them to move their gaze point to where they are actually looking, as shown in Figure 5.6. This offset is applied statically to the x- and y- coordinates of the raw gaze points from the back-end.

During testing, we noticed that the gaze accuracy varied greatly depending on the screen position and height of the user. Feit et al. and others previously showed that the tracking quality of eye trackers deviates from numbers obtained in the manufacturer's lab and varies widely across different tracking conditions and users [31]. In terms of position-based precision, Feit et al. found that the gaze points were most variable on the right and bottom margins of the screen. We ran a similar test and collected gaze samples from three users staring at evenly spaced out points on the screen. The screen was separated into six columns labelled using numbers 1-6 and five rows labelled using letters A-E, as seen in Figure 5.7. The aver-

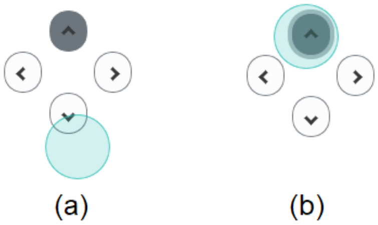
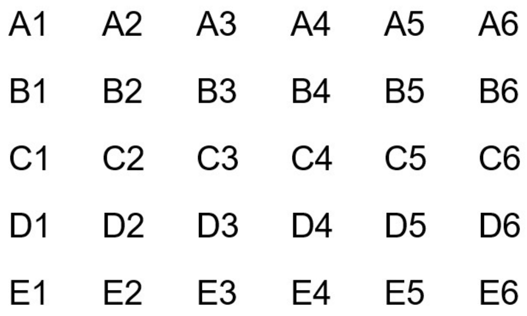


Figure 5.6: Manual gaze calibration shifts users can use before starting the application to change the gaze offset. (a) Shows the initial offset and (b) shows the gaze point after manual calibration.



A1	A2	A3	A4	A5	A6
B1	B2	B3	B4	B5	B6
C1	C2	C3	C4	C5	C6
D1	D2	D3	D4	D5	D6
E1	E2	E3	E4	E5	E6

Figure 5.7: Enlarged image showing the screen position of labels used for manual offset calibration.

	1	**2**	**3**	**4**	**5**	**6**
A	-5.0	77.3	135.7	165.0	233.7	274.3
B	43.3	75.7	153.0	185.0	223.0	274.7
C	22.7	78.3	130.0	176.3	232.3	274.0
D	21.7	68.0	116.0	172.7	250.7	292.7
E	12.3	75.3	109.0	176.0	225.0	282.0

Table 5.2: The average horizontal offset in pixels.

	1	**2**	**3**	**4**	**5**	**6**
A	-69.0	14.7	-2.3	15.0	-5.3	-12.3
B	21.3	10.0	26.3	29.7	40.3	15.3
C	19.3	23.0	30.7	47.0	47.7	49.0
D	41.7	54.0	54.0	76.0	62.7	68.0
E	85.3	95.7	93.0	86.7	104.7	107.0

Table 5.3: The average vertical offset in pixels.

aged deviation based on the positioning in the horizontal direction is shown in Table 5.2 and the vertical is shown in Table 5.3. Similar to Feit et al., the gaze points on the right columns (5-6) on average varied 200 px horizontally to the right than the actual gaze location. While the lowest row (E) averaged to more than 85 px lower than in the vertical direction. Using the average offset from the collected sample, we calculated the variable offset adjustments by calculating the average horizontal change (slope) in offset value as a function of horizontal position, using a domain of [0,1920], where 1920 px is the width of the screen, and obtained the following offset adjustment formula:

$$x = x - 0.09895 \times x \tag{5.3}$$

Similarly, in the y direction, we calculated the average rate of change in the vertical offset value as a function of vertical position, using a range of [0,1080], where 1080 px is the height of the screen. Additionally, there is an additional vertical upward drift due to people's reading habits as they move horizontally across texts which has been noted by other researchers [25, 46, 71]. To account for that upward drift, we introduced a second term to adjust the vertical offset which is dependent on horizontal gaze position. The formula obtained is as follows:

$$y = y - 0.1204 \times y - 0.0208 \times x \tag{5.4}$$

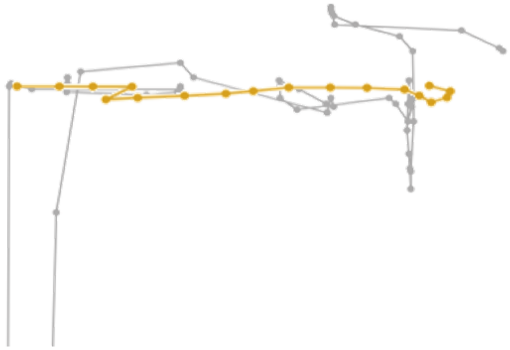


Figure 5.8: The raw gaze is shown in grey and the resulting filtered gaze is shown in yellow.

5.2.2 Gaze filtering

Gaze points after introducing offsets at this point are more accurate than the raw gaze data obtained initially, however, gaze points contain jitters and microsaccades due to human's inherent oculomotor movements. Therefore, there needs to be a smoothing step applied to the gaze points so that the gaze moves smoothly across the screen when users choose to have their gaze visible. We use a weighted average filter followed by a saccade detector, as suggested by Feit et al. [31]. The weighted average filter computes the filtered point $\hat{x_t}$ at time t over the last N points,

$$\hat{x_t} = \sum_{i=0}^{N} \frac{w_i}{\sum_j w_j} \cdot x_{t-i} \tag{5.5}$$

where x_{t-i} is the point at time $t - i$and $\frac{w_i}{\sum_j w_j}$ is the corresponding normalized weight. The weight w_i is calculated by assigning a weight of 1 to the least recent point (N), 2 to the next point ($N - 1$), and so on such that $w_i = N - i + 1$. From testing with an eye tracker running at 60 Hz, Feit et al. determined that the optimal window size (N) for horizontal and vertical should be set to 36 and 40. Due to the lower sampling frequency, we set our window size to 15 and 19 for horizontal and vertical, respectively. Saccade detection can be used to extend the weighted average filter, it defines a threshold s used to detect saccades so as long as the distance between successive gaze points is larger than s. We used a threshold of 0.7 cm, which correlates to 26 px. The results after passing the raw gaze points through the filters is shown in Figure 5.8.

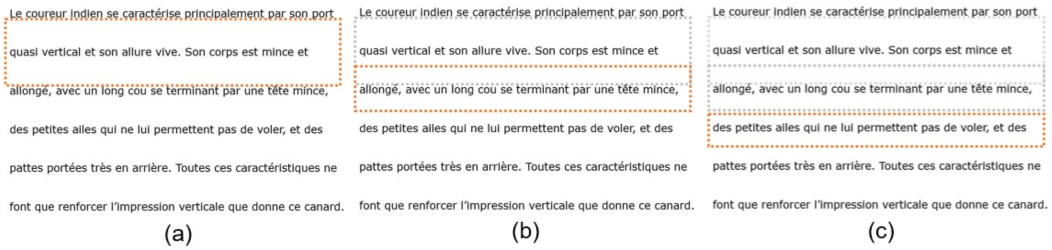
5.2.3 Word detection

Generally, during reading, readers read through one line at a time, sequentially, with occasional regression to previous lines [25]. Initially, we used a large loop to check which word the user was looking at by checking all the possible words shown on the screen, however, this caused lags and errors due to insufficient processing power. To narrow down the number of words processed for every fixation, we took inspiration from iDict [46], where the system keeps track of the line the user reads, as shown in Algorithm 3.

Algorithm 3 New line detection algorithm.

```
function NEWLINEEVENT($x, y, current\_line,$ allRawFixations)
    $min\_y\_shift \leftarrow 0.15$
    $max\_forward \leftarrow 5$
    $max\_regression \leftarrow 10$
    $width \leftarrow$ getTextWidth($current\_line$)
    $height \leftarrow$ getTextHeight($current\_line$)
    $margin\_left \leftarrow$ left($margin$)
    $N \leftarrow$ length(allRawFixations)
    if $current\_line > 0$ and $N > 0$ then
        $x\_shift \leftarrow$ allRawFixations$[N].x -$ allRawFixations$[N-1].x$
        $y\_shift \leftarrow$ allRawFixations$[N].y -$ allRawFixations$[N-1].y$
        if allRawFixations$[N-1].x < margin\_left$ and $x\_shift < 0$ and
math.abs($y\_shift$) $> min\_y\_shift * height$ then        ▷ Regression with large enough
vertical movement occured
            $weight \leftarrow 0.5$                ▷ line mask size increase by 50% at the start
            if $y\_shift < 0$ then                        ▷ Backward movement
                $start \leftarrow current\_line + 1$
                $end \leftarrow current\_line + max\_forward$
            else                                          ▷ Forward movement
                $start \leftarrow current\_line - max\_regression$
                $end \leftarrow current\_line$
            end if
            $current\_line \leftarrow$ setCurrentLine($x, y, start, end, weight$)    ▷ Algorithm 4
        end if
    end if
    return $current\_line$
end function
```

To keep track of the line, we implemented a new line detection algorithm to keep track of the line. The input variables to this algorithm include the gaze position as coordinates (*x,y*), the current line number (*current_line*), and an array containing the unfiltered raw fix-



(a)

Le coureur indien se caractérise principalement par son port
quasi vertical et son allure vive. Son corps est mince et
allongé, avec un long cou se terminant par une tête mince,
des petites ailes qui ne lui permettent pas de voler, et des
pattes portées très en arrière. Toutes ces caractéristiques ne
font que renforcer l'impression verticale que donne ce canard.

(b)

Le coureur indien se caractérise principalement par son port
quasi vertical et son allure vive. Son corps est mince et
allongé, avec un long cou se terminant par une tête mince,
des petites ailes qui ne lui permettent pas de voler, et des
pattes portées très en arrière. Toutes ces caractéristiques ne
font que renforcer l'impression verticale que donne ce canard.

(c)

Le coureur indien se caractérise principalement par son port
quasi vertical et son allure vive. Son corps est mince et
allongé, avec un long cou se terminant par une tête mince,
des petites ailes qui ne lui permettent pas de voler, et des
pattes portées très en arrière. Toutes ces caractéristiques ne
font que renforcer l'impression verticale que donne ce canard.

Figure 5.9: The influence of weight value on the size of the line height window, (a) shows a line height increased by a weight of 0.5, (b) shows a line height increased by a weight of 0.25, and (c) show a line height increased by a weight of 0.

ation coordinates (`allRawFixations`). A new line event is detected when the horizontal difference between the previous and most recent fixation is negative, meaning a regression has occurred. We initially introduced a horizontal threshold for the saccade, where new line events are only triggered when the horizontal regression exceeds 60% of the line width. However, during testing, it caused errors because there would be accidental fixations between line changes causing inaccurate new line detections. People often paused in the middle of a line before moving from the end of the previous line to the start of the next line. We found that having the program trigger the detection after a regression worked more accurately in determining which line the user was on. Due to the risk of triggering a new line detection when people regress to read previous words on the same line, we detect if a large enough vertical shift (*y_shift*) has occured. The combination of a regression (*x_shift*<0) and sufficiently large vertical shift triggers the new line detection algorithm to detect which line the user is on. So, if we suppose the vertical distance between the previous and newest fixation is negative, then the user is reading a previous line and we start detecting from the current line minus the maximum possible line regression based on the number of lines visible on the screen (*max_regression*). If the vertical distance is positive, we start looking for their line starting from the current line plus one. The algorithm uses Algorithm 4 to set the new current line number and uses the variable *weight* to increase the height of the line mask used to detect the line. In the new line detection, more weight is given to the lines nearby the most current line, as they are the most likely to be read next. In the new line detection algorithm the weight is initially set to 0.5, meaning that the window for detecting the next line is increased by 50%, see Figure 5.9 for a visualization. Moreover, after scroll events and when the user presses the keyboard button `F`, the line the user's gaze is on is re-detected using Algorithm 4 with *weight* set to 0.

At the word level, when people fixate upon a word or a point on the screen, that fixation is often made up of multiple fixations, with small saccades in between. Thus, we need to create a word boundary box which allows room for error, Buscher et al. suggest extending the target area to at least 50 px by 50 px [17]. In the application, each word with a gloss has a 30 px margin extending on all sides to accommodate for shifts in fixations.

Algorithm 4 Detecting and setting the current line.

```
function setCurrentLine($x, y, start, end, weight$)
    $current\_line \leftarrow start$
    for $idx \leftarrow start, end$ do
        if isOnLine($x, y, idx, weight$) then
            $current\_line \leftarrow idx$
        end if
        if $weight > 0$ then
            $weight \leftarrow weight -$ math.abs($1/(end - start)$)
        end if
    end for
    return $current\_line$
end function
```

6 Evaluation

Human-computer interaction (HCI), which is the main field of study this thesis falls into, focuses on designing with people in mind people and centres on the user experience. To assess a system's functionality, how people feel and their thoughts on the system we created. We need to investigate which gaze controlled glossing method provides the best assistance for French L2 learners while reading French text. Through a participant study, we want to address the latter two questions posed in Chapter 1, namely whether (**RQ2**) the placement of glosses affects the perceived and measured reading comprehension while reading in L2 and (**RQ3**) what kind of gaze intervention do users prefer. To truly answer these two questions, we would need to conduct a longitudinal study with a group of participants and seek aid from researchers focusing on SLA research to analyze learning gain from each gloss type, which is out of the scope of this thesis. Therefore, we conducted a smaller-scale participant study to test the preference for gloss placements using gaze data and self-reported ratings to evaluate system performance. Using a reading comprehension quiz and interview transcripts as guidelines for participant understanding of text, we compare their perceived versus measured comprehension. Additionally, we utilized interviews and open-ended questions in usability questionnaires to prompt participants for information on the types of glossing methods they found useful. We present the results from our participant study in the latter half of this chapter.

6.1 Participant Study

Eye-tracking studies in the language learning context generally follow a similar format, borrowing many components from eye-tracking studies in psychology. The main components of our study include participants, eye trackers, tracking environment, equipment setup, tasks to complete, procedure, and data collection. We collected gaze data, reading comprehension question answers, usability questionnaire responses, and interview responses from our participants. In the original study setup, the participant completed 5 tasks testing 4 different gloss positions (footer, marginal, interlinear, and glossary). However, footer and glossary

CEFR level	Number of participants
A1	7
A2	4
B1	8

Table 6.1: Participant's self-reported French proficiency.

glosses showed to lead to statistically insignificant changes from no gloss conditions [1, 2, 86], so they were excluded from the study. Our study involved testing four conditions: no help, interlinear gloss, marginal gloss, and both interlinear and marginal gloss. Our control method, which act as a baseline for comparison, is the no help condition.

6.1.1 Participants

We recruited a total of 20 participants; however, due to data collection error, we obtained data from 19 participants (N=19). The study lasted 1.5 hours and each participant was compensated $30 for their time. When recruiting for the study we required participants to read through the following exclusion criteria to minimize eye tracking error [10]:

- wear bifocal or progressive glasses
- has dyslexia, strabismus, nystagmus
- have lazy eyelids, droopy eyes, or prosthetic eyes
- has uncorrected astigmatism
- have photosensitive epilepsy; and,
- have eye surgery, including corneal (e.g., LASIK, RK), cataract, and intraocular implants

We also mentioned to participants that they should not be advanced or fluent in French (CEFR level C) and be either beginner (A1 and A2) or intermediate (B1 and B2) French learners. At the start of the study, our participants filled out a modified LEAP-Q questionnaire [64] to gather data on their L1 and L2 proficiencies (see Appendix A for the survey questions). We did not require them to complete a more comprehensive L2 proficiency test because we are evaluating differences in reading comprehension using glosses, therefore the main focus is their performance change from their own baseline. All participants wrote that they know English at a fluent or native-like level. Table 6.2 shows a summary of all the prior

	French		
	Mean	*SD*	*Range*
Critical age (years)			
Start learning	10.7	8.08	4-30
Start reading	11.4	7.86	4-31
Proficiency (1-5)			
Speaking	1.89	0.88	1-3
Understanding	1.84	0.83	1-3
Writing	2.47	1.02	1-4
Contributing factors (0-10)			
Friends/Family	1.63	2.75	0-10
Media consumption	3.42	2.76	0-8
Reading	3.79	2.80	0-8
Formal classes	3.63	3.39	0-10
Language learning apps	6.89	3.30	0-10

Table 6.2: Study participant modified LEAP-Q results information.

language proficiency knowledge of the participants. We correlated the proficiency (1-5) to the standard CEFR levels, where 1 refers to A1 and 5 refers to any C level. The distribution of the participant's French proficiency can be seen in Table 6.1. In addition to English and French, many participants noted they spoke Mandarin Chinese, Portuguese, Hindi, Bengali, Tamil, Italian, and Arabic see Table 6.3 for a more detailed breakdown.

6.1.2 Setup

We set up the tracking environment in a room lit with both natural light from a window on the right and artificial overhead lighting. To accommodate participants at various heights, participants were seated in a chair where the height could be adjusted, and the position of the chair wheels was marked using two pieces of white tape, as seen in Figure 6.1. During testing, we noticed that user height played a big role in tracking accuracy as the distance they were away from the tracker varied significantly. Therefore, we tracked the participant's distance from the eye tracker. The horizontal distance d and the net distance l of the eye tracker to the participant were measured for each session, as displayed in Figure 6.2. The horizontal distance was kept from 50 to 70 cm, and l was adjusted using the chair's height. The mean and standard deviation of the initial distance measurements are shown in Table

Language	Number of participants	Average fluency
Arabic	1	B1
Bengali	2	C1/C2
Creole	1	B2
English	19	C1/C2
French	19	B1
Hindi	4	A2
Italian	2	A1
Mandarin	3	B2
Spanish	4	B1
Portuguese	1	C1/C2
Tamil	1	C1/C2
Urdu	2	B2

Table 6.3: Languages spoken by participants and their average fluency based on CEFR levels.

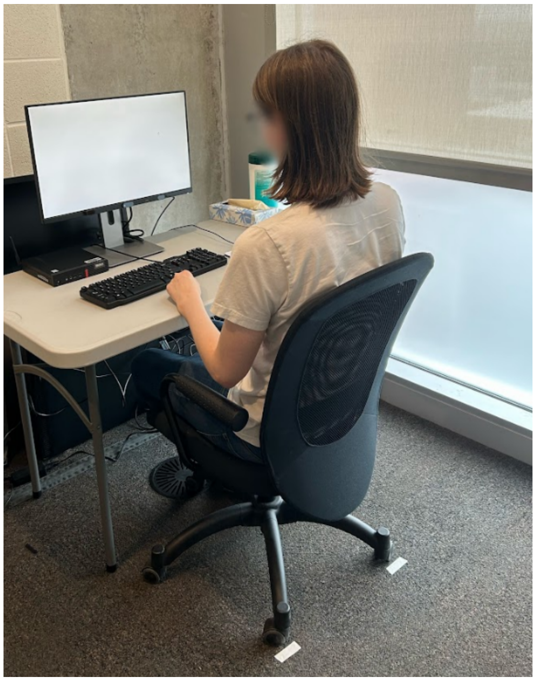


Figure 6.1: Study set up as viewed from the back.

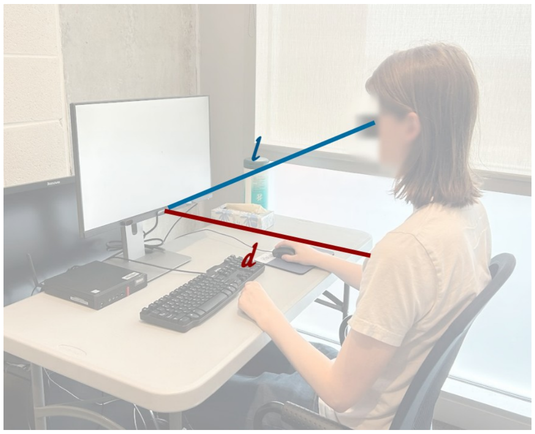


Figure 6.2: Diagram of how participant distances were measured.

6.4. With this setup, a visual angle of 1°corresponds to a movement of 1.05 cm. We used a desktop computer with an Intel Core i7-6700T CPU operating at 2.8 GHz with 8 GB RAM. The monitor used was 23.8 in with a resolution of 1920x1080. The eye tracker was magnetically attached to the bottom of the monitor using materials provided by the manufacturers. We used a Tobii 5 eye tracker, which operates at 33 Hz. During initial calibration, participants were to hold their heads still and only move their eyes to look at the corner of the screens.

	Mean	*SD*
l	63.7	4.54
d	60.1	4.47

Table 6.4: Study participant distances from the eye tracker in cm.

6.1.3 Task and Procedure

The study consisted of four reading tasks followed by a semi-structured interview. The four reading tasks each involved using a different reading intervention system or reading condition: no help (NH), marginal gloss (MG), interlinear gloss (IG), and both marginal and interlinear glosses (FS). For each reading task, the article order was randomized to remove potential learning or order effect. Additionally, after each reading task, the participants were allowed to take a break and look away from the screen to reduce potential eye strain.

The order of the articles were randomized so that they were used evenly throughout the study. Additionally, to reduce the influence of each participant's prior knowledge, we asked the participants if they had prior knowledge about the topic covered in any of the articles during the pre-study questionnaire. Articles where the participants had prior knowledge of were exchanged out for other article texts.

The procedure of each task went as follows; first, participants are shown the task screen, which contains a task number and a start button. Clicking the start button will refresh the screen to show the manual gaze calibration page, and participants can adjust their gaze manually using arrow buttons. A more detailed description was given previously in Chapter 5. After they are given a short tutorial on how to use the reading glossing method, the sentence they use to test is *"Portez ce vieux whisky au juge blond qui fume"* meaning *"Take this old whisky to the blond judge who is smoking"*, which is a French pangram, or a phrase that uses every letter in the alphabet. After the tutorial, each participant read a French article, and then completed a reading comprehension quiz. See Appendix A for a sample of the reading comprehension questions. Each question was created by hand based on the article texts and reviewed by a native French speaker for grammar errors. Except for the no-help case, the participant also completes a usability questionnaire based on USE [61] and SUS [15], see Appendix A. For ordering, the first intervention used is no-help, which is used as their baseline. Following, the order of the two glossed reading conditions, IG and MG, were counterbalanced to minimize order and learning effects. The last intervention system tested is the system where the participant has access to both the interlinear and marginal glosses.

6.2 Results

We used a mixture of quantitative and qualitative data to evaluate the effectiveness of gloss position and gaze action preferences. The data was gathered in various ways, including oral semi-structured interviews, anonymized written feedback, and real-time gaze data. The participants' gaze behaviour during the study gives us information about their proficiency and interaction with each glossing method, we found clear differences between beginner and intermediate learners when using different methods. The differences also materialized in the reading comprehension scores, with a more overall subjective comprehension of the text when using glossed text and a significant difference when using marginal glosses. Additionally, learners had both personal and proficiency-based preferences regarding the position of the gloss.

	NH		**IG**		**MG**		**FS**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
total duration (min)	6.02	2.75	11.73	8.20	9.47	3.63	7.09	3.43
total saccade count	3173	1358	4966	2742	5840	2308	3684	1409
average saccade velocity (°/s)	224.4	13.8	242.4	39.8	335.2	51.4	278.5	40.8
average fixation duration (ms)	407.6	28.8	360.4	23.3	137.1	10.9	149.0	22.8
maximum fixation duration (s)	3.26	2.63	5.49	2.56	3.87	2.14	4.91	2.70

Table 6.5: Gaze feature summary for all glossing methods.

6.2.1 Gaze behaviour

The gaze of each participant was gathered for all four tasks at 33 Hz, meaning a sample is taken every 30.3 ms. We processed the gaze samples to extract the gaze features of each participant during their session. Following previous works in determining the L2 proficiency level of learners [6, 7, 16, 50, 65, 88], gaze features measured in this study include the total duration for reading each text, average saccade velocity, saccade count, average fixation duration, and maximum fixation duration. Although maximum saccade velocity, scanpath length, and blink features have shown to be useful, due to the lower accuracy of commercial eye trackers and noise, it is difficult to determine those values, so they were not used. Additionally, to track the usage of each type of gloss for the three glossing methods (IG, MG, FS), we recorded the id of the gloss interacted with, duration of fixation on a single gloss, number of times the gloss was triggered, and what type of gloss it was (interlinear or marginal). The fixation duration on the gloss was recorded, and the average duration was calculated.

Method comparison

A summary of the gaze features with the average of all participants and the standard deviation is shown in Table 6.5. Overall, participants spent the most time using the system, showing interlinear glosses (*Mean = 12.61 min*) followed by marginal glosses (*Mean = 10.53 min*). During pilot testing, the average time taken to read an article text with no help by a B1 level French learner was around 6-8 minutes, which aligns with the duration of the no help method, and the full system fell within the expected time. The standard deviation and range for the duration of the IG method could be attributed to increased fixation on individual words to interact with glosses. Although, interlinear glosses were meant to appear automatically after implicit gaze actions, after participants knew that glosses would appear after each word was fixated on, fixation duration became an explicit gaze control method. Participants

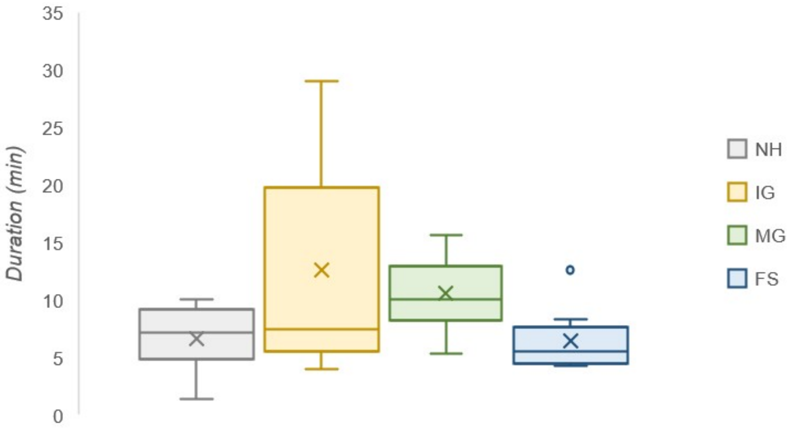


Figure 6.3: Comparison of reading duration for each method.

	IG	MG	FS
interlinear gloss	240	-	29
marginal gloss	-	339	254

Table 6.6: Average number of gloss triggered for each glossing method.

with lower L2 proficiency spent more time reading through by fixating longer on each word to trigger the gloss. In contrast, participants with higher proficiencies trigger fewer glosses, leading to a shorter reading duration. A graph showing a more detailed comparison of the total reading duration can be seen in Figure 6.3.

Average saccade velocity ranges from 300-400°/s during normal tasks, however, while reading in a foreign language, the saccadic velocity may lower to range between 150-300°/s [55]. By inspection, there are differences between the average saccadic velocities of each glossing method. The saccadic velocity for the MG and FS conditions are both higher than the NH and IG conditions, likely due to the presence of marginal glosses. The difference in saccadic velocity is more apparent when comparing the NH and MG case since the FS case also allows for interlinear and marginal glosses. For the MG condition, the range of saccadic velocities is larger, and the average speed is higher than the NH condition since the user needs to shift their gaze from the text to the right margin to activate the glosses.

Looking at the maximum duration of a fixation on a single gloss, we can see a clear difference between interlinear and marginal glosses due to the nature of the interaction methods, as shown in Figure 6.4. Interlinear glosses require users to fixate on a word for a threshold amount of time before triggering the appearance of the gloss. In contrast, marginal glosses only require a fixation of 200 ms to be rendered in the right margin. There are no signif-

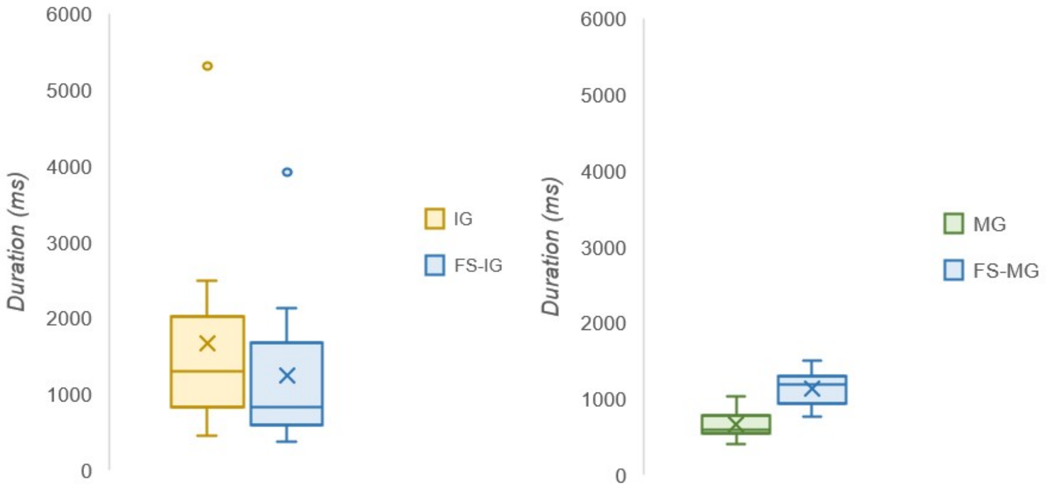


Figure 6.4: Maximum word fixation duration comparison for *(left)* interlinear glosses in the IG and FS systems, and for *(right)* marginal glosses in the MG and FS systems.

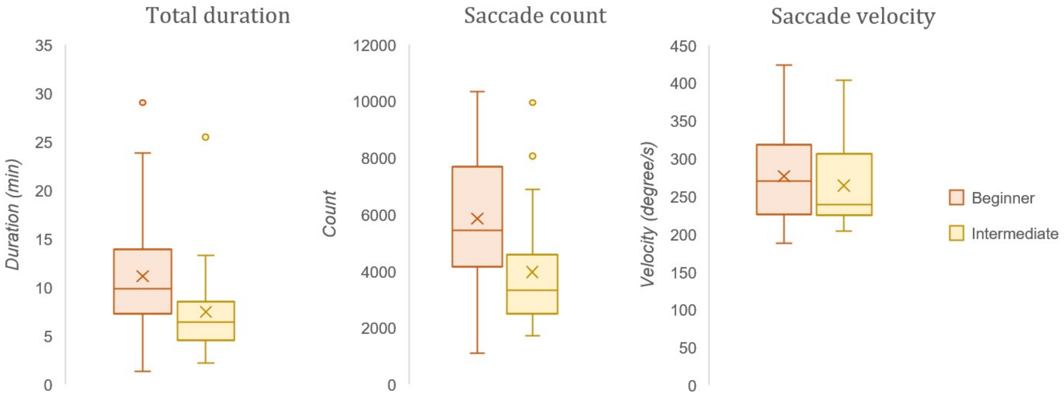


Figure 6.5: Differences in gaze behaviour of beginner and intermediate learners.

icant differences between interlinear gloss fixation duration in the IG and FS system, and the slightly lower fixation duration may be due to the users becoming more familiar with the system. In comparison to the range of fixation duration for interlinear glosses, the range for marginal glosses is much smaller for both MG and FS. For marginal glosses, there are also slight differences in the average maximum word fixation duration between MG and FS. Table 6.6 shows the average number of glosses triggered while using each glossing method. Overall, more marginal glosses were triggered by participants during separate gloss conditions, and when given the choice of marginal or interlinear, a significantly higher number of marginal glosses were used. This may be due to the quicker response time of marginal glosses, needing only a 200 ms fixation duration, compared to interlinear glosses that may have taken too long to trigger. Another factor which could favour marignal glosses is the fact that participants mentioned that they were unsure which word had interlinear glosses in the text, while with marginal glosses you could check if the word before or after had glosses. Additionally, marginal glosses do offer a higher amount of user agency, which some participants mentioned to prefer.

Proficiency level differences

There were high levels of variation between the saccade count and the total duration for each method. The saccade count and total duration have been shown to correlate to L2 proficiency in SLA eye tracking research [6, 7, 50, 88]. Since the reading comprehension scores shown in Table 6.7 compiles data for participants without regarding their L2 proficiency, the standard deviation is quite large. After separating the beginner (A1 and lower) and intermediate (high-A2, B1, B2) participants and examining the different gaze behaviour, we can see a clear difference between the two groups in Figure 6.5, in accord with prior works. Overall, participants who self-identified to be A1 and lower had higher averages and a broader range for total duration, number of saccades, and saccade velocity. Moreover, there are also distinct differences in how the two groups interacted with the text for each condition. Figure 6.6 shows the gaze movement of two participants, one beginner and one intermediate, reading texts. For the NH condition, we see that the beginner participants have more regressions, fixations, and saccades compared to the intermediate learner. Although there are some differences between the two, the differences are not obvious without close inspection. However, looking at the gaze movements during the IG and MG, we can see distinct differences in the gaze patterns while interacting with the glosses. From the gaze movement in IG, we can see that the beginner participants interacted with interlinear glosses much more frequently than the intermediate participant. Similarly, the beginner seems to have looked for the gloss of

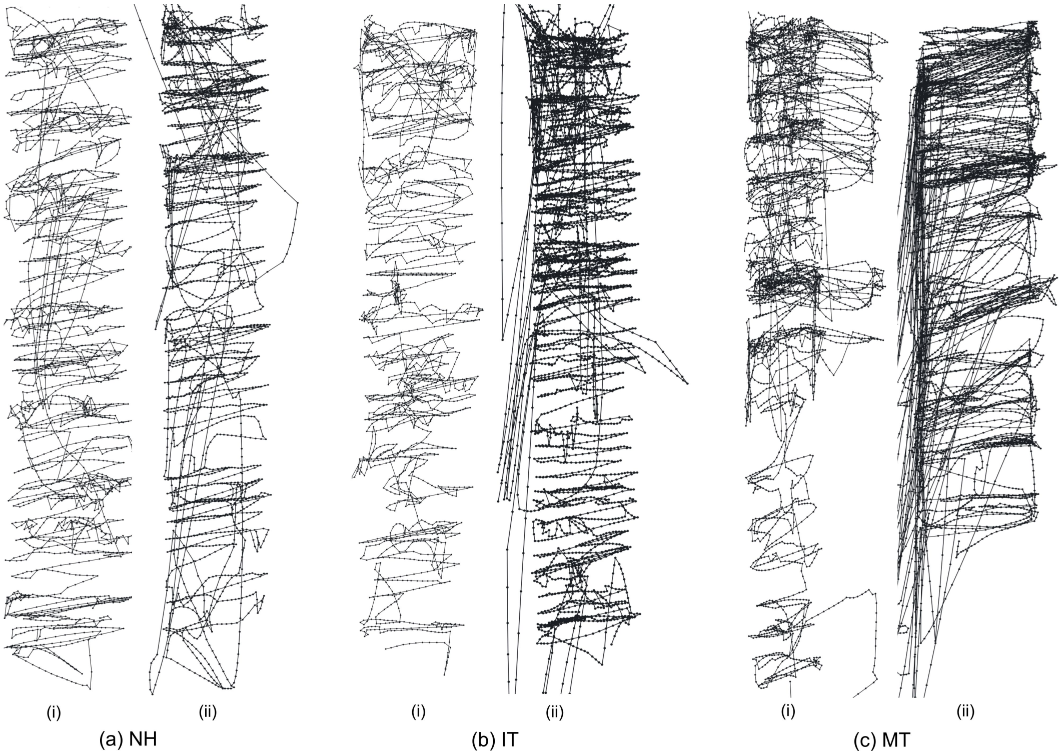


Figure 6.6: Gaze movement of (i) a beginner learner and (ii) an intermediate French learner reading with (a) no aid, (b) interlinear glosses, and (c) marginal glosses.

every word on a line, while the intermediate participant shows clear signs of looking at the margin only when needed. Additionally, as the intermediate reader continued through the text, they looked up fewer words.

6.2.2 Reading comprehension

After reading each article text, participants complete a reading comprehension quiz composed of a summary of the text in 2–3 sentences written in English, 4 multiple choice reading comprehension questions about the text, and 8 vocabulary matching questions with a dropdown list. Their score refers to the portion of correct answers for each section and is taken separately. A total quiz score consisting of the score from the multiple choice and vocabulary matching sections is calculated. The total score is used as a quantitative index, while the summary section is used as a qualitative index to gauge the participant's understanding of the text. The vocabulary matching section score is used to measure L2 vocabulary acquisition in relation to gloss usage.

	NH		**IG**		**MG**		**FS**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Multiple Choice	47.9	7.34	41.7	7.20	41.6	9.47	33.3	11.7
Vocabulary Matching	42.4	3.88	59.0	4.01	71.5	7.13	56.3	7.13
Total	41.8	3.12	52.8	3.43	60.9	5.98	52.3	6.59

Table 6.7: Average reading comprehension scores in percent for each gloss method.

Understanding the text

The summaries the participants wrote in the reading comprehension quiz varied in length and detail. There was no correlation between the length of the summary and comprehension. For the IG method, many summaries included specific terms that had been glossed in the text, and this phenomenon was not as detectable in the MG or FS methods. Using the total score, we conducted one-way ANOVA and found no significant differences (*p=.059, F=2.58*) between the different methods from each other and the baseline (NH). The mean and standard deviation of the scores for multiple choice questions, vocabulary matching, and total scores can be seen in Table 6.7.

When asked to rate their understanding of the text using the assistance provided using a Likert scale of *disagree (1)* to *agree (5)*, the distribution of the ratings is shown in Figure 6.7. From the graph, we can see that the opinion towards IG was split between really disliking the system or finding it somewhat useful to use in assisting them in understanding the test better. Participants noted during interviews that the help either appeared too slowly or too quickly, so the timing set for the gloss appearance did not work for them. In comparison, most participants felt neutral to positive towards the marginal gloss system. But overall, the method that seemed to help with understanding the most is FS. Although the timing of the interlinear gloss threshold boundaries update staying the same for IG and FS, participants felt that the timing of interlinear glosses functioned the best in FS. This may be due to being more familiar with the system and learning how to trigger interlinear glosses better. In addition, many felt that having marginal glosses as supplementary support was helpful.

L2 vocabulary acquisition

We conducted one-way ANOVA on the vocabulary matching score and found that there were significant (*p<.01, F=4.60*) differences between each glossing method. However, no two methods difference passed the Scheffe test. Further testing was conducted on the vocabulary matching section after separating beginner and intermediate results. One-way ANOVA

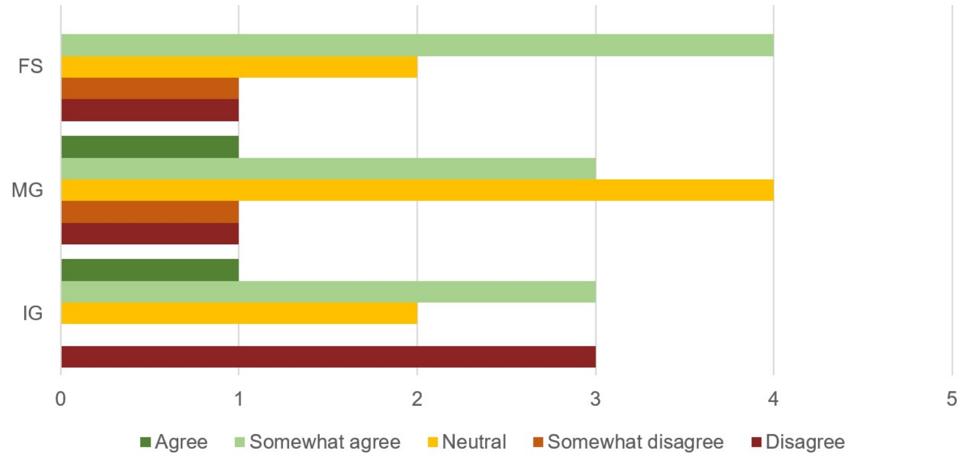


Figure 6.7: Rating each glossing method on whether they provided assistance for reading comprehension.

showed that beginner vocabulary score variance was not significant between each gloss presentation method (*p=.739, F=0.42*), but intermediate vocabulary scores reached significance (*p<.01, F=9.19*). Using the post-hoc Scheffe test, we obtained a Scheffe critical value of 0.243 and found that marginal glosses (*MD=0.438*) led to improvement in vocabulary scores when compared to no gloss (NH). In contrast, other glossing methods (IG and FS) did not lead to significant improvements (*MD<0.243*). Additionally, after conducting effect size analysis on the score in the vocabulary matching section, we found that having glosses has a large effect for intermediate learners ($\eta^2 = 0.385$) and medium effect for beginner learners ($\eta^2 = 0.050$).

6.2.3 Gloss position

We measured the participants' feelings about the usability of each glossing method using a combination of a 5-point Likert scale ranging from *disagree (1)* to *agree (5)* and short answer responses. To gather data on participants' preferences and compare the different gloss positions, we collected qualitative data through a semi-structured interview at the end of each participant study. Questions asked during the interview are detailed in Appendix A. Following the steps used to evaluate systems using the USE questionnaire, we summed all the questions that were positive aspects of the system and subtracted the negatively phrased items and obtained the following usability scores for each method. Table 6.8 shows the participant preference count for the glossing methods tested during the study and the usability

Method	Usability score	Preference count
IG	6.44	5
MG	6.48	8
FS	7.20	6

Table 6.8: Participant preferences for each glossing method.

scores. From interviews with participants, we found that there was a clear preference from one gloss position to the other depending on their L2 proficiency.

Starter and elementary level participant preferences

Participants who were level A1 in French found that marginal glosses caused headaches and eye strain because the text contained mostly words that they did not know and thus had to glance to the margins often. They felt that using marginal glosses distracted them from the text and caused them to lose their train of thought. Therefore, having interlinear glosses is useful to them as it provides them with help within the text area and allows them to comprehend the text better.

> *"It's a bit disruptive to have to look so far over to the right to see the definitions"*
> *- Participant 3*

> *"It was a lot better than the other one where you had to look to the right, as it did not draw your attention away from what you were reading"*
> *- Participant 6*

They found that the glosses appeared too slow when testing the IG method and appeared at a good speed towards the end of the text. Additionally, because most of the A1 level participants were only starting to learn French, they found that glosses of only the keywords from the text was insufficient and wanted to see entire sentences or paragraph translated. Many A2 level participants also found interlinear glosses to be preferable to marginal glosses due to easier access. Additionally, many were unsure which word had interlinear gloss available, so liked having marginal glosses where they would know exactly which word had glosses. Many noted that compared to true beginners, and they had a larger vocabulary base, so they did not need as much help compared to A1 level participants. For instance, one participant said,

> *"I was able to translate words and able to get a good idea on what the article was without having to translate the whole sentence"*
> *- Participant 12*

Generally, for A2 participants, having the L1 glosses was sufficient to aid them during the reading process. They also found that the threshold time for the gloss appearance was set at a good level but would have preferred a little longer, so they could try to guess what the meaning of the word first. Overall, for all beginner participants, having glosses interlinearly is the most preferred method for gloss positioning.

Intermediate level participant preferences

Compared to lower beginners, most intermediate (B1) levelled participants preferred marginal glosses. Many of the intermediate participants felt that interlinear glosses slowed down their reading, partially because they felt distracted by the glosses and, in other parts, due to the long fixation time needed to trigger the gloss. Additionally, the interlinear glosses provided help too easily and, similar to A2 levelled participants, the system did not give the participant enough time to process the word before providing help.

> *"...words would take too long to show up and it felt like I was straining my eyes"*
> *- Participant 11*

> *"I found that because the definition often showed over words that I already knew, I would end up looking at those and would miss the definition of words that I didn't know."*
> *- Participant 15*

During interviews, many intermediate-level participants specifically mentioned how much they preferred marginal glosses over interlinear glosses. However, a few intermediate participants felt that interlinear glosses were useful as they learned new words that they would have otherwise skipped over while reading. Besides preferring the position and control method of marginal glosses, intermediate-level participants felt that having an L1 gloss and L2 definition was helpful. They preferred the combination of the L2 definition over the L1 definition and wished that there were contextually relevant synonyms shown as well.

> *"It's great to see both the one-word English translation as well as the definition explained briefly in French!"*
> *- Participant 2*

In summary, intermediate-level participants found marginal glosses to be more useful and provide better help to them because they already had prior knowledge of the text and only had to look up certain words in the text. Therefore, having the marginal glosses, which offered a higher level of explicit control, was preferable.

Providing both interlinear and marginal glosses

The opinion on the full system, which provided both interlinear and marginal glosses, was split, and there was no consensus between the L2 proficiencies of the participants. Many participants felt hopeful about having both available before using. However, after using they felt that it was cumbersome. There was too much new information presented at one time, from the L2 text to all the glosses, which caused the participants to feel cognitive overload.

> *"Initially thought having both would be good but it was more annoying"*
> *- Participant 11*

> *"There was a lot going on to have both types of text flashing at me."*
> *- Participant 17*

On the contrary, some participants felt they preferred this system the most because if interlinear glosses did not show up, they would seek out marginal glosses. We found that participants who preferred higher levels of control over gloss appearance disliked the system, as it provided them with too much information. While participants who preferred the combined system primarily liked the ease of access provided by interlinear glosses and then the controllability of marginal glosses second.

6.2.4 Usability of the system

Overall, all participants felt that glossed texts were more helpful and that using gaze to control is better compared to looking up the definition in a dictionary or translation. However, gaze precision issues and inconsistencies in tracking made the system hard to use at times. The results from the usability questionnaire for all three glossing methods are shown in Figure 6.8. Based on the results, opinions on the usability of each glossing condition were divided, with no clear consensus on which system was the best.

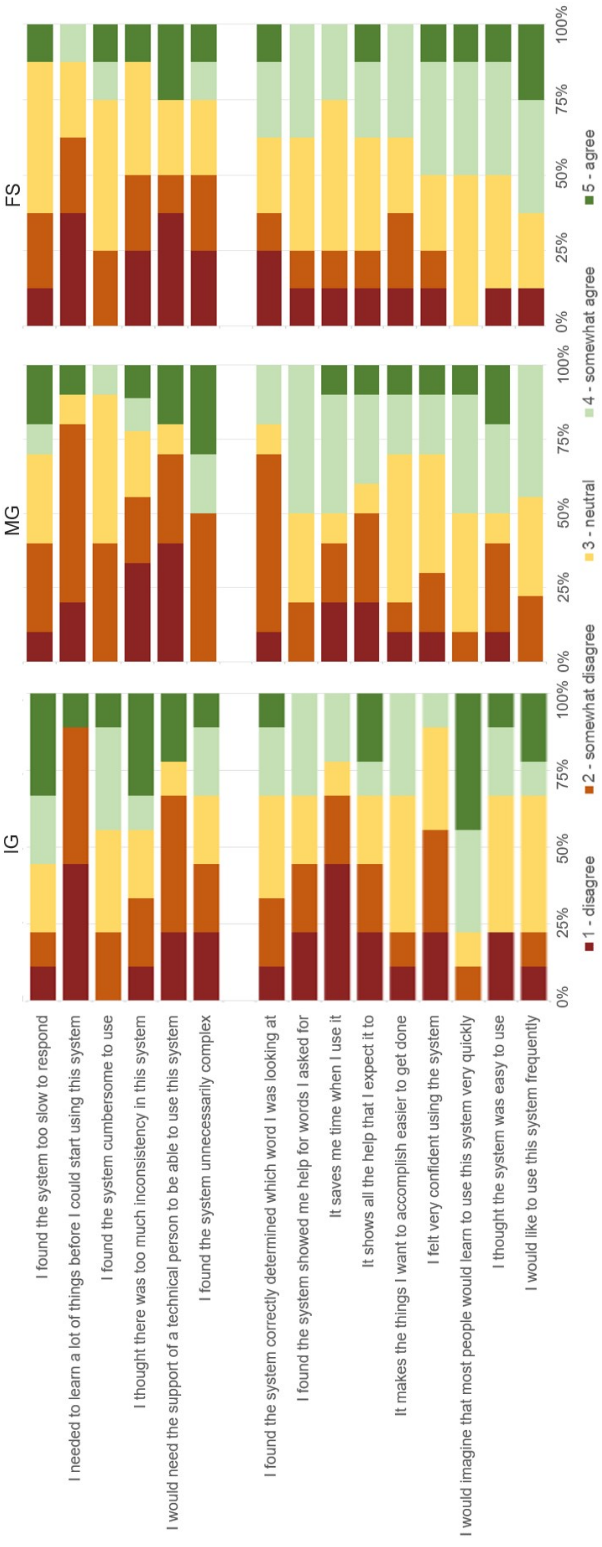


Figure 6.8: Usability questionnaire feedback for each system.

	Mean	*SD*	*Min*	*Max*
x	-12.66	34.06	-60	95
y	-13.98	49.29	-165	85

Table 6.9: Manual gaze adjustment summary.

Moreover, from the questionnaire, we found that participants felt that although there was a slight learning curve when troubleshooting gaze inaccuracies, the systems did not need much prior training. In addition to gaze precision difficulties, participants had varying opinions on the different types of assistance the gloss should provide and how many words should be included in a single L1 gloss.

Gaze accuracy and trigger threshold

Many made remarks about eye-tracking inaccuracies but saw potential in the glossing method. They felt that the eye tracking was too sensitive to head movements. Some participants expressed that if the interlinear glosses took too long to respond to words they wanted glosses for, and felt strained when staring at each word for a long duration. They felt that if the system responded faster, it would help their reading flow. The duration of the gloss threshold was continuously updated throughout each task. However, after inspecting values after each task, most of the values of the upper and lower boundaries for the trigger threshold fell within two bins. On the lower end, the lower threshold limit was set to between 150-200 ms and the upper was set to 300-350 ms. On the upper end, the lower threshold was around 300-400 ms and the upper threshold was between 650-800 ms. Only one participant fell between those ranges, with their range set to 490 ms and 600 ms for their lower and upper boundary.

During gaze data collection, we also collected the offset that each participant manually inputted to calibrate their gaze point before the start of each reading task. The horizontal and vertical offset values are plotted on a graph in Figure 6.9 to show the dispersion of the eye-tracking device. From the plot, we can see that there is no consensus on how much the offset should be since it varies from participant to participant. There is a more significant portion of offset values on the left side of the plot, showing a general trend of gaze points being too far right. Table 6.9 shows that the horizontal offset values ranged from -55 px to 35 px, while the vertical offset values ranged from -165 px to 35 px. Given that the vertical height of a monitor is less than its width, the vertical offset having such an extensive range of different offset values is a significant source of gaze inconsistency.

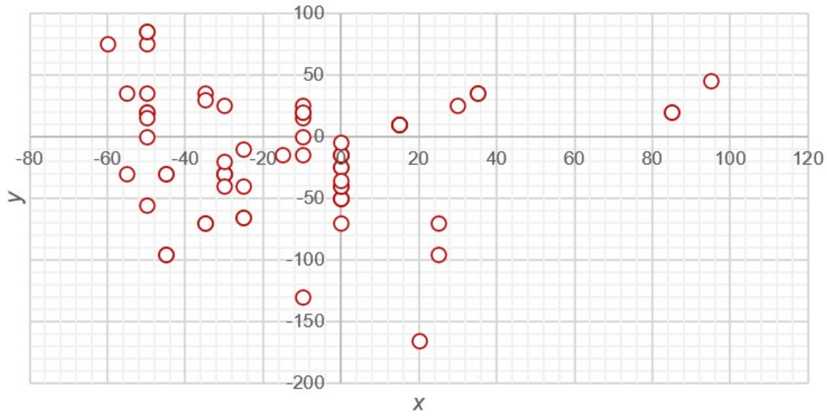


Figure 6.9: Plot of participant's manual gaze adjustments.

Variety of help provided by glosses

There was a split between the participants' preferences for the help provided by glosses which were dependent on their purpose behind reading. During interviews, participants who say they do not usually use reading for pleasure as methods prefer methods involving listening to L2 audio, like podcasts and watching TV shows. Those participants prefer reading to understand the general idea behind the article and are not concerned about personal vocabulary. Their purpose for reading is to gain more knowledge and immerse themselves in their L2. They also prefer to know what the words and pacing in the language sound like. Therefore, they favoured the IG or FS condition, where the system more implicitly provided them with aid within the text area.

In comparison, participants who mentioned that they use reading for pleasure to learn their L2 prefer marginal glosses, which would provide them with the definition in French. They liked to have their text area relatively minimal so that they could reread the text and use contextual information to guess the meaning of an unknown word. They favoured the ability to control marginal glosses explicitly and disliked having the system guess when they needed help. In the interview, when asked about the potential of including different varieties of help like synonyms and example sentences in L2, they were enthusiastic about including L2 phrases that use the glossed word and L1 translations.

Length of translated phrases

Participants also had varying opinions about how much text should be translated into a single gloss. Participants at lower proficiencies, A1, expressed preferences for entire sentences translated because the portion of the words they do not know in the text vastly outnumbers the terms they know. Therefore, being able to see the entire sentence translated and having the ability to compare the parallel text is more beneficial.

> *"I wish the system was able to translate sentence if looking at it for too long"*
> *- Participant 4*

For A2 and B1 levelled participants, some felt that having keywords translated in the text was sufficient; however, a more significant majority mentioned they would prefer if the gloss could provide more contextual information. Participants said they wanted information about immediate words surrounding the glossed word.

Preferred learning method

Participants had varied preferences for what they used for learning new languages. Some participants preferred using curated instructional content like YouTube videos created by language instructors, podcasts focused on teaching French, and mobile applications like Duolingo. In comparison, other participants enjoyed more passive learning methods and immersing themselves in their L2. We found that participants who were actively seeking French instructional content were looking to increase their French knowledge and had a clear goal for why they were learning. Some learned because of school, and others for an upcoming trip to a French-speaking place. In contrast, those that were using passive learning materials were more focused on maintaining French knowledge learned previously. Additionally, when asked about if and what types of reading comprehension questions participants prefer having at the end of the text, most participants said that they do prefer having a few questions at the end to check their comprehension of the text. We found that even participants who were focused on getting through the text would prefer having a quick multiple-choice question at the end.

7 Discussion

In this thesis, we first explored resources available for French, which can be used to easily annotate text to use in web applications. Using those resources, we selected articles of similar complexity levels and annotated words and phrases with glosses using an additional dictionary API. The articles were then used in a gaze-enabled system which provided two types of glosses: interlinear and marginal. We conducted a participant study to evaluate the different gloss positions and found that there were differences in their reception due to L2 proficiency and personal preferences. There were also insights such as when gaze should be used during reading assistance, the threshold for triggering glosses, types of gaze gestures to allow for use, and when to introduce other input modalities like mouse clicks. We compiled a series of lessons learned from our findings and suggestions for future researchers working with gaze-enabled systems and glosses for L2 reading aids. In addition, we will discuss the potential applications and future directions that could benefit from the findings outlined in this thesis.

7.1 Gaze precision and accuracy

A key feature of eye movements distinguishing them from other input types is that they remain constantly active. Gaze-enabled systems use gaze in two main ways: implicitly and explicitly. Systems that use gaze implicitly determine preferences and information about users based on what and how they interact with elements in a system. There is always a trade-off between complexity and preciseness when working with gaze-enabled systems that use explicit gaze commands. Gaze commands are often based on dwell time, or fixation duration, the idea seems promising in theory, but when put into practice, we encounter the Midas Touch problem. The Midas Touch problem exists because every fixation on an interface element in a gaze-enabled could be treated as an activation event despite that not being the user's intention. To deal with this problem, many existing systems use gaze gestures—a set of predefined gaze movements—to trigger a response from the system [67]. However, the problem that arises with gaze gestures is the rise in complexity of the gesture and the large learning curve. Additionally, during reading L2 text which already requires higher cognitive processing, performing additional gaze gestures could lead to cognitive overload.

Our third research question of this thesis pertains to the preferred gaze intervention method for users of gaze-enabled reading aid systems. We found that in order to address the intervention method preferred, we need to first consider the limitations of eye-tracking precision and accuracy. Prior works have made remarks about the inconsistency of commercially available eye-tracking devices and how their accuracy deviates from manufacturer specifications [31]. Moreover, other works have worked with sentence-level L2 glosses due to accuracy issues with widely available eye-tracking devices [44]. Gaze inaccuracy has two sources: one from the actual hardware used to capture the gaze position and the other from innate human gaze movement error. As we design technology for human use, we must acknowledge that humans may use their eyes in ways that are beyond our control. Therefore, it's important to build systems that anticipate and minimize errors. If we want to work with eye gaze, we will need to use methods to compensate for the inaccuracy of the eye-tracker, which only provides us with one to two degrees of visual angle precision. While our techniques were not without fault and resulted in some errors based on gaze position, we improved precision through the following measures during implementation: filtering the gaze, introducing variable and manual offsets, and tracking the line that the user was on.

We initially attempted to use a model of gaze during reading, presented in iDict [46], to detect which line the reader was on. This model simplifies reading a single line of text as a series of events involving fixations, medium-length saccades, several short regressions within a line of text, and a vertical shift upwards as the reader moves horizontally along the line. Although applicable to some extent, in the case of an L2 learner, there are many pauses when moving to the next line of text. Hence, we needed to adjust the algorithm to be triggered every time a regression occurs with a large enough vertical shift. By determining the line users are on, we can also filter out words that are not on the line, so we can better determine which specific word the user was fixated on and detect when the fixation had passed the word's threshold. The threshold duration was calculated using a word's frequency in the FLElex dataset, with parameters for the lower and upper threshold duration boundaries adjusted for each user. The parameters were tuned as users read, however, due to the short duration of the study and wide range of participant proficiency, we could not come to a conclusion about what the threshold boundary values should be set to or how they should be properly updated to best suit learner use.

During each intervention condition testing, we allowed participants to view where their gaze was at any time so that they could troubleshoot when glosses did not show up. Troubleshooting involved resetting the system's line detection algorithm by rerunning the line setting algorithm (Algorithm 4) with *start* set to 1 and *end* set to the length of the text, so that the system would re-detect which line the user's gaze was on. Participants were quickly able to learn how to use the system troubleshooting tools without much assistance from the researchers, as shown in the usability feedback questionnaire results where most participants felt that there was not a large learning curve to using the system and did not feel a need for technical assistance while using it. Participants mentioned that adapting to a gaze-based system requires some time as its functionalities are more unfamiliar. Although when errors occurred too often, causing participants to continuously check which line the system thought they were reading, the system increased the amount of work put in to read L2 text and caused frustration. We did not account for this during the system design as it falls outside of the scope of this project. However, this is a major flaw in the design and concerns about the ease of troubleshooting eye tracking methods need to be addressed in future works involving gaze-enabled systems.

7.2 Interplay of novel input modalities and ease of use

Developing new interaction methods with different input modalities is often about creating a well-balanced application which uses new input modalities to enhance the entire experience. We hope to show with our system that using gaze-enabled reading technology is a possibility that can be further explored. Eye movements as input for applications should not assume that eye movements follow some idealized movement as each person interacts with applications in unique ways. However, this is not to say that generalizations about gaze movements cannot be made. For example, we found that there are general trends in what beginners prefer in terms of reading assistance compared to intermediate learners. When users know that a system has gaze enabled, they will assume that their gaze will activate something when they interact with system elements. During our study, even when participants went through the tutorial for each system quickly, they were able to quickly how realize that the system responded to their gaze. After knowing that the system responds to their gaze, the participants will apply their preconceived ideas about how it should respond. As such, participants either said that they thought every word had a gloss or every keyword had a gloss. Inconsistencies in the annotation led to participants feeling confused about what actually had help provided for it, which was one of the reasons why some participants preferred the marginal method as it will show the most recently looked at word with a gloss along with the nearest neighbour words with glosses. Participants noted that the margins served as a check to see if the words they were wondering about even had a gloss. Additionally, when the glosses for certain words did not show up, the participant thought that it was a gaze error and said that they would often switch the gaze cursor on to see where they were looking at.

Thus, one may think that increasing the number of glosses may be the solution, however, the participant's feedback about the interlinear and combination system suggests that there would be additional problems if glosses were provided for everything. The problem with providing interlinear glosses for every word, as requested by some participants, is information overload. Beginner participants have often suggested including sentence-level translations of texts as glosses which could serve as a parallel text for them to reference as they read. Although showing more L1 gloss may be useful to users who are just starting to learn their L2, they may rely too heavily on the gloss and forgo the L2 text altogether. Therefore, it may be more beneficial to read texts which are appropriately levelled for them, as suggested by Chiang et al. [24]. Since the goal of reading assistance is to help learners read more text and aid comprehension, having the translations might motivate them to read more for pleasure in their L2. We believe that it will be beneficial for reading aid systems to alert readers when they are reading texts which contain too many unknown words and is above their current L2 proficiency level. Reading aids should be aware of their user's L2 proficiency and provide help that is suitable for them.

Additionally, having a system that uses gaze for the sake of increasing the number of input modalities may actually hinder usability. There should be a consideration of when the gaze is an appropriate input for devices and when it is not. Systems that allow for gaze interventions should always allow users to set their preferences and limits for when they want their gaze to be used as a source of input. Many of the participants noted that they would have liked to be able to click on the words that they wanted help with since some words are short and would require more fine-tuned precision to maintain fixate on the word. Moreover, staring at a singular point to trigger a word's gloss breaks the reading flow and is not as implicit of a method as we intended the design to be.

7.3 Gloss position

Through our study, it has come to our realization that user preference for gloss position is due to a combination of L2 proficiency and personal preferences. For most lower proficiency L2 learners, we found that there was no lower proficiency L2 learner which preferred the marginal gloss only system, their preferences were either interlinear glosses only or the system with both interlinear and marginal glosses. We found that interlinear glosses are preferred as it leads to lower cognitive load compared to marginal glosses since the user's gaze does not need to be shifted away from the text they are reading. Moreover, additional information like L2 definitions and L2 synonyms are useless to lower levelled learners as the information provided is too high level for them. In contrast, intermediate learners mainly prefer the marginal only gloss system as it is less distracting and allows them a higher degree of control with additional information to provide them with more contextual information. No matter which gloss position is used in a system, one must be aware of the benefits and implementation issues that each method comes with.

7.3.1 Usability of marginal glosses

Through our results with the vocabulary matching portion of the reading comprehension quiz, we can see that marginal gloss could have a significant effect on vocabulary learning during the immediate posttest. Other research has suggested that glosses also increase recall during delayed posttests. So, in terms of the position that helps to learn the most, marginal glosses alone perform superior compared to interlinear glosses and the combination of both. Marginal glosses are useful for learners who are interested in active learning while learning their L2 and aid them in looking up definitions of words in their L2. For these users, a larger variety of help in marginal glosses would be beneficial. Showing too much help in the text area may overwhelm the users, but when the assistance is isolated in the margin, it allows users to have higher amounts of control over what and when they receive help.

Despite its benefits, there were a few design issues with participants which could be improved in the system. The right margins were set to be around 400 px wide with 50 px margins making the entire size of the margin 450 px, this was initially designed so that the text area would not be affected by the presence of glosses and reduce input error. However, due to the design choice of making the gloss harder to trigger, the distance the eye must travel to move from the starting word on a line of text to the right margin is large. This large distance introduces additional vertical variance in the gaze position, increasing the chance and a higher chance of accidental fixation on other words in the text during the gaze shift to the margins. Additionally, when the user shifts their gaze back from the margin to the text, the increased vertical variance and loss of text context would cause them to forget and lose their place in line. Despite our efforts to highlight the glossed word in the text, on return movement from the margin, the user may accidentally fixate on another word causing the highlight to fade. Many mentioned how the right margin would be difficult to use in actual practice because of the large necessary saccade needed to activate it.

7.3.2 Usability of interlinear glosses

The increased accessibility and quick response of interlinear glosses make them useful for learners who are less inclined to up unknown words during reading. Having the glosses in the text, one is reading also reduces the need to remove themselves from the text and lose their focus. It encourages people who would not otherwise read L2 text to interact with it more. If the system's goal is to aid readers who want to read through L2 text as quickly as possible and understand the general idea presented, then interlinear glosses are preferred. However, as mentioned previously, if the goal was speed, then learners may be tempted to translate the entire article to their L1. Ultimately, the effectiveness and usability of interlinear glosses rely on the level of effort that learners are willing to invest in order to enhance their understanding of their second language.

The duration of the fixation, or dwell time, needed to trigger glosses need to be explored further. The threshold time needed to trigger an interlinear gloss was variable and updated during reading. Initially, we set the threshold's lower and upper boundaries to 350 and 600 ms. However, these values were too high for A1 level participants, who lowered the boundaries to the minimum allowed values, and too low for A2 and B1 level participants, who found it distracting to have the glosses appear so fast. Further work must be done to determine a better method for updating the boundary values. Past research by Karolus et al. has determined that C2 and A1 levelled readers have an average fixation duration of 0.8 s and 1.04 s, respectively. In comparison, during our studies, we found that the fixation duration of French L2 learners who speak English fluently was on average, shorter. This may be due to the similarity in structure of English and French, which allows learners to understand the text through cognates. Therefore, researchers in this area should be aware of the native language of their participants and explore the average fixation duration for their target audience.

7.4 Suggestions from research findings

To build a usable system, the system design should involve the appropriate level of proactivity and transparency for the system. Especially for gaze-enabled systems, the system state should be clearly visible since occasional misinterpretation of gaze movement is inevitable. If the user interacts with static objects, the system should be able to detect and provide appropriate feedback. The implicit gaze intervention techniques, like showing interlinear glosses, should be more customizable and allow users to make the decision on whether to show them or not. This will reduce triggering undesired functionalities and help users focus on their main goal. Finding the perfect balance between automated and user-directed actions depends on personal preference, and therefore, the user should have the ability to adjust the system's sensitivity accordingly. There is also a threshold of error that users will accept to be part of the system since gaze movements are hard to control. However, the system to recover from errors and allow users to take back control should be made easily available. It is important to ensure that feedback from the system is easy to understand and aligns with the common understanding of gaze functions. It is important for the user to feel in control at all times.

It is impossible to build a system that suits all learner needs at every level because learner needs at every level is different. Every language learner has characteristics that affect their language learning progress and preferences toward learning methods. During our participant study, we found that many of the participants preferred listening to content in their L2 rather than reading because it is simply more practical to learn how to understand spoken language. There could be a possibility of including additional input modalities like sound clips in the system, however, that is outside the scope of this thesis.

The design of each gaze intervention method should be evaluated not only for effectiveness compared to other gaze control methods but also more commonly seen control methods like using mouse clicks. For practical use, actions which require accuracy and precision may be better left for input methods that reduce error, like using a mouse. From our study, gaze is more useful for informing the system about the user state and gathering information about them to use for later purposes. It can be used for interactable glosses while reading, however, the option of using a mouse should be available to the user. Additionally, the distance from the text to the right margin should also be considered when designing for gaze-enabled systems. Requiring users to move a large distance using their eye causes strain and people's gaze do not move perfectly horizontally. Hence, the back-and-forth movement introduces additional vertical errors that must be accounted for. For interlinear glosses, the window size for words included in a single interlinear gloss should be customizable. Users should have the option to translate keywords and up to a certain number of words in their vicinity. This would allow for real-time applications to provide more accurate assistance for not only singular words but also phrases. Furthermore, the controllability of glosses is important. There should be three preset levels for threshold boundaries, each catering to a different L2 proficiency level. Beginners should have lower threshold boundaries compared to intermediate learners.

7.5 Limitations and future directions

This thesis scope included examining gloss positions and the usability of gaze intervention techniques when reading French text as a French L2 learner. The timeline of a master's thesis limits the duration and amount of studies conducted to explore our research questions. As the goal of this project was to compare the usability of gaze-enabled gloss intervention methods, we did not focus our efforts on system usability as it falls outside the scope of this particular project. As thus, we did not employ a user-centred design approach or use codesign sessions to detect possible usability issues when creating the system. Additionally, during our study, we gathered participants with varying language proficiencies. For a more comprehensive understanding of language proficiency preferences, it is recommended to conduct a detailed study with a more strict separation of language levels. In the text processing chapter, we've shown how to utilize available tools to annotate French text, however, were not able to fully explore using the available tools to annotate French text in real-time for web applications. Therefore, future research can examine the feasibility of using those tools and creating glosses for French texts automatically.

Furthermore, future research can build off this thesis to explore different methods for improving accuracy in gaze position for more precise tracking during reading tasks. There also needs to be further explorations of the ideal fixation duration as a function of reading speed for L2 learners at different language proficiencies in an application that uses gaze as a source of implicit control. In our system, we were able to track which words people sought glosses for and how many refixations they made on the same words. Future works can combine detecting words of interest with generative text tools to automatically generate personalized reading comprehension quizzes. Additionally, with advances in the computational powers of commercial computers, there is the possibility of running pre-trained machine learning models to automatically detect a user's L2 proficiency as they are reading. To realize this, work must be done to gather data and generalize gaze feature information for readers at different proficiency levels. Although, the data exists to a limited extent, including only the average fixation duration of individuals at A1 and C2 proficiencies in English [50], more multilingual content should be explored.

As our work shows the potential of gaze intervention techniques adapting for explicit and implicit commands while reading, future work can explore using different forms of eye tracking like webcam-based or mobile phone infrared eye tracking. Although, as shown in this work and previous works using commercial eye trackers [31, 44] that accuracy issues with devices specifically created for eye tracking still exists, these problems will be aggravated when using other devices not intended for eye-tracking purposes. However, the potential of gaze-enabled applications for everyday devices is exciting and opens the door to new innovative research. Gaze-enabled reading aid technology has implications outside of language learning and could be used in the assistive context for people with reading difficulties or limited mobility.

8 Conclusion

Although one of the best ways to acquire vocabulary in one's L2 is by reading, there are problems with the notion that vocabulary growth does not require effort. It is challenging to infer the meaning and learn new words from reading unless one is already familiar with most of the vocabulary in the text. In addition, people prefer certain types of text, and it may be that a reader who mainly reads fiction has little chance of learning words essential for writing academic papers, for example. Electronic glosses can help readers read texts in their L2 which may be inaccessible to them due to the number of unknown words. We initially started this thesis project to explore how gaze can be used as an additional input mode to determine when L2 learners encounter difficult words implicitly. However, as we delved into the project, we discovered a gap in knowledge investigating user preference in gloss placement and gaze intervention techniques during reading. Our study found that language proficiency and personal preferences affect which type of control modality and gloss positioning are the most useful for a person. Therefore, a reading aid needs to provide more personalization for users through customization options for modes and adapt continuously through their actions.

A Appendix A: Participant Study Materials

A.1 Pre-study questionnaire

Shortened and adapted from the Language Experience and Proficiency Questionnaire (LEAP-Q) [64]

1. Please list all the languages you know in order of dominance and your overall level of proficiency on a scale of 1-5 (1-beginner, 2-elementary, 3-intermediate, 4-upper-intermediate, 5-native/native–like)
   You can use the CEFR self-assessment table provided by the Council of Europe to help identify your level

2. Please list all the languages you know in order of acquisition (i.e. your native language first):

3. When did you start learning French?

4. When did you start reading French?

5. Please select your level of proficiency in speaking, understanding, and reading in French on a scale of 1-5 (1-beginner, 2-elementary, 3-intermediate, 4-upper-intermediate, 5-native/native–like)

   a) Speaking

   b) Understanding spoken language

   c) Reading

6. On a scale of 0-10, please select how much the following factors contributed or aided to you learning French:

a) Interacting with friends and family

b) Consuming media (includes watching videos, music/podcasts, and social media content)

c) Reading

d) Language-lab/self-instruction

e) Language learning app

A.2 Reading comprehension questions

Sample reading comprehension questions with answers for Article B.

Part I: Summary
Write a 2-3 sentence summary of the article in English

Part II: Multiple choice reading comprehension questions

1. Qu'est-ce que le Carnaval de Québec ?

    a) Un festival de musique en été

    b) Un carnaval d'hiver à Gaspé

    c) Un carnaval d'hiver à Québec

    d) Un carnaval de printemps à Montréal

    ANS: c) Un carnaval d'hiver à Québec

2. Quel est l'emblème du Carnaval de Québec ?

    a) Un caribou

    b) Un castor

    c) Un bonhomme de neige

    d) Un homme avec une ceinture fléchée

    ANS: c) Un bonhomme de neige

3. Quel est le but principal du Carnaval de Québec ?

    a) Fêter l'hiver

   b) Attirer les touristes

   c) Collecter de l’argent pour la ville de Québec

   d) Célébrer la culture française

   ANS: b) Attirer les touristes

4. Quelles sont les traditions qui ne contribuent pas à l’ambiance du Carnaval de Québec?

   a) Le port de la ceinture fléchée

   b) La pêche sur glace

   c) Les chansons carnavalesques

   d) Le port du rouge

   ANS: b) La pêche sur glace

Part III: Vocabulary matching
Match the following French words to their English equivalent

1. *carême* lent
2. *remontent* go back
3. *réchauffer* warm up
4. *mondiale* world
5. *devenu* became
6. *événement* event
7. *cible* target
8. *syndique* unionize

A.3 Usability questionnaire

Usability questions for each technique (based on the USE [61] and SUS [15])
Rate the following from 1-5 (1 - disagree, 5 - agree)

1. I understood everything that was said in the article with the assistance provided
2. I have prior knowledge of the topic covered in the article

3. I knew almost all the words tested in the vocabulary matching quiz prior to this study
4. I would like to use this system frequently
5. I found the system unnecessarily complex
6. I thought the system was easy to use
7. I would need the support of a technical person to be able to use this system
8. I thought there was too much inconsistency in this system
9. I would imagine that most people would learn to use this system very quickly
10. I found the system cumbersome to use
11. I felt very confident using the system
12. I needed to learn a lot of things before I could start using this system
13. It makes the things I want to accomplish easier to get done
14. It shows all the help that I expect it to
15. It saves me time when I use it
16. I found the system showed me help for words I asked for
17. I found the system too slow to respond
18. I found the system correctly determined which word I was looking at

Feedback questions

1. Did you feel like this intervention technique supported you while reading?
2. Was the intervention method distracting? If so, how?
3. What do you wish was included in this system that was not?
4. Any suggestions on how this system can be improved?
5. Other comments?

A.4 Interview questions

Application usage questions

1. Which system did you like using the most? Why?
2. What do you feel are the most important to you: variety of help (e.g. definitions, translations, synonyms, etc.), automatic help, or help embedded within the text itself? Why?
3. Was there an article that you enjoyed reading the most? (Article A, Article B, Article C, Article D)
4. How do you think having an article that you enjoy reading affects how you read? Would it motivate you to look up more words?
5. Did you think that having the translation help you? If yes, how? If not, what would be better?
6. Would you rather an application that quizzes you to get you to learn vocabulary? If no, why?
7. When do you think you would use an application like the ones you tested today?
8. Who do you think would benefit the most from applications like these? Beginner, Intermediate, Advanced?

Feedback questions

1. What are some methods that you use when reading in languages you are learning?
2. What are some applications that you currently use for language learning?
3. Are there any language learning applications that you have seen or heard about but not tried?
4. Other comments?

Bibliography

1. A. F. AbuSeileek. "Hypermedia annotation presentation: The effect of location and type on the EFL learners' achievement in reading comprehension and vocabulary acquisition". *Computers & Education* 57, 2011, pp. 1281–1291. DOI: 10.1016/j.compedu.2011.01.011.

2. A. F. M. AbuSeileek. "Hypermedia Annotation Presentation: Learners' Preferences and Effect on EFL Reading Comprehension and Vocabulary Acquisition". *CALICO* 25, 2 2008, pp. 260–275.

3. S. Ahn, C. Kelton, A. Balasubramanian, and G. Zelinsky. "Towards Predicting Reading Comprehension From Gaze Behavior". *ACM Symp. on Eye Tracking Research and Applications (ETRA)*. Short Papers. Association for Computing Machinery (ACM), Stuttgart, Germany, 2020. ISBN: 9781450371346. DOI: 10.1145/3379156.3391335.

4. P. J. L. Arnaud and H. Béjoint. "Vocabulary and Applied Linguistics". 1992.

5. S. Arnott, M. Masson, and S. Lapkin. "Exploring trends in 21st century Canadian K-12 French as second language research: A research synthesis". *Canadian Journal of Applied Linguistics* 22:1, 2019.

6. O. Augereau, H. Fujiyoshi, and K. Kise. "Towards an automated estimation of English skill via TOEIC score based on reading analysis". *Proc. of Int. Conf. on Pattern Recognition* 0, 2016, pp. 1285–1290. DOI: 10.1109/ICPR.2016.7899814.

7. O. Augereau, H. Fujiyoshi, K. Kunze, and K. Kise. "Estimation of english skill with a mobile eye tracker". *Proc. of the ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing (UbiComp)*, 2016, pp. 1777–1781. DOI: 10.1145/2968219.2968275.

8. O. Augereau, C. Jacquet, K. Kise, and N. Journet. "Vocabulometer: A web platform for document and reader mutual analysis". *Proc. of IAPR Int. Workshop on Document Analysis Systems (DAS)*, 2018, pp. 109–114. DOI: 10.1109/DAS.2018.59.

9. D. Bauer and C. Cavonius. "Improving the legibility of visual display units through contrast reversal". *Ergonomic aspects of visual display terminals*, 1980, pp. 137–142.

10. J. R. Bergstrom and A. Schall. *Eye tracking in user experience design*. Elsevier, 2014.

11. Y. Berzak, C. Nakamura, S. Flynn, and B. Katz. “Predicting Native Language from Gaze”. *Proc. of the Conf. of the Association for Computational Linguistics (ACL)* 1, 2017, pp. 541–551.

12. R. Biedert, G. Buscher, S. Schwarz, J. Hees, and A. Dengel. “Text 2.0”. *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2010, pp. 4003–4008. DOI: 10.1145/1753846.1754093.

13. J. Bingel, G. H. Paetzold, and A. Søgaard. “Lexi: A tool for adaptive, personalized text simplification”. *Proc. of the Int. Conf. on Computational Linguistics*, 1 2018, pp. 245–258.

14. I. V. Blinnikova, M. D. Rabeson, and A. I. Izmalkova. “Eye movements and word recognition during visual semantic search: Differences between expert and novice language learners”. *Psychology in Russia: State of the Art* 12, 1 2019, pp. 129–146. ISSN: 20746857. DOI: 10.11621/pir.2019.0110.

15. J. Brooke. “SUS: A quick and dirty usability scale”. *Usability Eval. Ind.* 189, 1995.

16. A. Bulling, J. A. Ward, H. Gellersen, G. Tröster, J. Karolus, P. W. Wozniak, L. L. Chuang, and A. Schmidt. “Robust recognition of reading activity in transit using wearable electrooculography”. *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems* 2017-May, January 2015 2008, pp. 19–37. ISSN: 03029743. DOI: 10.1007/978-3-540-79576-6_2.

17. G. Buscher, A. Dengel, L. V. Eist, and F. Mittag. “Eye movements as implicit relevance feedback”. 2008, pp. 2991–2996. DOI: 10.1145/1358628.1358796.

18. S. Canada. *While English and French are still the main languages spoken in Canada, the country’s linguistic diversity continues to grow*. 2022. URL: https://www150.statcan.gc.ca/n1/daily-quotidien/220817/dq220817a-eng.htm.

19. C. des niveaux de compétence linguistique canadiens. *NCLC : Français langue seconde pour adultes*. 3rd ed. 2018. ISBN: 978-1-100-99313-3. URL: https://www.language.ca/product/pdf-f-001-nclc-francais-langue-seconde-pour-adultes/.

20. J. B. Carroll. “Chapter I: Linguistics and the Psychology of Language”. *Review of Educational Research* 34:2, 1964, pp. 119–126. DOI: 10.3102/00346543034002119.

21. I. J. Chen. “Hypertext glosses for foreign language reading comprehension and vocabulary acquisition: effects of assessment methods”. 29, 2 2014, pp. 413–426. ISSN: 17443210. DOI: 10.1080/09588221.2014.983935.

22. I. J. Chen and J. C. Yen. "Hypertext annotation: Effects of presentation formats and learner proficiency on reading comprehension and vocabulary learning in foreign languages". *Computers & Education* 63, 2013, pp. 416–423. ISSN: 0360-1315. DOI: 10.1016/J.COMPEDU.2013.01.005.

23. Y.-H. Cheng and R. L. Good. "L1 Glosses: Effects on EFL Learners' Reading Comprehension and Vocabulary Retention." *Reading in a Foreign Language* 21, 2 2009, pp. 119–142. ISSN: -1539-057.

24. M. H. Chiang. "Effects of varying text difficulty levels on second language (L2) reading attitudes and reading comprehension". *J. of Research in Reading* 39, 4 2016, pp. 448–468. ISSN: 14679817. DOI: 10.1111/1467-9817.12049.

25. C. Clifton, F. Ferreira, J. M. Henderson, A. W. Inhoff, S. P. Liversedge, E. D. Reichle, and E. R. Schotter. "Eye movements in reading and information processing: Keith Rayner's 40 year legacy q". *J. of Memory and Language* 86, 2016, pp. 1–19. DOI: 10.1016/j.jml.2015.07.004.

26. J. Coady and T. Huckin. *Second language vocabulary acquisition: A rationale for pedagogy*. Cambridge University Press, 1997.

27. K. Conklin and A. Pellicer-Sánchez. "Using eye-tracking in applied linguistics and second language research". *Second Language Research* 32, 3 2016, pp. 453–467. ISSN: 02676583. DOI: 10.1177/0267658316637401.

28. J. Ding, B. Zhao, Y. Huang, Y. Wang, and Y. Shi. "GazeReader: Detecting Unknown Word Using Webcam for English as a Second Language (ESL) Learners". *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. Extended Abstracts. Association for Computing Machinery (ACM), Hamburg, Germany, 2023. ISBN: 9781450394222. DOI: 10.1145/3544549.3585790.

29. C. of Europe. *The CEFR Levels*. 2023. URL: https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions.

30. "Facilita: Reading assistance for low-literacy readers". *Proc. of SIGDOC the Int. Conf. on Design of Communication*, 2009, pp. 29–36. DOI: 10.1145/1621995.1622002.

31. A. M. Feit, S. Williams, A. Toledo, A. Paradiso, H. Kulkarni, S. Kane, and M. R. Morris. "Toward Everyday Gaze Input: Accuracy and Precision of Eye Tracking and Implications for Design". *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. Association for Computing Machinery (ACM), Denver, Colorado, USA, 2017, pp. 1118–1130. ISBN: 9781450346559. DOI: 10.1145/3025453.3025599.

32. T. François, N. Gala, P. Watrin, and C. Fairon. “FLELex: a graded lexical resource for French foreign learners”. *Int. Conf. on Language Resources and Evaluation (LREC)*. 2014.

33. U. Garain, O. Pandit, O. Augereau, A. Okoso, and K. Kise. “Identification of Reader Specific Difficult Words by Analyzing Eye Gaze and Document Content”. *Proc. of the Int. Conf. on Document Analysis and Recognition (ICDAR)* 1, 2017, pp. 1346–1351. DOI: 10.1109/ICDAR.2017.221.

34. T. Gedeon, B. S. Mendis, L. Copeland, and S. Mendis. “Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error”. *Article in Artificial Intelligence Research* 3, 3 2014. DOI: 10.5430/air.v3n3p35.

35. V. A. Ginsburgh, I. Ortuño Ortı⊠n, and S. Weber. “Why do people learn foreign languages?”, 2004.

36. R. Godwin-Jones. “Riding the digital wilds: Learner autonomy and informal language learning”, 2019.

37. A. V. González-Garduño, G. Garduño, and A. Søgaard. “Using gaze to predict text readability”, 2017, pp. 438–443.

38. S. Gooding, E. Kochmar, S. M. Yimam, and C. Biemann. “Word Complexity is in the Eye of the Beholder”, 2021, pp. 4439–4449. DOI: 10.18653/V1/2021.NAACL-MAIN.351.

39. A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. “Coh-Metrix: Analysis of text on cohesion and language”. *Behavior research methods, instruments, & computers* 36:2, 2004, pp. 193–202.

40. A. C. Graesser, D. Greenberg, A. Olney, and M. W. Lovett. “Educational Technologies that Support Reading Comprehension for Adults Who Have Low Literacy Skills”. *The Wiley Handbook of Adult Literacy*, 2019, pp. 471–493. DOI: 10.1002/9781119261407.ch22.

41. R. H. Hall and P. Hanna. “The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention”. *Behaviour & Information Technology* 23:3, 2004, pp. 183–195. DOI: 10.1080/01449290410001669932.

42. J. M. Harackiewicz, J. L. Smith, and S. J. Priniski. “Interest Matters: The Importance of Promoting Interest in Education”. *Policy Insights from the Behavioral and Brain Sciences* 3:2, 2016. PMID: 29520371, pp. 220–227. DOI: 10.1177/2372732216655542.

43. R. Hiraoka, H. Tanaka, S. Sakti, G. Neubig, and S. Nakamura. “Personalized unknown word detection in non-native language reading using eye gaze”. *Proc. of ACM Int. Conf. on Multimodal Interaction (ICMI)*, 2016, pp. 66–70. DOI: 10.1145/2993148.2993167.

44. T. Y. Ho, H. C. Wang, and S. H. Lai. “Non-native language reading support with display of machine translation based on eye-tracking and sentence-level mapping”. *ACM Int. Conf. Proceeding Series*, 2018, pp. 57–63. DOI: 10.1145/3202667.3202675.

45. K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

46. A. Hyrskykari. *Eyes in Attentive Interfaces: Experiences from Creating iDict, a Gaze-Aware Reading Aid*. Tampere University Press, 2006. ISBN: 951-44-6643-8.

47. A. Hyrskykari, tivi Majaranta, and A. Aaltonen. “Design Issues of iDict: A Gaze-Assisted Translation Aid”. *Proc. of the Symp. on Eye Tracking Research & Applications (ETRA)*, 2000. DOI: 10.1145/355017.

48. S. Ishimaru, K. Kunze, T. Dingler, K. Kise, and A. Dengel. “Reading interventions-tracking reading state and designing interventions”. *Proc. of the ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing (UbiComp)*, 2016, pp. 1759–1764. DOI: 10.1145/2968219.2968271.

49. H. Jiang, S. Xu, F. C. Lau, J. Karolus, P. W. Wozniak, L. L. Chuang, H. Jiang, S. Xu, and F. C. Lau. “Capturing user reading behaviors for personalized document summarization”. *Proc. of the Int. Conf. on Intelligent User Interfaces (IUI)* 23-27-Octo, 2011, pp. 355–358. DOI: 10.1145/1943403.1943464.

50. J. Karolus, P. W. Wozniak, L. L. Chuang, and A. Schmidt. “Robust gaze features for enabling language proficiency awareness”. *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems* 2017-May, 2017, pp. 2998–3010. DOI: 10.1145/3025453.3025601.

51. C. Kramsch. “Language and culture”. *AILA review* 27:1, 2014, pp. 30–55.

52. S. Krashen. “The comprehension hypothesis and its rivals”. *Selected Lectures from Int. Symp. on English Teaching and Pan-Asian Conf.* Citeseer. 2002, pp. 395–404.

53. S. Krashen. “We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis”. *The Modern Language J.* 73, 4 1989, p. 440. ISSN: 00267902. DOI: 10.2307/326879.

54. S. D. Krashen. “Explorations in language acquisition and use explorations in language acquisition and use”. *Heinemann*, 2003, p. 87.

55. V. Kuperman, N. Siegelman, S. Schroeder, C. Acartürk, S. Alexeeva, S. Amenta, R. Bertram, R. Bonandrini, M. Brysbaert, D. Chernova, S. M. D. Fonseca, N. Dirix, W. Duyck, A. Fella, R. Frost, C. A. Gattei, A. Kalaitzi, K. Lõo, M. Marelli, K. Nisbet, T. C. Papadopoulos, A. Protopapas, S. Savo, D. E. Shalom, N. Slioussar, R. Stein, L. Sui, A. Taboh, V. Tønnesen, and K. A. Usal. *Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus*. 2022, pp. 1–35. ISBN: 0272263121000. DOI: 10.1017/S0272263121000954.

56. M. M. A. Latif. “Eye-tracking in recent L2 learner process research: A review of areas, issues, and methodological approaches”. *System* 83, 2019, pp. 25–35. ISSN: 0346-251X. DOI: 10.1016/J.SYSTEM.2019.02.008.

57. P. Lightbown and N. Spada. *How Languages are Learned 4e*. 4th ed. Oxford University Press, 2013. ISBN: 978-0-19-4442224-6.

58. J. Liu. “L1 Use in L2 Vocabulary Learning: Facilitator or Barrier”, 2008.

59. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. “Roberta: A robustly optimized bert pretraining approach”. *arXiv preprint arXiv:1907.11692*, 2019.

60. S. Logan, E. Medford, and N. Hughes. “The importance of intrinsic motivation for high and low ability readers’ reading comprehension performance”. *Learning and Individual Differences* 21:1, 2011, pp. 124–128.

61. A. Lund. “Measuring Usability with the USE Questionnaire”. *Usability and User Experience Newsletter of the STC Usability SIG* 8, 2001.

62. M. F. Lungu, L. van den Brand, D. Chirtoaca, and M. Avagyan. “As We May Study: Towards the Web as a Personalized Language Textbook”. *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. Association for Computing Machinery (ACM), Montreal QC, Canada, 2018, pp. 1–12. ISBN: 9781450356206. DOI: 10.1145/3173574.3173912.

63. V. Marian, J. Bartolotti, S. Chabal, and A. Shook. “CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities”. *PloS one* 7, 2012, e43230. DOI: 10.1371/journal.pone.0043230.

64. V. Marian, H. K. Blumenfeld, and M. Kaushanskaya. “The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals”, 2007.

65. P. Martínez-Gómez and A. Aizawa. “Recognition of understanding level and language skill using measurements of reading behavior”. *Proc. of Int. Conf. on Intelligent User Interfaces (IUI)*, 2014, pp. 95–104. doi: 10.1145/2557500.2557546.

66. P. Martínez-Gómez, T. Hara, and A. Aizawa. “Recognizing personal characteristics of readers using eye-movements and text features”. *Proc. of the Int. Conf. on Intelligent User Interfaces (IUI)*, 2012, pp. 1747–1762. doi: 10.1145/2557500.2557546.

67. O. Namnakani, Y. Abdrabou, J. Grizou, A. Esteves, and M. Khamis. “Comparing Dwell Time, Pursuits and Gaze Gestures for Gaze Interaction on Handheld Mobile Devices”. *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. Association for Computing Machinery (ACM), Hamburg, Germany, 2023. isbn: 9781450394215. doi: 10.1145/3544548.3580871.

68. A. Okoso, K. Kunze, and K. Kise. “Implicit gaze based annotations to support second language learning”. Association for Computing Machinery (ACM), Inc, 2014, pp. 143–146. isbn: 9781450330473. doi: 10.1145/2638728.2638783.

69. G. H. Paetzold and L. Specia. “Anita: An intelligent text adaptation tool”. *Proc. of COLING Int. Conf. on Computational Linguistics System Demonstrations*, 2016, pp. 79–83.

70. A. Pintard and T. François. “Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words”. English. *Proc. of the Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*. European Language Resources Association, Marseille, France, 2020, pp. 85–92. isbn: 979-10-95546-45-0.

71. K. Rayner, A. Pollatsek, and E. R. Schotter. *Reading Word Identification and Eye Movements*. Vol. 4. 2012, pp. 548–577.

72. L. Rello, R. Carlini, R. Baeza-Yates, and J. P. Bigham. “A plug-in to aid online reading in spanish”. *Proc. of W4A the Web for All Conf.*, 2015. doi: 10.1145/2745555.2746661.

73. J. C. Richards. *Moving Beyond the Plateau From Intermediate to Advanced Levels in Language Learning*. Cambridge University Press, 2008. isbn: 978-0-521-97597-1.

74. F. Robertson. “Show, Don’t Tell: Visualising Finnish Word Formation in a Browser-Based Reading Assistant”. *Proc. of the Workshop of Natural Language Processing for Computer Assisted Language Learning (NLP4CALL)* 175, Nlp4call 2020, pp. 37–45. doi: 10.3384/ecp2017537.

75. S. Rott, J. Williams, and R. Cameron. “The effect of multiple-choice L1 glosses and input-output cycles on lexical acquisition and retention”. *Language Teaching Research* 6:3, 2002, pp. 183–222. doi: 10.1191/1362168802lr108oa.

76. N. Sagarra and M. Alba. “The Key Is in the Keyword: L2 Vocabulary Learning Methods With Beginning Learners of Spanish”. *The Modern Language J.* 90, 2 2006, pp. 228–243. ISSN: 1540-4781. DOI: 10.1111/J.1540-4781.2006.00394.X.

77. M. S. Schmid. *Language attrition*. Cambridge University Press, 2011.

78. A. Sherstinsky. “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network”. *Physica D: Nonlinear Phenomena* 404, 2020, p. 132306.

79. M. Shimabukuro, J. Zipf, M. El-Assady, and C. Collins. “H-Matrix: Hierarchical Matrix for Visual Analysis of Cross-Linguistic Features in Large Learner Corpora”. *Proc. of the IEEE Conf. on Information Visualization (short papers)*. 2019.

80. A. Tack, P. Desmet, C. Fairon, and T. François. *Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers*. eng. 2021-06-25.

81. A. Taylor. “The Effects of CALL Versus Traditional L1 Glosses on L2 Reading Comprehension”. *CALICO J.* 23, 2 2006, pp. 309–318.

82. A. I. Tweissi. “The Effects of the Amount and Type of Simplification on Foreign Language Reading Comprehension”. *Reading in a Foreign Language* 11, 1998, pp. 191–206.

83. R. Wilkens, D. Alfter, X. Wang, P. Pintard, A. Tack, K. Yancey, and T. François. “FABRA: French Aggregator-Based Readability Assessment toolkit”. *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC)*, 2022.

84. P. Winke, S. Gass, and T. Sydorenko. “Factors influencing the use of captions by foreign language learners: An eye-tracking study”. *Modern Language J.* 97, 1 2013, pp. 254–275. ISSN: 00267902. DOI: 10.1111/j.1540-4781.2013.01432.x.

85. B. Wolter. “Lexical Network Structures and L2 Vocabulary Acquisition: The Role of L1 Lexical/Conceptual Knowledge”. *Applied Linguistics* 27, 4 2006, pp. 741–747. ISSN: 0142-6001. DOI: 10.1093/APPLIN/AML036.

86. A. Yanagisawa, S. Webb, and T. Uchihara. “How do different forms of glossing contribute to L2 vocabulary learning from reading?” *Studies in Second Language Acquisition* 42, 2 2020, pp. 411–438. ISSN: 0272-2631. DOI: 10.1017/S0272263119000688.

87. Y. Yano, M. H. Long, and S. Ross. “The Effects of Simplified and Elaborated Texts on Foreign Language Reading Comprehension”. *Language Learning* 44, 2 1994, pp. 189–219. ISSN: 14679922. DOI: 10.1111/j.1467-1770.1994.tb01100.x.

88. K. Yoshimura, K. Kise, and K. Kunze. "The eye as the window of the language ability: Estimation of English skills by analyzing eye movement while reading documents". *Proc. of the Int. Conf. on Document Analysis and Recognition, ICDAR* 2015-Novem, 2015, pp. 251–255. DOI: 10.1109/ICDAR.2015.7333762.

89. R. Yu. "Culture in Second or Foreign Language Acquisition", 2020. DOI: 10.17507/jltr.1106.10.