

CALL-me MT: a Web Application for Reading in a Foreign Language

Madeleine Comtois

A Dissertation

Presented to the University of Dublin, Trinity College
in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Intelligent Systems)

Supervisor: Yvette Graham

August 2022

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Madeleine Comtois

August 18, 2022

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Madeleine Comtois

August 18, 2022

Acknowledgments

I would like to thank my supervisor Yvette Graham for her guidance during this project. Our mutual passion for the machine translation domain motivated the different approaches taken in this dissertation project. Additionally I would like to thank my friends and family for their support and encouragement throughout this entire process. They were always there for me when I needed to take a step back and look at the bigger picture. Finally, I would like to thank my mother, whose writing and editing advice has helped me for many years and, hopefully, for many years to come.

MADELEINE COMTOIS

*University of Dublin, Trinity College
August 2022*

CALL-me MT: a Web Application for Reading in a Foreign Language

Madeleine Comtois, Master of Science in Computer Science
University of Dublin, Trinity College, 2022

Supervisor: Yvette Graham

Computer-Assisted Language Learning (CALL) applications are tools that use a wide range of technology to provide learning resources for foreign language learning. This dissertation describes the development of a CALL platform that uses Machine Translation (MT) tools to assist students reading in their target language. Highlighting a word or phrase pulls the translation up right onto the screen, increasing the reading speed and fluency of the learner. Testers of the platform showed overall positive results in this tool increasing their enjoyment and skills for reading in a foreign learning.

Additionally, this thesis analyses the performance of Google Translate, Microsoft Translator, and DeepL Translate to measure translation accuracy with or without the addition of contextual information. This accuracy was measured using BLEU, TER, and ChrF scores. Although resulting in mixed performances regarding translator and language, DeepL Translate outperformed the other two translators the majority of the time and, therefore, was chosen as the MT tool to integrate into the CALL platform.

Summary

The purpose of this thesis is to develop a web application that serves as a useful tool for the language learning community. The web application provides texts in multiple languages for learners to choose from. When reading the selected story, the learner highlights a word or phrase they do not know. The translation to this highlighted text is then displayed on the screen next to the selected word or phrase. Integrating the translation into the text itself helps keep users engaged and focused on their work instead of becoming slowed down and frustrated by a start-stop process. This platform was tested by 10 users in order to gain feedback on the design and effectiveness of the application.

The second purpose of this thesis is to compare the accuracy and performance of three machine translation tools. Google Translate, Microsoft Translator, and DeepL Translate were analysed not only by how accurate the machine-translated text was to a human-translated reference, but also how accurate it was with or without the presence of additional contextual information. Each sentence in the text was translated both in isolation, as well as within the context of the entire document. These two translations were compared using BLEU, TER, and ChrF scores to see how additional text affected the accuracy of the translations. Of the three translators tested, there was no significant difference in the addition of textual information, and performance varied depending on which language or translator was used. Overall, DeepL provided the most accurate results, and was therefore used as the machine translation tool in the web application.

Contents

Acknowledgments	iii
Abstract	iv
Summary	v
List of Tables	ix
List of Figures	x
Chapter 1 Introduction	1
1.1 Background	1
1.2 Motivation	2
1.3 Research Question and Objectives	3
1.4 Thesis Structure	4
Chapter 2 Literature Review	5
2.1 Technology for Language Learning	5
2.2 Machine Translation to Assist Reading	6
2.3 Machine Translation	9
2.3.1 Statistical Translation	9
2.3.2 Neural Translation	10
2.3.3 Using Textual Context	13
2.3.4 Evaluating Machine Translation	15

Chapter 3 Methodology	19
3.1 Comparing Machine Translation Tools	19
3.1.1 Analysing the Use of Textual Context	20
3.1.2 Data set	21
3.1.3 Translating In and Out of Context	21
3.2 CALL Platform Design	22
3.2.1 Textual Data	22
3.2.2 User Experience	23
Chapter 4 Implementation	24
4.1 Context Translation Analysis	24
4.1.1 Data processing	24
4.1.2 Translating the Text	26
4.2 CALL Platform Implementation	27
4.2.1 Framework and Architecture	27
4.2.2 Front-end User Interface	28
4.2.3 Back-end Server	31
4.2.4 Deployment and Testing	34
Chapter 5 Evaluation and Discussion	35
5.1 Analysis of Translator Performances	35
5.2 User Feedback on CALL Platform	39
Chapter 6 Conclusions	41
6.1 Conclusion	41
6.2 Limitations	42
6.3 Future Work	43
Bibliography	44
Appendices	48
.1 Link to CALL Application	49
.2 GitHub code	49
.2.1 Code for the web application	49

.2.2	Code for context analysis	49
.3	Survey responses	49

List of Tables

2.1	Example calculation of clipped precision for BLEU score	16
4.1	Language pairs for each source file used for evaluation	25
5.1	BLEU, TER, and ChrF scores for translating sentences in and out of context with DeepL Translator	36
5.2	BLEU, TER, and ChrF scores for translating sentences in and out of context with Microsoft Translator	36
5.3	BLEU, TER, and ChrF scores for translating sentences in and out of context with Google Translate	37
5.4	Summary of translator performances	38

List of Figures

2.1	Difference in skill performance of audio vs. book-based foreign language course for Grade 3 students [1]	7
2.2	The model architecture of Google’s Neural Machine Translation system showing the encoder, decoder, and attention model [2]	12
2.3	The relationship of the five aspects of context [3]	14
2.4	Formula for calculating the BLEU score [4]	17
2.5	Formula for calculating the ChrF score [5]	18
4.1	Processing source and reference files into arrays of sentences	25
4.2	Translating source text in and out of context	26
4.3	System Architecture of the Application	28
4.4	User flow of the web application	30
4.5	User Interface	31
4.6	User Interface - Story Menu	32
4.7	User Interface - Story Text	32
4.8	User Interface - Survey Button	32
4.9	User Interface - Help Button	33

Chapter 1

Introduction

This chapter provides the background and motivation for the project, as well as stating the goals and objectives of the undertaken research.

1.1 Background

Learning a foreign language is a challenge many people undertake during their lifetime. Some learn a second language for school, some for work, some for travel, some for enjoyment, and some because it is a necessary part of life. As communication is one of the main factors that makes us human, we have been learning languages for centuries. For many of the world's languages, four core skills are acquired in order to master a language: listening, speaking, reading, and writing [6]. One of these skills in particular—reading—is the focus of this project.

Reading is an accessible way to practice a foreign language. Given the depth of the internet, it is relatively simple for learners of widely-spoken languages to find content to read on the web. This provides a convenient way for learners to practice and improve their language skills. These skills are easier for independent learners to achieve given the abundance of modern technology that can assist their learning, especially if no teachers or native speakers are available to answer questions.

Many different digital platforms, such as CALL (Computer-Assisted Language Learning) applications, provide tools for language learners. Machine translation is one of these tools widely used in foreign language learning to look up unknown con-

tent. Online translators can provide translations to readers when no native speakers are available to help. Although these tools have evolved significantly over the years to become even more accurate and reliable, they are still not perfect in comparison with native speakers. One issue in particular with machine translation is the use of context when providing a translation. Words or phrases might be translated differently depending on what information precedes or follows them; therefore, without this knowledge machine translators do not always provide the best translations. This issue makes it more difficult for learners to obtain accurate knowledge when they encounter words they do not know in their reading.

1.2 Motivation

One of the challenges in reading a foreign language is the interruption created by stopping to look up a word the learner does not know before coming back to the text. This start-stop process interrupts the flow of reading and can become frustrating to the learner. If learners are frustrated, they are more likely to become discouraged and quit practicing.

Fortunately, technology can be used to reduce these interruptions so that readers can engage with the material and learn, increasing their reading speed to that of a native speaker. This is achievable by bringing translation to the forefront of their learning. If a learner is reading online text, instead of being interrupted by having to look up a word (either in a dictionary or online resource), the translations should be provided on the screen adjacent to the word or phrase the learner does not know. This saves time and energy, keeping the learner less frustrated and engaged, as well as increasing reading speed.

In providing these types of tools, the technology needs to be accurate so learners get the best experience. Machine translation tools, such as Google Translate, are especially popular for reading tasks, as they provide direct translations of unknown words and phrases. However, these tools have pros and cons regarding accessibility and translation accuracy. Only the best tools should be used to provide a learner with the best experience.

1.3 Research Question and Objectives

Given the omnipresence of technology, especially in the education domain, language learners have many tools at their disposal to assist in their foreign language reading. However, to keep learners engaged in their reading, it is important that these tools are readily available and easy to use. Unfortunately, many of these tools are very complex and involved, which can distract from the learning experience. This project hopes to put the focus of language learning back on the reader by providing a simple, interactive application without all the distracting bells and whistles that often accompany these platforms.

Three main issues in particular arise when creating an environment for reading in a foreign language:

1. Learners need an easy, effortless way to read in order to make their learning more efficient.
2. Learners need encouragement to continue practising their skills without becoming discouraged by their lack of knowledge.
3. Learners need access to accurate information so they learn the language correctly.

These issues can be addressed by harnessing the power of advanced machine translation tools to create a simple application for learners to use in an enjoyable manner. This thesis therefore aims to investigate whether an interactive application that integrates popular language translation technology encourages learners to read in a foreign language so that their overall learning experience is enjoyable and effective, as well as to determine if these tools are efficient for the job. Thus, the research for this thesis is split into two distinct processes:

1. Evaluate and benchmark different machine translation tools.
2. Develop and test a CALL application to assist foreign language reading using the tool best selected from the previous evaluations.

For this thesis three different popular machine translation tools are evaluated to identify one that works best for language learners to achieve their goals: Google Trans-

late,¹ Microsoft Translator,² and DeepL Translate³. These evaluations were done on the basis of translation accuracy: accuracy in regard to translation quality as well as accuracy in regard to translating text in and outside the context of an entire document.

Using these evaluations to determine which translator performed the best, a CALL web application was developed using this translator to create an interactive reading environment for learners to practice reading. Learners tested the application in order to provide feedback and insight for evaluation and future work on the application.

1.4 Thesis Structure

This thesis is divided into six chapters. Chapter one contains the introduction and motivation behind the project, as well as the goals of the research. Chapter two discusses previous literature on foreign language acquisition, technology for language learning, and different machine translation techniques and evaluations. Chapter three discusses the methodology of the project, including the experiment design of the analysis of the machine translation tools as well as the CALL platform design. Chapter 4 delves deeper into the implementation of the analysis and CALL application development. Chapter 5 provides evaluations and discussions on these experiments, such as the scoring techniques used for the different translation tools and the survey results for testers of the CALL application. Chapter 6 finishes with conclusions, limitations, and future work on the project.

¹<https://cloud.google.com/translate>

²<https://www.microsoft.com/en-us/translator/business/translator-api>

³<https://www.DeepL.com/en/docs-api>

Chapter 2

Literature Review

This section discusses the history, science, and importance of reading in a foreign language and how technology can be applied to this process. It introduces popular CALL platforms and the background of different machine translation techniques.

2.1 Technology for Language Learning

Computer-Assisted Language Learning (CALL) platforms are a growing industry in a contemporary world where it is common to learn a foreign language. The abundance of technology makes language learning easier, engaging, and even more attainable. Not only do these tools enhance the language learning experience, but they also serve to keep endangered languages alive and more accessible [7][8], as well as make the learners more confident and autonomous [9].

A popular CALL system used by many language learners is the platform Duolingo¹. Duolingo is an app launched in 2012 that offers language courses where users can practice vocabulary, grammar, and pronunciation. A study conducted by Queens College shows that the app is effective in boosting confidence and motivation for learning a language [10]. This study states that a beginner language learner would need an average of 34 hours to master the material equivalent to a semester of college. Although effective for learning new words and phrases, it does not provide all the benefits of engaging with a native speaker and culture [11]. For independent learners, however, it

¹<https://www.duolingo.com/>

is an invaluable tool.

Another online language learning platform is FluentU². Founded in 2011, this platform provides authentic, foreign-language media clips for teaching or learning a foreign language. FluentU's digital library is available in 10 languages, and each video is paired with interactive subtitles and exercises. This platform helps users learn pronunciation and vocabulary, along with training the ear [12]. Like Duolingo, it is effective for the independent learner.

Quizlet is a CALL application used for effective vocabulary learning and practice [13]. Its online flashcards are equipped with audio to help students learn pronunciation along with vocabulary. Quizlet also provides games and exercises to practice grammar concepts and verb conjugations [14][15]. Like other CALL applications, it is an effective tool for independent learners or as an additional aid in a classroom setting.

Effectively obtaining a second language has many theoretical approaches, such as cognitive linguistic, psycholinguistic, human-learning, and social context learning [16]. These theoretical approaches must be taken into account in the development of CALL systems so as to provide the most effective language learning experience. These aforementioned platforms are only a few examples of the multitude of technological resources available to language learners. While these platforms focus on a holistic, well-rounded approach to learning a foreign language, this dissertation looks into the development of reading skills, one aspect of the language learning process that is not the main focus of these other platforms. By focusing on this particular aspect of learning, we gain valuable insight into readers' preferences and experiences to improve CALL applications for the future.

2.2 Machine Translation to Assist Reading

A common tool students use when reading in a foreign language is automatic machine translation. Faster than looking up words in a dictionary and advanced to the point of providing document-level translations, these tools are at the core of assisting reading comprehension. Reading is an invaluable skill, and consistent practice enables the learner to read fluently in their target language, just like a native speaker. Two types

²<https://www.fluentu.com/en/>

of reading are commonly found in education in regard to practising the skill: intensive reading and extensive reading [1]. Intensive reading involves students working with short texts to identify main ideas and enhance vocabulary and grammar skills. Extensive reading, however, exposes the language learner to extensive quantities of reading material. The prolonged nature of reading in this way makes language skills acquisition easier to master as the student progresses. Although quite different, both approaches are important in developing strong reading skills, and this dissertation aims to address both of these techniques.

A paper by Nanyang Technological University [1] describes a research experiment [17] in comparing the results of audiolingual vs. book-based language learning. In the study, Grade 3 students who had a book-based course outperformed the students of the audio-based course on an end-of-year exam. The results of this study are shown in Figure 2.1.

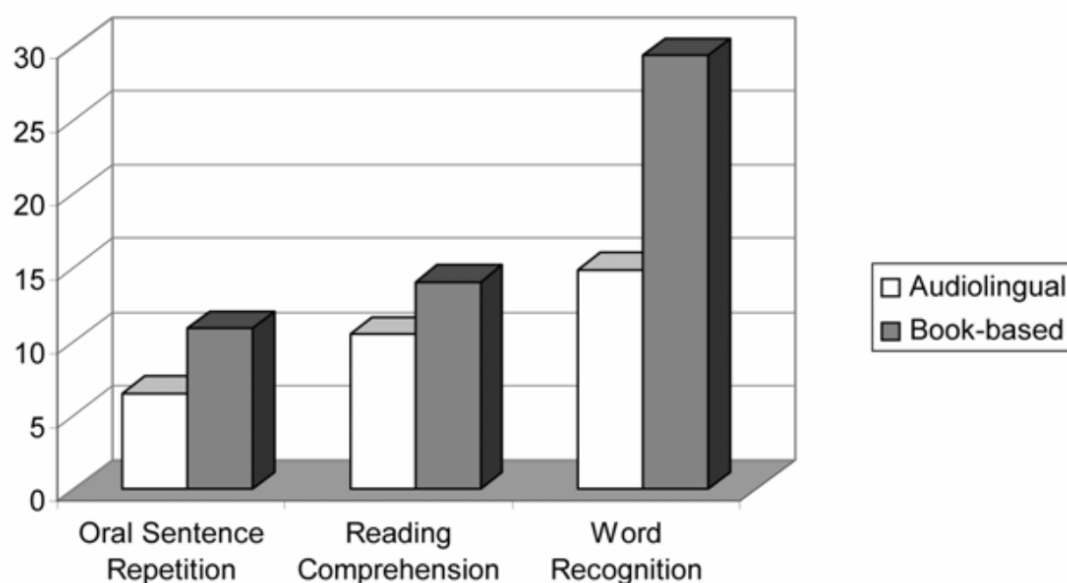


Figure 2.1: Difference in skill performance of audio vs. book-based foreign language course for Grade 3 students [1]

Given that textual input is an important aspect of learning a foreign language, machine translation tools have been introduced into classrooms to assist this process. Machine translation can be used not only to look up an unknown word, but also for reading comprehension, comparing/contrasting input and output, error spotting,

grammar practice, and much more. Although practical tools, they are not always perfect for the learning environment. Some students cannot always pick out the good and bad results from machine translated text, and they do not know how to solve translation problems on their own. However, machine translation can be great for those who find manual translation too difficult and time consuming.

In a study by The University of Manchester, researchers conducted several surveys among teachers and language learners to learn about the advantages and disadvantages of applying machine translation to traditional language learning [18]. Out of the 30 teacher participants, 70% used machine translation in the classroom, with the main objectives being reading comprehension, essay writing, revision, translation practice, accuracy training, and assessment. However, the majority of the participants stated that they would only use these tools in advanced classes, as students who have less knowledge of the language can easily turn to them for quick answers without learning the reasoning behind the translations. When asked about the limitations of machine translation, the majority stated that the translations can sometimes be of low quality and not always reliable.

Although machine translation has greatly improved since the undertaking of this study, these limitations are still an issue that affect the use of machine translation today. Using machine translation as the sole resource in understanding foreign text has the possibility of teaching learners incorrect words or syntax. The syntactic priming effect (syntactic/structural alignment) is a phenomenon where people repeat a syntactic structure they have previously seen, heard, or read. This effect was tested in a study by Resende et al. [19] on Brazilian Portuguese speakers. The speakers read a machine translated English caption of a picture, and then they were asked to describe the picture verbally. The study then concluded that participants who had previously seen a given translation more likely used the same syntactic structure as the translation system than participants who did not see a previous translation. Because of effects like these, it is easy to learn incorrect information, which is why it is important that machine translators be as accurate as possible. This dissertation therefore aims at selecting an accurate translator so learners can be confident that they are provided with the best tools.

2.3 Machine Translation

Machine translation is the process of automatically translating text from one language to another without human intervention. It is a very challenging task for artificial intelligence given the amount of vocabulary, rules, and fluidity of human language. With the widespread use and convenience of the internet, there is a greater need than ever before for quick, reliable, online translations. The idea of machine translation can be traced back to the seventeenth century, but the first practical implementations did not occur until the 1930s with the invention of mechanical multilingual dictionaries [20]. Using the concepts of cryptography, statistics, logic, and information theory, machine translation systems then evolved to process even more complex aspects of human language, such as syntax and vocabulary ambiguity.

The first machine translation models were rule-based (RBMT). These models used rules developed by linguistics for systematically translating text from one given language to another. These rules included information on the lexical, syntactic, and even semantic characteristics of a particular language. Although this approach guarantees complete control on the outcome of the translation model, it requires great amounts of time and expertise in linguistic knowledge [21].

Even if rule-based machine translation models provide a solid foundation for automatic translation, newer models have evolved to better address the fluidity and ambiguity of human language. Statistical models replaced rule-based models by translating from example, and neural translation models evolved by using artificial neural networks to learn and predict the translation output.

2.3.1 Statistical Translation

Statistical machine translation models, first developed in the 1980s, replaced the original rule-based systems by utilising a large corpus of examples [22]. The idea is to maximise the probability of an output sequence given the input sequence following an explicit suite of candidate translations. This probability can be modelled by the equation:

$$Pr(S|T) = \frac{Pr(S)Pr(T|S)}{Pr(T)} \quad (2.1)$$

Brown et al. use this model to build a statistical model to translate from French to English [23]. Given a sentence T in English, the model looks for a sentence S in French that could have produced T . Error is minimised by choosing the S that is most probable given T . Therefore, the model chooses S to maximise the probability $Pr(S | T)$.

This statistical approach is driven by data, as it applies a search process to select the most likely translation from the model's probability distribution. Given the large available corpus of examples in both the source and target language, linguists are no longer necessary to specify all the rules of translation.

These models work well for sequence-based translations, such as a sequence of words. They are able to parse phrases by breaking down the text and translating sub-sequences of words [22]. However, given the finite corpus of words and phrases to translate from, statistical models are not always able to capture the broad, fluid nature of human language. Certain grammar rules, which could be produced by linguists, can be overlooked since the translations are only data-driven. They also need to be updated and tuned as language and data evolve.

2.3.2 Neural Translation

The current state-of-the-art machine translation models are built using neural networks. Neural machine translation (NMT) is able to address the limits of statistical translation, such as the vast knowledge of linguistic rules and grammar exceptions required to produce a human-like translation. Neural translation models are used to learn a statistical model and only require training on source and target language text instead of using specialised systems.

Early neural translation models were based on Multi-layer Perceptron (MLP) neural networks. These networks are fully connected feed-forward artificial neural networks that pass data through at least one hidden layer [24]. The input values are fed through multiple layers in order to update their values; this is called “feeding forward”. For MLP, an input layer of nodes is fed in feature values. These values are then fed forward into a certain number of hidden layers, where they are finally fed to the final output layer. The hidden layers are given their name due to the fact that their activation values are not accessible directly from outside the network, only as they pass from one

layer to the next.

$$o_j = f\left(\sum_i w_{i,j} a_i + b_i\right) \quad (2.2)$$

In equation 2.2, the output layer nodes o equal the sum of input layer nodes a , each multiplied by a weight w and added bias b , which is passed to an activation function f .

Although these models are effective neural networks, they are limited in that the input sequence must be the same length as the output sequence. These models have thus been replaced by recurrent neural networks which allow any size of input and output sequences. This is done by implementing an encoder-decoder method for a given source and target language pair, where an encoder reads a source sentence and encodes it into a vector of fixed length [25]. A decoder is then used to output the translation using the encoded vector by producing the translated words one at a time with the help of an attention mechanism. This helps in processing the semantics of very long input sources. The encoder and decoders are trained together in order to maximise the probability of a correct translation. This architecture is used in Google’s Neural Machine Translation system (Google Translate) [2]:

$$Pr(X|Y) = \prod_{i=1}^N P(y_i|y_0, y_1, y_2, y_3, \dots, y_{i-1}; x_1, x_2, x_3, \dots, x_M) \quad (2.3)$$

where (X, Y) is a source and target sentence pair, $x_1, x_2, x_3, \dots, x_M$ is the sequence of M symbols in the source sentence, and $Y = y_0, y_1, y_2, y_3$ is the sequence of N symbols in the target sentence. The probability of the next symbol is calculated given the source sentence encoding and the decoded target sequence.

Google is not the only entity that implements a neural network for its automatic translation. Originally developed as a statistical machine translation model, Microsoft Translator now uses 24 Transformer encoder layers and 12 decoder layers in the feed-forward neural network for their multilingual translation model [26]. DeepL Translate is another neural network model that performs highly in translation. A study by the University of Geneva uses both automated and human evaluations to show the effectiveness of DeepL and how it performs better than traditional statistical models [27].

In addition to training a model on a single pair of source and translation languages,

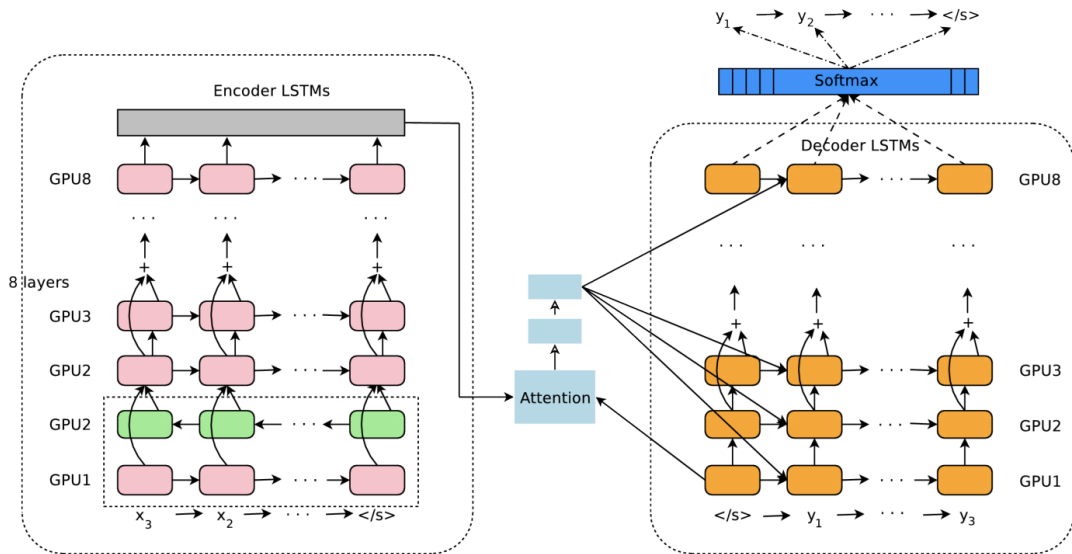


Figure 2.2: The model architecture of Google’s Neural Machine Translation system showing the encoder, decoder, and attention model [2]

neural machine translation can also be used for Multilingual Neural Machine Translation (MNMT). As a vast majority of language corpora for machine translation is in English, this translation ability is important for minority languages or language pairs that do not involve English. These models are trained like the original NMT models, except they use a joint set of bilingual corpora from different languages. Using this model, translations can be substituted by using the translation from one language as the input for another to get the translation in a language lacking in data. These models are very effective and perform just as well as standard NMT models [28] [29].

Although neural machine translators are the most accurate models to date, they are not perfect. Issues regarding these systems include scaling to larger vocabularies, including rare words, the speed it takes to train the models, as well as the use of context [2]. Textual context in particular is important for a translation model as it provides the background information necessary to produce more human-like translations [30]. These issues are currently being addressed in the state-of-the-art models, but there is much more development to be done in order to produce high quality, human-like translations. All of the translators tested in this dissertation are neural machine based translators.

2.3.3 Using Textual Context

One of the major differences between human and machine translation is the knowledge and understanding of context. For a translation to be completely accurate, a machine needs background knowledge to navigate any ambiguity or unknown content that could be present in a given input text [31]. Textual context includes information that helps readers accurately interpret the meaning of a text, such as background information or specific grammatical references. Translations of text can vary greatly depending on the information available.

Melby and Foster [3] describe five aspects of context that are necessary to identify in order to understand the source text and accurately produce the target text:

- Co-text: Surrounding text of the source word or phrase, such as definitinal text or extra information
- Rel-text: Monolingual resources for the source text, such as a dictionary
- Chron-text: The chronological changes of the source text as different versions are updated or edited over time
- Bi-text: Bilingual resources for the source text, such as bilingual glossaries and similar texts to the source text that have already been translated
- Non-text: The “real-world” setting of a document that could have a cultural or linguistic significance

All five of these concepts are important to keep in mind when developing machine translators. The more resources available to supply this knowledge, the more accurate the target translations will be.

Another theory by House sees text and context even more dynamically related with a reflexive relationship between linguistic and non-linguistic components [32]. She introduces the theory of re-contextualisation, which is the idea of taking text out of its original context and placing it within a new context of relationships and culture. This process is completed by the means of overt and covert translation. In overt translation, the original context is added with the target context so that both are present in the target context. For covert translations, only the target context is taken

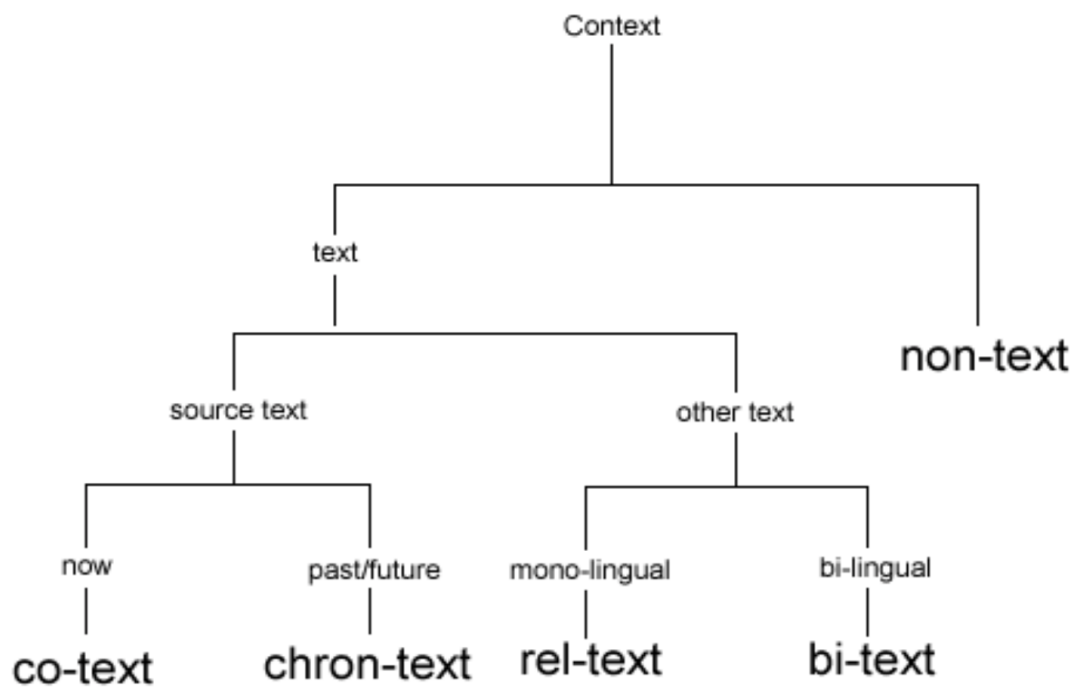


Figure 2.3: The relationship of the five aspects of context [3]

into account, making the translation more affected by any contextual differences. This re-contextualisation is important in order to translate not only words, but cultural ideas and expressions that would otherwise get lost in translation.

Neural machine translation helps address these contextual issues by introducing the idea of document translation. Instead of training the encoder-decoder network on sentence-level pairs, the network is trained on a document with methods to remember and take advantage of textual context. There are many ways to go about training such models. Wang et al. [33] do this by adding an additional encoder. Another approach by Tu et al. [34] uses a cache-based memory network which stores past context as a set of words that stay learned by the network during the translation process. Miculicich et al. [35] propose a hierarchical attention network by modelling the contextual information with word-level and sentence-level abstractions. In these trials, the models performed better with the additional context compared to the baselines. A similar approach is taken in this dissertation by comparing popular neural machine translation tools and analysing how they perform given different amounts of context.

2.3.4 Evaluating Machine Translation

It is important to evaluate the outputs of machine translation in order to assess the needs and areas of improvement. One way to evaluate translations is by using human translators. Human translators are able to identify what constitutes a good translation and what does not sound natural or correct. However, professional translators are not the only people qualified for evaluating translation systems. A paper by Graham et al. conducted an experiment to show how crowd sourcing is also an effective way to quickly provide quality evaluations on machine translation [36]. There are many ways to manually evaluate translations, such as ordinal level scales, point systems, and subjective judgements. This study employed volunteers to use continuous scales to rank the adequacy and fluency of translated text. In order to ensure the responses of the experiment are of a high quality, the translation task was reduced to monolingual evaluation of the translated text, and biases were filtered out along with inconsistent participants. This process provided more conclusive results on the difference in performance of various machine translators.

Although manual assessment provides the highest quality evaluation of machine

translation, and even if crowd sourcing provides significant data for mass translation, it is not always accessible or practical to use. This problem can be addressed using various metrics for automatic evaluation, with the goal of producing scores as close as possible to human ones. One of these scores is the Bilingual Evaluation Understudy (BLEU) [4]. This score is calculated using a weighted average of n-gram matches against a reference translation. *N-gram* is a term used to describe a set of n consecutive words in a sentence. Clipped precision is the core metric used for this score, and it is calculated by comparing each word from the candidate sentence with all of the words in the reference sentences. These matches are counted up, but the number is limited, or ‘clipped’, to the maximum number of times that the word occurs in any reference sentence. The final precision score equals the number of correct predicted words over the number of total predicted words. The following example adapted from [4] shows how the candidate sentence receives a precision score of 2/7 using 1-gram intervals:

Candidate: *the the the the the the the.*

Reference 1: *The cat is on the mat.*

Reference 2: *There is a cat on the mat.*

Candidate Word	Reference	Matched Count	Max Limit
the	Reference 1	7	2
the	Reference 2	0	2

Table 2.1: Example calculation of clipped precision for BLEU score

To calculate the final BLEU score, clipped precision is first calculated for (normally) 1, 2, 3, and 4 n-grams. These values are then used to calculate the Geometric Average Precision value. To combat perfect scores that could arise from 1-gram scores of short sentences (as in the example in Table 2.1), a Brevity Penalty is applied:

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{1-r/c}, & \text{if } c \leq r \end{cases} \quad (2.4)$$

where c is the number of words in the candidate sentence, and r is the number of words in the reference sentence. Finally, the Geometric Average Precision value is multiplied with the Brevity Penalty to get the final BLEU score:

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

Figure 2.4: Formula for calculating the BLEU score [4]

where typically $N = 4$ and the uniform weights are $w_n = N / 4$. The score results in a value between 0 and 1, where a score of 1 indicates a perfect translation (i.e. the candidate sentence matches the reference translations word for word). Although a straightforward and simple metric to implement, given that BLEU is a parameterised metric, it cannot always be used to compare two different systems, as each system might calculate the metric differently [37]. In addition, multiple systems might tokenize the candidate and reference sentences differently, resulting in different n-grams and therefore different scores for the same sentences.

In addition to BLEU, other machine translation metrics can be used to evaluate different systems. The Translation Edit Rate (TER) score measures the number of edits needed for a translation to exactly match the closest reference sentence regarding fluency and semantics [38]. These edits include word deletion, addition, and substitution. The score falls between 0 and 1 and is calculated by dividing the minimum number of edits necessary by the average length of the reference sentence. The lower the TER score, the less editing is needed, resulting in a better performing model. While a good metric for measuring accuracy, it is also a good indicator of the amount of post-editing effort that would be necessary for human translators using machine translation. Although not a perfect indication of the performance of a machine translation system, it generally correlates well with human translations and helps eliminate subjective human judgement.

Another metric that can be used for evaluation is the character n-gram F-score (ChrF). This metric calculates the F-score averaged on all character and word n-grams, and it has proven to outperform both the BLEU and TER scores in terms of accuracy and correlates well with human scores [5]. Because it analyses text on a character level, it is able to pick up on morphological components of language that would otherwise go unnoticed in traditional n-gram evaluations. The ChrF score can be calculated using the following formula:

$$\text{ChrF}\beta = (1 + \beta^2) \frac{\text{CHRP} \cdot \text{CHRR}}{\beta^2 \cdot \text{CHRP} + \text{CHRR}}$$

Figure 2.5: Formula for calculating the ChrF score [5]

where *CHRP* stands for the percentage of n-grams in the candidate sentence that appear in the reference sentence, and *CHRR* stands for the percentage of character n-grams in the reference sentence that also appear in the candidate sentence. These values are used for character n-gram precision and recall, which is arithmetically averaged over all n-grams. The β parameter is used to assign β times more importance to recall than to precision. This metric performs well in comparison to other automatic evaluations and, in comparison with the other metrics, is a good indicator of how well a translation system performs.

Given the caveats with automatic metrics, such as BLEU, other evaluations have been tested in hopes of providing a more accurate metric for evaluating systems. A paper by Specia et al. creates a data set for assessing machine translation evaluation metrics [39]. They use what they call ‘quality estimation metrics’, which are metrics that use machine learning techniques to learn quality estimates from human annotated data. These human evaluated features are not present in metrics that test overlapping between n-grams, such as BLEU, so they are able to provide more insight on the quality of a translation. This study used expert translators to annotate the source and target text. These annotations were used as features for the Partial Least Squares machine learning technique. Using the Root Mean Squared Prediction Error to quantify the amount by which the model differed from the expected scores, the metrics showed that the machine-learned features outperform using solely n-gram metrics.

Other metrics, such as the Metric for Evaluation of Translation with Explicit Ordering (METEOR) also exist for analysing the accuracy of machine translation [40]. METEOR is based on the precision and recall of analysing unigram outputs of translations. This metric also takes into account word stems and synonyms, therefore validating translations that provide an accurate translation but do not match a reference translation word for word. For the scope of this dissertation, the BLEU, TER, and ChrF metrics are used to evaluate different systems given their universality.

Chapter 3

Methodology

This chapter discusses the background and design of the two main focuses of the dissertation: experiments to test the performance of different machine translation tools and the design of a CALL platform integrating these tools.

3.1 Comparing Machine Translation Tools

With increasing research in neural networks and big data, many machine translation tools exist that provide accurate translations given some type of textual input. These tools are widely available, and language learners usually have their preferences in which tools perform the best. Some tools perform better with certain languages, and some perform better given the amount of input data they receive. This dissertation aims to conduct a cross evaluation of three popular tools: Google Translate, Microsoft Translator, and DeepL Translate, evaluating them using the same textual input for a given language.

Translation is not always an objective task. Sometimes there are many ways to translate a given word or phrase, and different translators might provide different, accurate translations. Therefore these tools are compared to the same set of human-translated versions of the text to create a baseline for translation accuracy evaluation.

3.1.1 Analysing the Use of Textual Context

The main focus of comparison for analysing these tools is evaluating how they perform based on the use of textual context provided. For this dissertation, textual context is defined as additional, textual information provided along with a word or phrase to translate, which could be important in determining how a phrase is translated. For example, the phrase *How are you* in English can be translated into French in the two following ways:

1. *Comment vas-tu ?*
2. *Comment allez-vous ?*

French has two versions of the second person pronoun *you* in the singular form: *tu* and *vous*. Translation (1) utilises the *tu* pronoun, and Translation (2) utilises the *vous* pronoun. Although both translations are accurate ways of asking how someone is doing, the translations differ in how they would be used in particular situations. The *tu* pronoun is used in more familiar, less formal situations, such as with friends or children. The *vous* pronoun is used in more formal circumstances, such as with strangers or adults of particular status. If a machine is not provided the context of the situation when given a phrase to translate, it does not know which translation is best to provide.

One way textual context comes into play for providing the best translation is by including key words or phrases that indicate which translation is better to use. For example, if the sentence to translate were preceded with *Hello Mr. Smith*, the use of a formal greeting shows the need of a more formal translation of the following phrase, and therefore Translation (2) would be a better choice. This example is just one of many different ways in which the addition of textual information can enhance the performance of machine translation.

As previously discussed, machine translators are advanced to the point of making statistical decisions on which translations to provide based on machine learning and big data. The focus of this dissertation therefore is to evaluate the quality and accuracy of these decisions using a certain amount of contextual information. Sentences are taken from a document and are translated both out of context (in isolation) and in context

(within the entire document) to see how the two translations for the same sentence compare.

3.1.2 Data set

The data for this project were taken from the WMT-21 News Systems and Evaluations¹ data set. This data set consists of different news articles in several different languages and is provided as part of a task detailed in the 2021 Sixth Conference On Machine Translation [41]. The data set was chosen for the following reasons:

- Open source
- Source text offered in 13 languages
- Human-translated reference text offered for each source text translation pair

Using this data, different evaluations were carried out to test the accuracy of the machine translation tools on the source text by comparing the translations to the human-translated reference text.

3.1.3 Translating In and Out of Context

To evaluate the difference in performance for translating text in and out of context, the text from each language was sent to the translator two different times. The first time, the text was translated individually sentence by sentence. The second time, the text was sent as a whole with all the sentences translated within the context of each other.

The accuracy of these translations is evaluated using different scoring techniques. The outputs from each of the translators were compared with those of the human-translated reference sets to see how well each translator performed. These translations were scored using different machine translation scoring methods, with the highest scoring translator deemed the most accurate for this experiment.

¹<https://github.com/wmt-conference/wmt21-news-systems>

3.2 CALL Platform Design

Following the evaluation of the machine translation tools, an interactive web application was designed that integrates the highest performing translator. This application was developed to facilitate the reading process of learning a foreign language. The idea is to keep the learner engaged, so if they come across a word or phrase they do not know in their reading, instead of getting frustrated or losing focus by searching for a translation, the translation should be provided to them directly. The application must be simple and easy to use, as this layout is best to keep users engaged and focused on their work [42][43]. The design therefore meets the following criteria to benefit the most number of users:

- Simple design so readers are not confused or distracted
- Minimal user input to keep reading at the forefront
- Multiple language options
- Multiple sources of text in a particular language
- Easy access to translations of words, phrases, sentences, or paragraphs

The interface therefore provides the learner with a plethora of reading material so that they have a choice in what they want to read, increasing motivation. The platform provides language support for nine languages: Chinese, Czech, English, French, German, Italian, Japanese, Portuguese, and Spanish. For the scope of this project, the application is directed towards English speakers, so the translations provided are in English. The exception to this is the English text, with translations provided in French (chosen for ease of testing, as I have advanced knowledge of both).

3.2.1 Textual Data

The text for this project is provided from Project Gutenberg², an online library of over 60,000 free books in the public domain. Three books were chosen for each of the languages, totalling 27 books for the platform. All books were originally written

²<https://www.gutenberg.org/>

in the language provided, which adds cultural authenticity to the language learning experience. The books were chosen based on their popularity, originality, and ability to be downloaded as HTML text.

3.2.2 User Experience

The goal of the application is to quickly provide a translation while the learner is reading the text. In order to do this without interrupting the flow of reading, the translations are provided on the screen, adjacent to the unknown word or phrase. When the learner comes across a word or phrase they do not know, they highlight the text with their mouse to send the text to the translator. This simple process keeps the learner engaged with their reading while providing the necessary translation information.

To assess the response and success of the application, a survey was created that asks questions related to the user experience. Testers filled out the survey after using the application for a minimum of five minutes. This feedback provided invaluable data to analyse the success of the project as a whole, and to determine how helpful it would be for the language learning community.

Chapter 4

Implementation

This chapter describes in detail the technical implementation of the context translation analysis, as well as the technical development of the CALL application.

4.1 Context Translation Analysis

Translating text in and out of context provides insight into the performance of different machine translation models. The following sections outline the process undertaken to assess these differences. Given the number of linguistic libraries and resources available, Python was used for this evaluation.

4.1.1 Data processing

The WMT-21 data taken for this experiment consist of both source files and reference files. Source files include around 1000 lines of text, each taken from various news sources, with each line corresponding to a single sentence. For the scope of this project, 200 lines were analysed from each source file. A source file exists for each language pair, such as for English to German or German to English. Each source file has at least one corresponding reference file. The reference file contains the same structure and text as in the source file, except that the text has been translated by a human translator. In the case of multiple reference files, translations from other translators have been supplied. The use of multiple references helps address this issue of sometimes having more than one valid way of translating a piece of text. These reference files are used

as a basis of comparison in translation for the text generated by the different machine translation systems. Table 4.1 shows the language pairs of the source files used for this experiment, where the text in the source file is in the first language, and the text in the corresponding reference files is in the second language. The source text is different for every language pair, so no two texts are used for the same language.

Language Pair	Abbreviation
Czech → English	cs → en
English → Czech	en → cs
German → English	de → en
English → German	en → de
German → French	de → fr
French → German	fr → de
English → Japanese	en → ja
Japanese → English	ja → en
English → Russian	en → ru
Russian → English	ru → en
English → Chinese	en → zh
Chinese → English	zh → en

Table 4.1: Language pairs for each source file used for evaluation

Each language pair was processed and evaluated individually with no cross-referencing amongst the languages. For evaluating a language pair, the source and reference texts are read into arrays, where each entry in the array corresponds to one line of text (one sentence). When multiple references files are included, each sentence array for each reference is contained within another array.

```

Source: [ sentence_1,
            sentence_2,
            sentence_3,
            ... ]

References: [ [ refA_sentence_1,
                  refA_sentence_2,
                  refA_sentence_3,
                  ... ],
                [ refB_sentence1,
                  refB_sentence2,
                  refB_sentence3,
                  ... ] ]

```

Figure 4.1: Processing source and reference files into arrays of sentences

4.1.2 Translating the Text

For comparing the translations of in context vs. out of context, two different translations were performed for each source:

1. In context: Each sentence in the source array was concatenated to a string to create one long string of text. This string was passed into the selected translator to be translated as a whole. The translated text was then parsed back into an array of sentences to be used for evaluation.
2. Out of context: Each sentence in the source array was passed into the selected translator individually. Each translation was then added into an array of translated sentences to be used for evaluation.

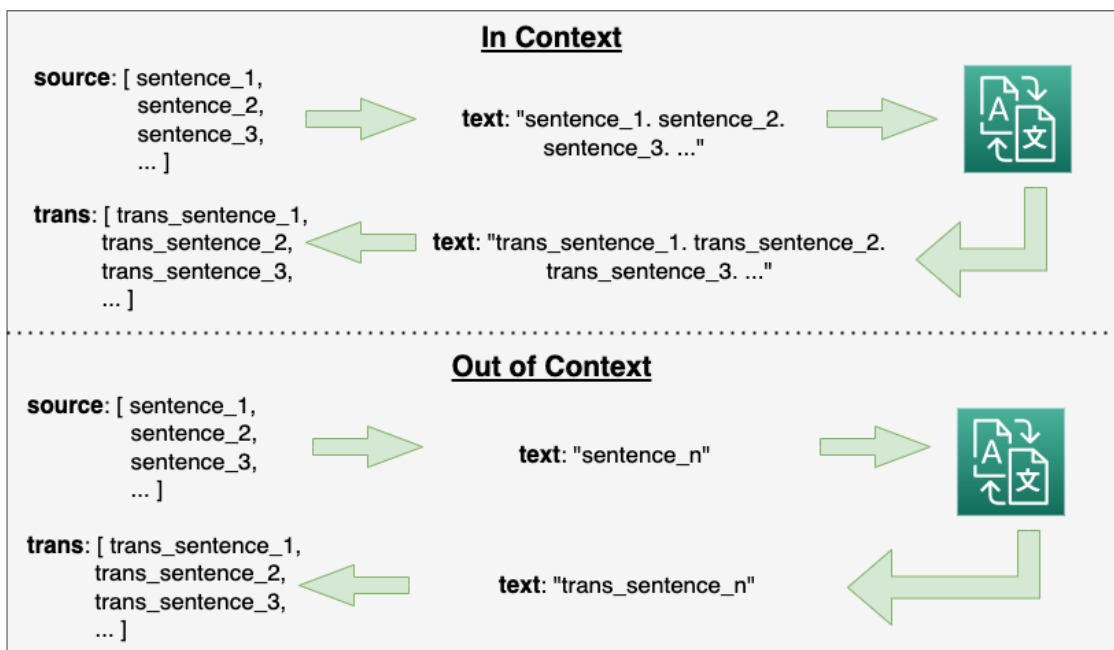


Figure 4.2: Translating source text in and out of context

Three different translators were used for this evaluation: Google Translate, Microsoft Translator, and DeepL Translate. To send the text to the translator, three corresponding APIs were used. The Google API was imported as part of the `google-cloud-translate` python module and was authenticated using a private key. The Microsoft

Translator was accessed using the Python `requests` module to send a post request to the Microsoft server and was also authenticated using a private key. Finally, the DeepL API was accessed using the `DeepL` module along with another private key for authentication. For each API call, the source text was sent along with parameters specifying the source and target languages¹. Each translator was tested on performance for both in and out of context.

To evaluate the performance of the models, three different metrics were selected: BLEU, TER, and ChrF. Originally, the BLEU score was calculated using the functions provided in the `NLTK.translate` python module, but this was changed to the SacreBLEU² implementation as this source is a standard for WMT data, and it provides calculations for the other metrics as well [37]. These scores were calculated by comparing the machine translated sentences to the reference sentences provided by WMT, and then calculating the BLEU, TER, and ChrF scores from their similarities. These final scores and evaluations are reported in Chapter 5.

4.2 CALL Platform Implementation

This section goes into detail about the development of the web application, including the architecture, technology stack, user flow, and user interface.

4.2.1 Framework and Architecture

This project was built as a web application in order to be accessible by the greatest number of users possible. There are many different tools and technologies available for building web applications. This project was developed as a Single Page Application (SPA) [44]. The SPA architecture consists of showing content on a main page following subsequent requests. Instead of getting and sending complete HTML pages, data is sent to and from the server in JSON format, speeding up the response time. The SPA architecture was chosen for its speed and simplified development.

This application consists of both a server side REST API and a client side framework. The server uses REST APIs to send data back and forth from the translators

¹NB: The abbreviation for English ‘en-us’ was changed to ‘en’ for all Google instances.

²<https://github.com/mjpost/sacrebleu>

as well as send these translations to the client side. The client side contains all the development code for the user interface. All of these frameworks are open source and based on JavaScript. Figure 4.3 depicts the architecture of the system.

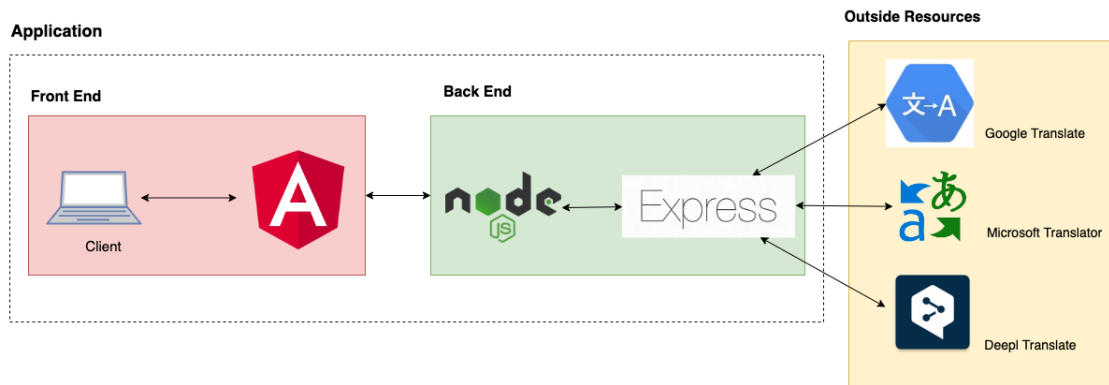


Figure 4.3: System Architecture of the Application

4.2.2 Front-end User Interface

The front-end client side of the application is developed using Angular³, which is a widely-used, open source framework based on TypeScript. The idea of Angular is to build an SPA using different components. Each component has a particular function, and it is loaded onto the main page as needed in the application, which keeps the app running quickly and smoothly. Each component consists of a TypeScript file for coding logic, a HTML file for page layout, and a CSS file for style. A service file is also included which contains the methods necessary for sending data back and forth to the back-end of the application.

Everything the user needs is provided on the main page of the application, so the user does not have to navigate through multiple pages. The application provides digital books in nine different languages for the user to pick from. These books are downloaded as plain text UTF-8 files so they can be uploaded into the HTML file of the application. These are accessible from a collapsible side menu on the page.

To translate a word or phrase, the user simply highlights the text with their mouse. Once the mouse button is released, the highlighted text will be sent to the back-end,

³<https://angular.io/>

where it is translated, and then sent back to the front-end. This translation is then displayed to the user.

An additional JavaScript library, PopperJS⁴, was implemented to create a popover for the translated text. This library provides a tooltip, or popover, relative to the text the user wants to translate. This popover is responsive in that it stays hovered next to the text while the user scrolls through their reading, staying in place until the user clicks away. This implementation brings the translation into the reading itself so the user has easy access. A diagram of the user flow of the application is shown in Figure 4.4.

The following figures depict the layout of the application after the user has agreed to the conditions. Figure 4.5 shows the entire page of the interface along with tags highlighting the important features of the application, as explained below.

The interface in 4.5 is broken down into three functional sections:

1. The collapsible menu: This menu shows all of the languages available to the user. When a user clicks on the language, the menu expands to show three different stories available for the user to choose from in that language. Clicking on a story automatically updates the text on the screen. This feature is shown in greater detail in Figure 4.6.
2. The source text: The main component of the page is the story text itself. The story selected from the menu is displayed in the centre of the screen, and the user can scroll in the box to view more of the text. Highlighting any part of the story text provides the translation as a popover next to the highlighted words. This popover follows the highlighted text even as the user scrolls and only disappears when the user clicks away. This feature is shown in greater detail in Figure 4.7.
3. Survey button: The survey button serves as a link to bring the user to a new page in order to fill out the survey associated with the study. Having the button on the screen makes it easy for the user to find this survey, as well increasing the chance of more responses. This button could be removed for non testing purposes. 4.8

⁴<https://popper.js.org/>

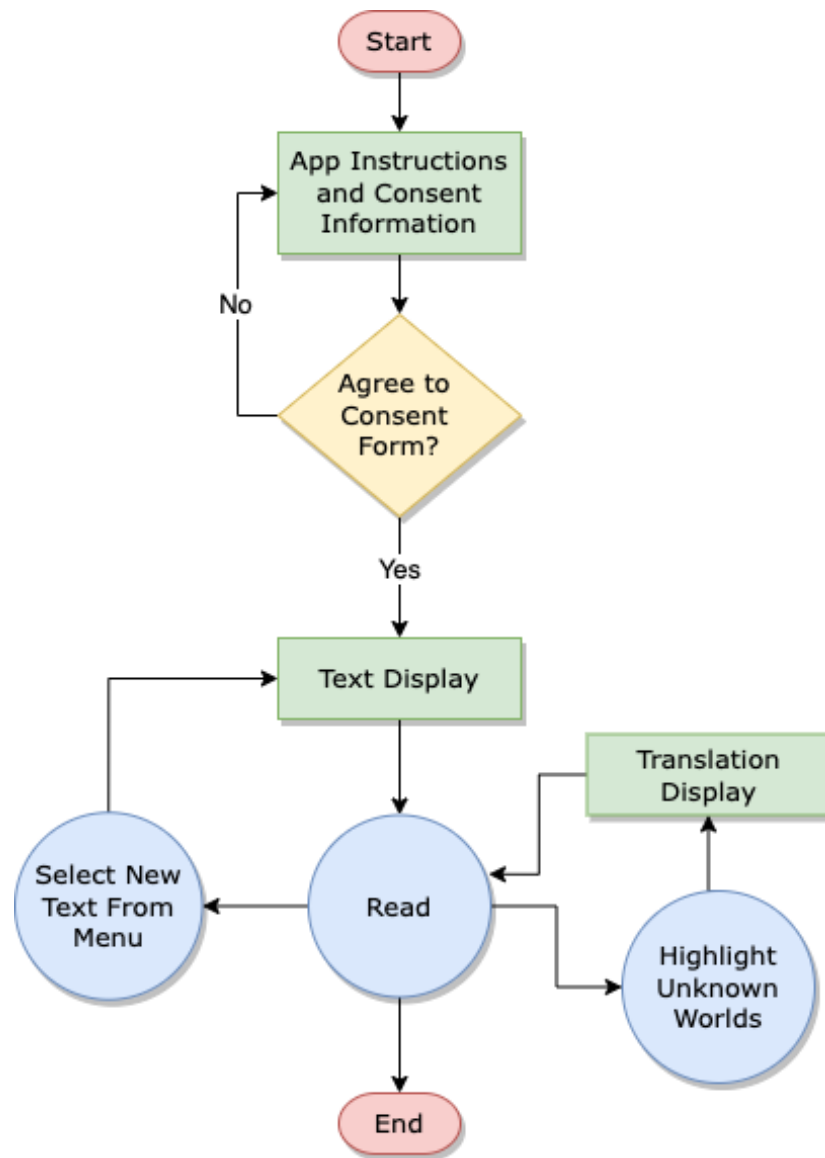


Figure 4.4: User flow of the web application

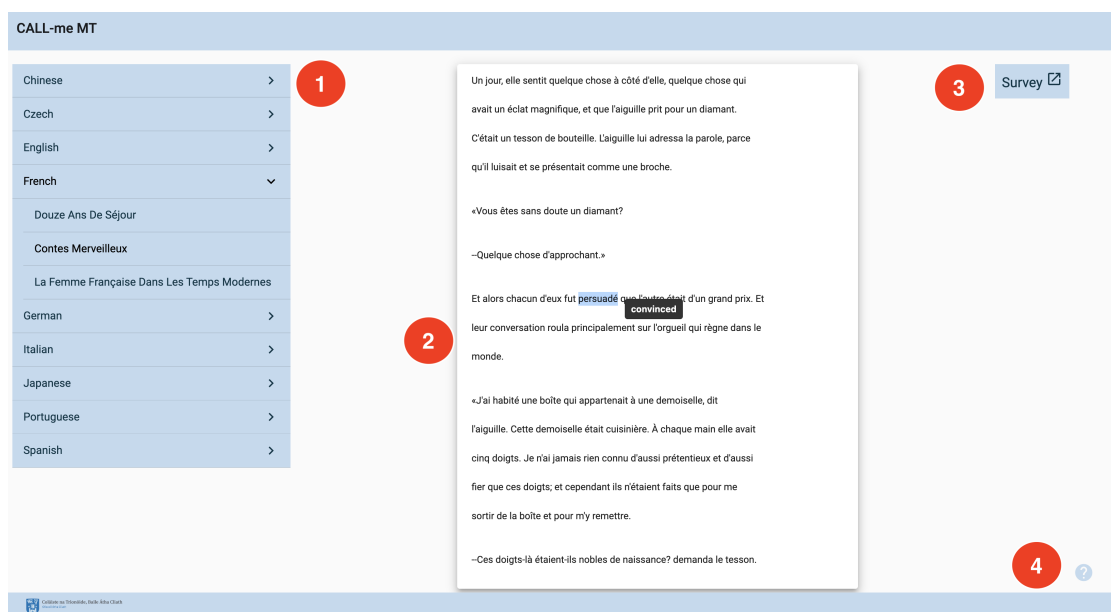


Figure 4.5: User Interface

4. Help button: The help button at the bottom of the screen brings up the initial instructions and consent information displayed when the user first visits the page. This functionality is important for users who need a reminder of how the application works or want to review the terms and conditions of the experiment. This feature is shown in greater detail in Figure 4.9

4.2.3 Back-end Server

The technology stack chosen for the back-end server consists of Node JS⁵ and Express⁶. Node JS is a server side JavaScript framework, and it was chosen for its light-weightness and ease of implementation. Express is a HTTP server framework frequently used with Node JS that provides routing ability. Express was chosen for its ease of sending data to and from the front-end application.

The back-end of the application is responsible for making API calls to the machine translator. All three machine translators (Google, Microsoft, and DeepL) were implemented for this application, but a final one was chosen for user testing purposes.

⁵<https://nodejs.org/en/>

⁶<https://expressjs.com/>

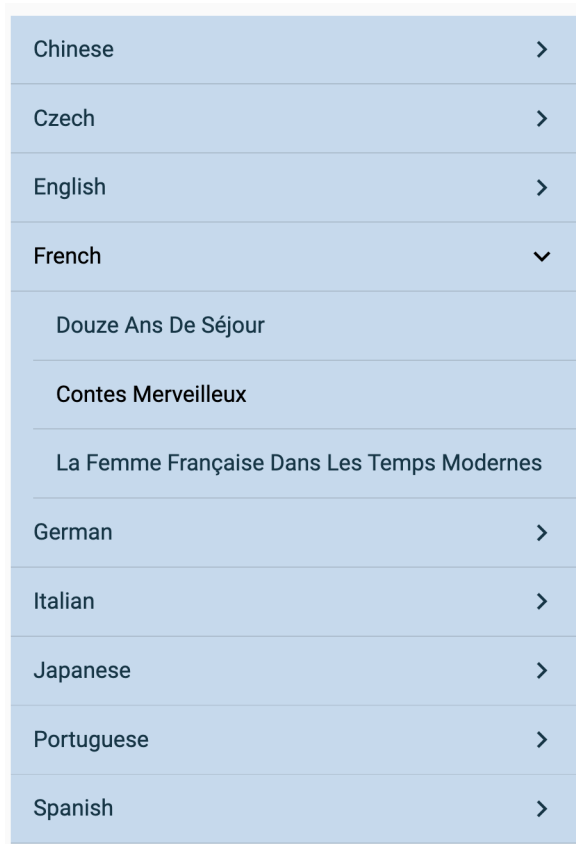


Figure 4.6: User Interface - Story Menu

Et alors chacun d'eux fut persuadé que l'autre était d'un grand prix. Et leur conversation roula principalement sur l'orgueil qui règne dans le monde.

Figure 4.7: User Interface - Story Text

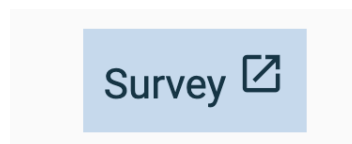


Figure 4.8: User Interface - Survey Button

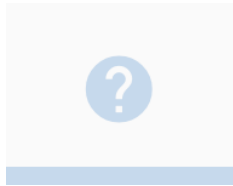


Figure 4.9: User Interface - Help Button

Similar to the API calls in the python translation analysis, the translators are called using HTTP POST requests as in the example Listing 4.1. Once the translated text is sent back to the server, the server forwards it on to the front-end to be processed and displayed to the user.

Listing 4.1: Example of a POST request to Google Translate using NodeJS and Express

```
app.post('/translateGoogle', function(req, res) {
  translateGoogle(req.body.textToTranslate,
                  req.body.targetLanguage)
  .then((translation) => {
    res.json(translation)
  }).catch((err) => {
    console.log(err);
  })
});

// Async function to call google translate
const translateGoogle = async (text, targetLanguage) => {
  try {
    let [response] = await translate
      .translate(text, targetLanguage);
    return response;
  } catch (error) {
    console.log('Google Translate Error: ${error}');
    return 0;
  }
}
```

4.2.4 Deployment and Testing

To make the application available for testing, the front-end code was compiled into a folder with all the files needed for a hosting service to serve the application. This code, along with the back-end code, were uploaded to Heroku⁷ for deployment. Heroku is a platform as a service (PaaS) that allows building, running, and operating applications on the cloud. The application is running as a NodeJS app and can be found at the URL <https://ancient-garden-11501.herokuapp.com/>.

To test the application, volunteers clicked on the link provided to view the application. These volunteers were all adults who were connected with the Trinity College Dublin community. They were instructed to spend at least five minutes reading and translating text in any language that they wished. When they were finished reading, the users were invited to fill out an anonymous survey that asked questions about their experience. This survey was conducted using Google Forms, and each question was optional and anonymous. The questions were centred around topics such as system ease of use, personal language experience, and the reading/translation process.

⁷<https://www.heroku.com/>

Chapter 5

Evaluation and Discussion

This section evaluates the results of the context analysis experiment as well as the user feedback for the CALL system.

5.1 Analysis of Translator Performances

Six language pairs were selected for translating in and out of context, resulting in a total of twelve languages tested. For each language, the source text was input into three different translators (Google Translate, Microsoft Translator, and DeepL Translate), and the results were compared to the sentences in the reference file of the pertaining target language. Three metrics were calculated both in and out of context:

1. Bilingual Evaluation Understudy (BLEU)
2. Translation Error Rate (TER)
3. Character n-gram F-score (ChrF)

The metrics for all three translators are shown in Tables 5.1, 5.2, and 5.3. An overall evaluation of the performances of these different translators is presented in Table 5.4.

One of the first interesting trends to note is that Microsoft Translator received the same scores for both in and out of context translation (Table 5.2). This suggests that the translator possibly does not use context and evaluates input at either the word or sentence level. Testing with more or fewer lines of text did not produce a difference

DeepL	In Context			Out of Context		
Language Pair	BLEU	TER	ChrF	BLEU	TER	ChrF
en ->de	0.322	0.617	0.586	0.335	0.596	0.584
de ->en	0.285	0.494	0.582	0.283	0.559	0.591
en ->cs	0.188	0.734	0.517	0.229	0.689	0.526
cs ->en	0.235	0.575	0.579	0.190	0.602	0.544
en ->ru	0.269	0.591	0.574	0.273	0.552	0.596
ru ->en	0.332	0.575	0.677	0.386	0.496	0.703
en ->zh	0.000	1.300	0.272	0.000	1.300	0.273
zh ->en	0.265	0.618	0.591	0.264	0.620	0.588
en ->ja	0.000	1.400	0.341	0.000	1.400	0.350
ja ->en	0.151	0.894	0.464	0.212	0.916	0.492
de ->fr	0.248	0.544	0.579	0.253	0.562	0.579
fr ->de	0.516	0.388	0.758	0.486	0.432	0.747

Table 5.1: BLEU, TER, and ChrF scores for translating sentences in and out of context with DeepL Translator

Microsoft	In Context			Out of Context		
Language Pair	BLEU	TER	ChrF	BLEU	TER	ChrF
en ->de	0.230	0.654	0.535	0.230	0.654	0.535
de ->en	0.406	0.437	0.656	0.406	0.437	0.656
en ->cs	0.161	0.781	0.475	0.161	0.781	0.475
cs ->en	0.277	0.557	0.596	0.277	0.557	0.596
en ->ru	0.278	0.607	0.591	0.278	0.607	0.591
ru ->en	0.338	0.507	0.655	0.338	0.507	0.655
en ->zh	0.000	1.000	0.374	0.000	1.000	0.374
zh ->en	0.243	0.657	0.556	0.243	0.657	0.556
en ->ja	0.000	1.200	0.378	0.000	1.200	0.378
ja ->en	0.123	0.965	0.387	0.123	0.965	0.387
de ->fr	0.326	0.477	0.608	0.326	0.477	0.608
fr ->de	0.463	0.449	0.731	0.463	0.449	0.731

Table 5.2: BLEU, TER, and ChrF scores for translating sentences in and out of context with Microsoft Translator

Google	In Context			Out of Context		
Language Pair	BLEU	TER	ChrF	BLEU	TER	ChrF
en ->de	0.270	0.600	0.581	0.270	0.601	0.577
de ->en	0.319	0.519	0.606	0.319	0.519	0.606
en ->cs	0.189	0.728	0.494	0.186	0.725	0.489
cs ->en	0.265	0.575	0.558	0.265	0.575	0.560
en ->ru	0.249	0.635	0.560	0.252	0.635	0.560
ru ->en	0.335	0.513	0.748	0.335	0.513	0.748
en ->zh	0.010	1.208	0.369	0.010	1.248	0.367
zh ->en	0.007	0.996	0.009	0.293	0.584	0.594
en ->ja	0.000	2.080	0.383	0.000	2.12	0.387
ja ->en	0.003	0.998	0.011	0.209	0.734	0.466
de ->fr	0.318	0.485	0.610	0.316	0.486	0.609
fr ->de	0.491	0.370	0.732	0.486	0.370	0.731

Table 5.3: BLEU, TER, and ChrF scores for translating sentences in and out of context with Google Translate

in metrics between in and out of context translations. Given that these values are the same for both in and out of context, it is not possible to determine if this translator performs better with additional input text.

For the other two translators, their performances are quite similar. To determine if a translator performed better in or out of context for a given language pair, all three metrics were compared for both the in and out of context translations. The BLEU scores were compared to see if they were higher with or without context. The same was applied to the ChrF score, and the inverse was applied to the TER scores by determining which one was lower of the two translations. The better translation (in or out of context) was determined by a ruling of which translation had the majority of higher scores (Table 5.4). Overall, Google Translate performed slightly better than DeepL in terms of using context. Google performed better using context 5/10 times (2 times there was no difference), and DeepL performed better using context 5/12 times. However, it is important to note that both of these translators performed better using context only about 50% of the time. These values suggest that the addition of context does not seem to have a strong effect on translation accuracy.

There are many possible reasons why some translators perform better using context than others, as well as why context does not seem to make a great impact on translation

	Better In vs Out			
Language Pair	DeepL	Microsoft	Google	Most Accurate
en ->de	OUT	SAME	IN	DeepL
de ->en	IN	SAME	SAME	Microsoft
en ->cs	OUT	SAME	IN	DeepL
cs ->en	IN	SAME	OUT	Microsoft
en ->ru	OUT	SAME	OUT	DeepL
ru ->en	OUT	SAME	SAME	DeepL
en ->zh	OUT	SAME	IN	DeepL
zh ->en	IN	SAME	OUT	DeepL
en ->ja	OUT	SAME	OUT	Microsoft
ja ->en	OUT	SAME	OUT	DeepL
de ->fr	IN	SAME	IN	Microsoft
fr ->de	IN	SAME	IN	DeepL

Table 5.4: Summary of translator performances

results. One possible reason is the amount and kinds of data that the translators are trained with. More data provides more examples the machine can learn to produce translations. Different languages might have different amounts of data available as well, meaning lower resourced languages might not perform as well as others where high amounts of data are available, so some translators might perform better than others depending on the data they use. Finally, the segmenting and parsing of the sentences can pose a problem for automatic evaluation. The translators mostly performed better out of context for languages that do not use the Latin alphabet. This could be due to the fact that sentences are parsed differently in different alphabets, or information is split between more than one sentence, so the translations would perform better in isolation if they are evaluated at a sentence level. Translating from English into both Chinese and Japanese produced very low BLEU scores across the different translators.

In addition to looking at context evaluations, the three translators were compared against each other to determine which one produced the best scores for each language pair overall. The best translator for each language pair was determined by comparing the BLEU, TER, and ChrF scores overall from both in and out of context (Table 5.4). The translator that produced the majority of higher performing scores across all metrics was deemed the best choice for that particular language pair. Overall, DeepL performed the best by producing the highest performing scores for 8 out of the 12 pairs.

Microsoft followed by producing the highest scores 4 out of 12 times. Google did not outperform either of these two translators for each language pair.

Although these metrics provide an easy basis of comparison, it is important to note that translation is a very subjective domain, and it is difficult to determine if one translation is better than another. These scores were the results of the text and translators used at a certain point in time, but as the translators are trained with more data, or if different texts were used, different scores could have been produced. In some cases one translator might produce a better BLEU score, but another might produce a better ChrF score. Because different metrics are being measured and compared, it could be the case that the highest performing translator is not necessarily the best out of the three. For the sake of this research, the results were based on a majority ruling of the different metrics produced among the translators.

Given that DeepL produced the highest scoring metrics out of the three translators, and it performed fairly well for context evaluation, this translator was chosen as the one to incorporate into the CALL platform.

5.2 User Feedback on CALL Platform

The link to the CALL platform was sent out to voluntary participants in order to gain feedback on the system. Participants practiced reading in their target language and then filled out a Google Form survey about their experience. Although only tested by a total of 10 participants, the feedback received provided valuable insight into the use of the application. Regarding the layout of the application, users found the application simple to use and did not have any major issues regarding its use.

For the language learning aspect of the application, all users practiced reading in one of the available romance languages, and their reading skill levels ranged from beginner to advanced. While none of the users had major issues with the translations, there were a couple instances where single words would not translate. This could be due to the fact that the translator did not have enough textual context to translate the word and simply returned the untranslated word back to the user. This feedback is helpful to addressing this type of behaviour in the future. For example, the system could send more text to the translator, in addition to what the user wants to translate, in order to provide this contextual information necessary.

Overall, the users found the platform helpful for better understanding their reading, keeping them engaged, and speeding up their reading by not having to access multiple sources. Using the machine translation proved much faster than referencing a dictionary or other paper-based tools. The answers to all questions of the survey can be found in Appendix 3.

Even if these insights to the applications are helpful in painting a picture of how this tool would be received in the community, the results represent only a small fraction of the language learning community. More studies would need to be carried out before analysing the use and effectiveness of the application.

Chapter 6

Conclusions

6.1 Conclusion

This dissertation looked into the differences in performance of three machine translation systems and how they performed with or without textual context. The machine translation systems chosen for this experiment were Google Translate, Microsoft Translator, and DeepL Translator. Twelve language pairs from the WMT 21 News data set were each tested individually to provide multiple examples of the different systems' performance. Sentences from each language were translated in isolation as well as in the context of the entire document they were taken from. These translations were compared to reference files created by human translators. To score the different outputs, three different metrics were evaluated: BLEU, TER, and ChrF. These scores were used to determine which translator would be used in a CALL application for reading in a foreign language.

According to the calculated metrics, all three systems performed relatively the same regarding the use of context. Google Translate scored a slightly higher BLEU and ChrF score and lower TER score using context than without context, suggesting that more textual information can help in translating text. Microsoft produced the same scores for both in and out of context. DeepL produced higher scores for both in and outside of context, but the values depended on the language pairs tested. Given the consistency and higher performance of the DeepL scores overall, this API was chosen to be implemented into the CALL system. Although these translators performed well

with the additional context, there is still room for improvement and research to be done in order to enhance machine translation performance.

The CALL web application developed for this dissertation consists of providing stories in several different languages for language learners to practice their reading. The goal of the application is to provide quick and painless translations of words and phrases that the learner does not know directly on the screen, by having the learner highlight the unknown phrase with their mouse. This convenient access to translations prevents the learner from becoming discouraged or slowed down during the stop-start process of looking up words.

Ten volunteers tested the platform for their own language learning reading goals. Feedback was recorded in a survey after the users tested the platform. The overall response was positive regarding the platform's ease of use and effectiveness. A few translations issues the users encountered show there is more room for improvement regarding the machine translation, especially how the translators respond with or without the addition of context.

6.2 Limitations

There were a few limitations regarding the contextual translation analysis. Given that the translation APIs have limits in the number of characters that can be translated every month, not all of the available data could be used for analysis, and some experiments had to be spread across several months. Testing the contextual translations with more data might provide a clearer distinction in performance between the use of context or not. In addition, not all of the translators offer support for all languages, so larger data sets would not be supported by all three platforms. Only one data set was used for this project, so it is possible that different types of textual data could influence the results in performance.

For the CALL application, only public domain text for all the languages was used for this experiment, limiting the resources for the user. The platform was also only developed to work on a PC, so it is not available for mobile use. Additionally, all participants read in English or one of the available romance languages, which does not provide insight in the translator performance for the other languages of non-Latin alphabets. Finally, given the small number of participants for testing, the feedback

we received on the platform represented only a small portion of the language learning community.

6.3 Future Work

There are many ways to continue improving these analysis and system. More languages and more translators can be implemented to gain further insight into the performance of such machine translation tools. In addition, more metrics could be taken as well to get a better depth of performance with the use of context. Future work regarding the user application includes allowing the users to pick which translator they want while they read, more language and stories options, and alternate translations regarding feedback from the survey. Even more tools could be implemented to add textual or situational context to a story in order to improve the automatic translation of its text.

Regarding the contextual analysis, there are many ways in which this experiment could be expanded upon in the future. Including more translators and more languages would help provide more data to analyse the overall performance and coverage of particular translators. In addition, testing the use of context with more and different types of data might impact the results. This project used news sources as textual data, but social media or other textual resources might produce a different output.

While there are many features to be added or directions to take, this project produces a reliable, simple platform that brings language learning back to the learner to enhance their experience and make language learning easier and more enjoyable.

Bibliography

- [1] W. A. Renandya, “The power of extensive reading,” *RELC journal*, vol. 38, no. 2, pp. 133–149, 2007.
- [2] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *CoRR*, vol. abs/1609.08144, 2016.
- [3] A. K. Melby and C. Foster, “Context in translation: Definition, access and teamwork,” *Translation & Interpreting, The*, vol. 2, no. 2, pp. 1–15, 2010.
- [4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
- [5] M. Popović, “chrf: character n-gram f-score for automatic mt evaluation,” in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, 2015.
- [6] R. Blake, “Technology and the four skills,” *Language Learning & Technology*, vol. 20, no. 2, pp. 129–142, 2016.
- [7] M. Ward and J. Genabith, “Call for endangered languages: Challenges and rewards,” *Computer Assisted Language Learning*, vol. 16, no. 2-3, pp. 233–258, 2003.

- [8] N. Ní Chiaráin, O. Nolan, M. Comtois, N. Robinson Gunning, H. Berthelsen, and A. Ni Chasaide, “Using speech and NLP resources to build an iCALL platform for a minority language, the story of an scéalaí, the Irish experience to date,” in *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, (Dublin, Ireland), pp. 109–118, Association for Computational Linguistics, May 2022.
- [9] A. A. Zarei and M. Hashemipour, “The effect of computer-assisted language instruction on improving efl learners’ autonomy and motivation,” *Journal of Applied Linguistics*, vol. 1, no. 1, pp. 40–58, 2015.
- [10] R. Vesselinov and J. Grego, “Duolingo effectiveness study,” *City University of New York, USA*, vol. 28, no. 1-25, 2012.
- [11] K. Teske, “Duolingo,” *calico journal*, vol. 34, no. 3, pp. 393–401, 2017.
- [12] G. Altynbekova and R. Zhussupova, “Mobile application fluentu for public speaking skills development,” in *SHS Web of Conferences*, vol. 88, p. 02008, EDP Sciences, 2020.
- [13] R. Stroud, “Student engagement in learning vocabulary with call.,” *Research-publishing. net*, 2014.
- [14] M. R. Setiawan and P. Wiedarti, “The effectiveness of quizlet application towards students’ motivation in learning vocabulary,” *Studies in English Language and Education*, vol. 7, no. 1, pp. 83–95, 2020.
- [15] M. Alharbi, “Quizlet: Promoting learners’ content literacy,” *Computer Assisted Language Learning*, vol. 22, no. 3, pp. 172–178, 2021.
- [16] C. A. Chapelle, “The relationship between second language acquisition theory and computer-assisted language learning,” *The modern language journal*, vol. 93, pp. 741–753, 2009.
- [17] P. De’Ath, “The niue literacy experiment,” *International Journal of Educational Research*, vol. 35, no. 2, pp. 137–146, 2001.

- [18] A. Niño, “Machine translation in foreign language learning: Language learners’ and tutors’ perceptions of its advantages and disadvantages,” *ReCALL*, vol. 21, no. 2, pp. 241–258, 2009.
- [19] N. Resende, B. Cowan, and A. Way, “Mt syntactic priming effects on l2 english speakers,” 2020.
- [20] J. Hutchins, “Machine translation: A concise history,” *Computer aided translation: Theory and practice*, vol. 13, no. 29-70, p. 11, 2007.
- [21] D. Torregrosa, N. Pasricha, M. Masoud, B. R. Chakravarthi, J. Alonso, N. Casas, and M. Arcan, “Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models,” in *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pp. 125–133, 2019.
- [22] P. Williams, R. Sennrich, M. Post, and P. Koehn, “Syntax-based statistical machine translation,” *Synthesis Lectures on Human Language Technologies*, vol. 9, no. 4, pp. 1–208, 2016.
- [23] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, and P. S. Roossin, “A statistical approach to machine translation,” *Computational linguistics*, vol. 16, no. 2, pp. 79–85, 1990.
- [24] C. M. Bishop, “Neural networks and their applications,” *Review of scientific instruments*, vol. 65, no. 6, pp. 1803–1832, 1994.
- [25] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.
- [26] J. Yang, S. Ma, H. Huang, D. Zhang, L. Dong, S. Huang, A. Muzio, S. Singhal, H. H. Awadalla, X. Song, *et al.*, “Multilingual machine translation systems from microsoft for wmt21 shared task,” *arXiv preprint arXiv:2111.02086*, 2021.
- [27] L. Volkart, P. Bouillon, and S. Girletti, “Statistical vs. neural machine translation: A comparison of mth and deepl at swiss post’s language service,” in *Proceedings of the 40th Conference Translating and the Computer*, pp. 145–150, 2018.

- [28] M. Freitag and O. Firat, “Complete multilingual neural machine translation,” *arXiv preprint arXiv:2010.10239*, 2020.
- [29] P. Vergés Boncompte and M. Ruiz Costa-Jussà, “Multilingual neural machine translation: Case-study for catalan, spanish and portuguese romance languages,” in *EMNLP 2020, Fifth Conference on Machine Translation: November 19-20, 2020, online: proceedings of the conference*, pp. 447–450, Association for Computational Linguistics, 2020.
- [30] R. Bawden, *Going beyond the sentence: Contextual Machine Translation of Dialogue*. PhD thesis, Université Paris-Saclay (ComUE), 2018.
- [31] S. J. Russell, *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.
- [32] J. House, “Text and context in translation,” *Journal of pragmatics*, vol. 38, no. 3, pp. 338–358, 2006.
- [33] L. Wang, Z. Tu, A. Way, and Q. Liu, “Exploiting cross-sentence context for neural machine translation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (Copenhagen, Denmark), pp. 2826–2831, Association for Computational Linguistics, Sept. 2017.
- [34] Z. Tu, Y. Liu, S. Shi, and T. Zhang, “Learning to Remember Translation History with a Continuous Cache,” *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 407–420, 07 2018.
- [35] L. Miculicich, D. Ram, N. Pappas, and J. Henderson, “Document-level neural machine translation with hierarchical attention networks,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 2947–2954, Association for Computational Linguistics, Oct.-Nov. 2018.
- [36] Y. Graham, T. Baldwin, A. Moffat, and J. Zobel, “Can machine translation systems be evaluated by the crowd alone,” *Natural Language Engineering*, vol. 23, no. 1, pp. 3–30, 2017.

- [37] M. Post, “A call for clarity in reporting bleu scores,” *arXiv preprint arXiv:1804.08771*, 2018.
- [38] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A study of translation edit rate with targeted human annotation,” in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231, 2006.
- [39] L. Specia, N. Cancedda, and M. Dymetman, “A dataset for assessing machine translation evaluation metrics.,” in *LREC*, 2010.
- [40] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, (Ann Arbor, Michigan), pp. 65–72, Association for Computational Linguistics, June 2005.
- [41] WMT21, “Wmt21 news systems and evaluations,” 2021.
- [42] J. Webster and J. S. Ahuja, “Enhancing the design of web navigation systems: The influence of user disorientation on engagement and performance,” *Mis Quarterly*, pp. 661–678, 2006.
- [43] H. L. O’Brien and E. G. Toms, “What is user engagement? a conceptual framework for defining user engagement with technology,” *Journal of the American society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008.
- [44] S. Holmes *et al.*, *Getting MEAN with Mongo, Express, Angular, and Node*. Simon and Schuster, 2019.

Appendix

.1 Link to CALL Application

CALL-me MT platform hosted on Heroku:

<https://ancient-garden-11501.herokuapp.com/>

.2 GitHub code

.2.1 Code for the web application

<https://github.com/maddiecomtois/CALL-me-MT>

.2.2 Code for context analysis

https://github.com/maddiecomtois/CALL-me_MT_TranslatorAnalysis

.3 Survey responses

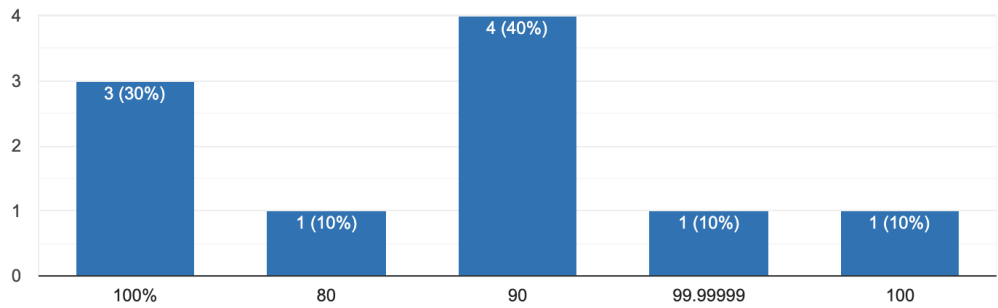
Responses to the survey questions asked after volunteers used the CALL-me MT platform.

Section 1: Ease of Use

To what percentage did you find the system easy to use and navigate? (1-100%)



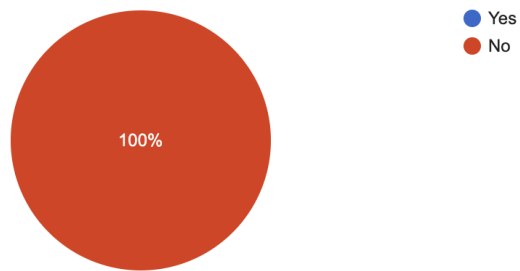
10 responses



Did you have any major issues or problems with using the system?



10 responses



If you had any issues, please explain

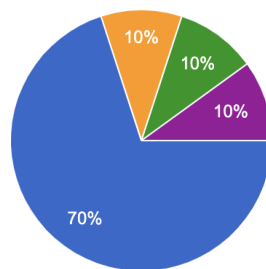
1 response

Didn't know you had to double click. Had to figure out how to do phrases. not all the words translated, sometimes they just appeared and were exactly the same in the same language

Did the application translate the text you wanted to translate?

 Copy

10 responses

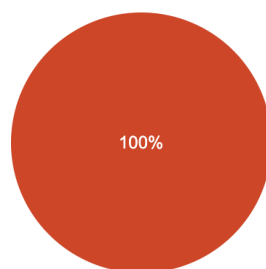


- Yes
- No
- Most of the time yes, but sometimes for shorter segments it would not, e.g. it would translate 'aux cimes' to 'aux cimes' and 'repose' to 'repose'.
- see above
- Sometimes the translation wouldn't work for some individual words, but once you highlighted the whole phrase, then it w...

Did the design of the system hinder your language learning in any way?

 Copy

10 responses



- Yes
- No

If yes, please explain

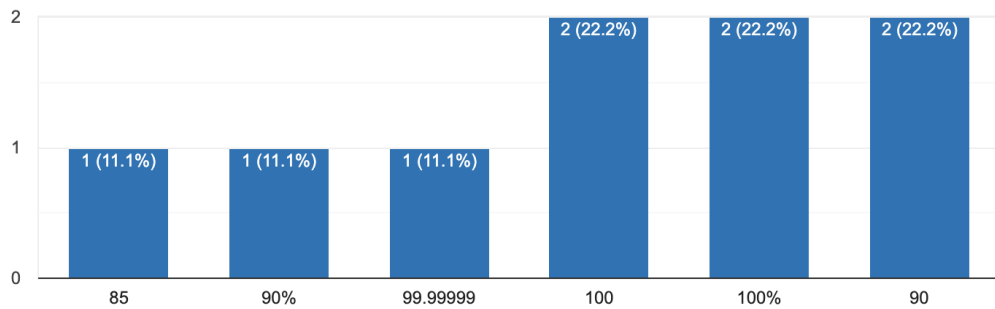
0 responses

No responses yet for this question.

To what percentage was it enjoyable to use the system?

 Copy

9 responses



Do you have any feedback regarding the use or layout of the system?

4 responses

It's great !

Good layout, not cluttered and it's easy to see how to navigate everything!

Need more instructions?

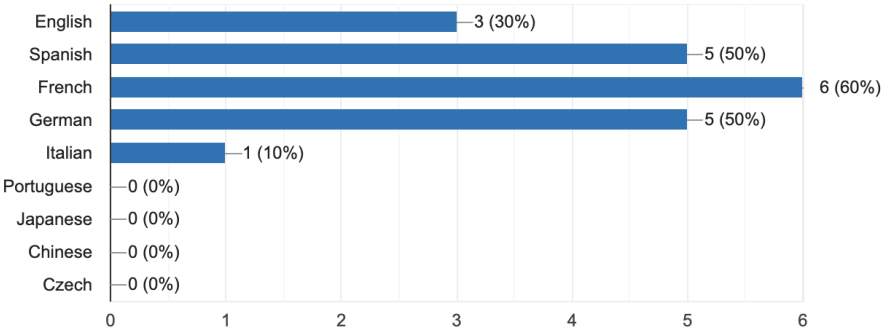
Suited me fine reading black on white, for others perhaps an option to change the background colour may help

Section 2: Language and Experience

Which language(s) did you read in?



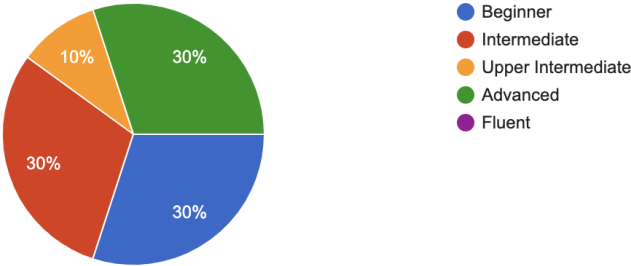
10 responses



What would you consider to be your reading level?



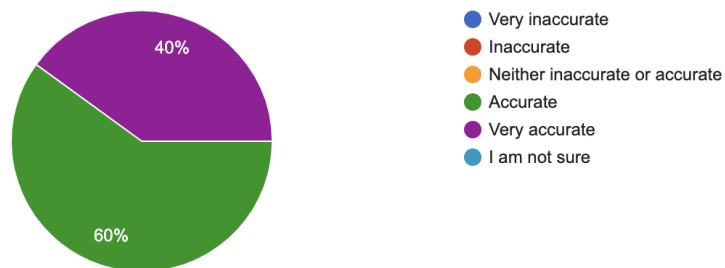
10 responses



How would you rate the accuracy of the machine translation?

 Copy

10 responses



Do you have any feedback regarding the foreign language and translation support of the system?

5 responses

Good translations! Would have said 'very accurate' except for the couple times it translated french -> french.


I guess you need to know more about the language if your trying to do whole phrases? Word translation seemed OK, but sometimes the answer given was the exact same word.

Excellent feedback once your proficiency was high enough to select the appropriate chunks of text, e.g. I tried out of curiosity to translate 'ont' and it returned 'ont', as if that was the English translation. A suggestion of chunks would be useful for less proficient readers, e.g. 'objet' -> 'object' but 'pour objet' -> 'for the purpose of'. Speed instantaneous, added greatly to the experience.

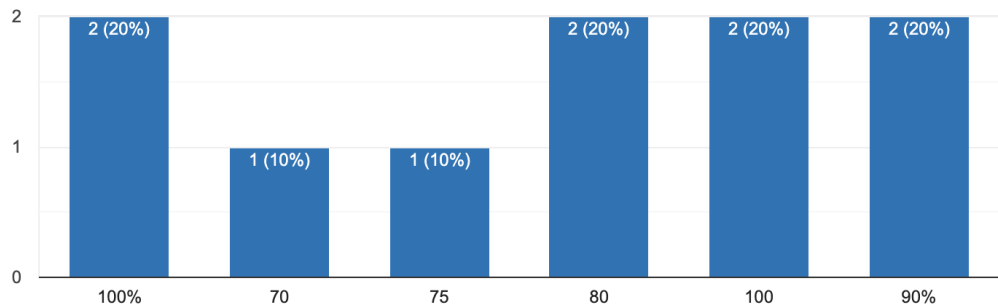
I thought it worked really well. The only thing that could possibly help even more would be to include alternate translations for words with multiple meanings (ex: derecho, esperar)


Single words sometimes didn't translate

Section 3: Reading and Translation

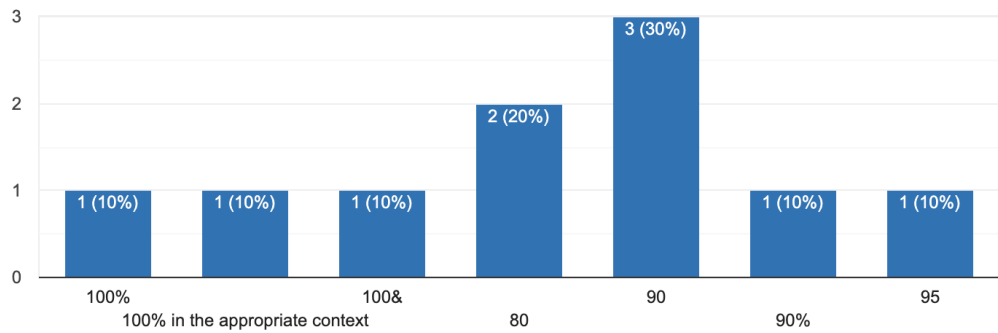
To what percent was the provided translation helpful for understanding the context of the reading?  Copy

10 responses



To what percentage do you think machine translation tools are helpful?  Copy

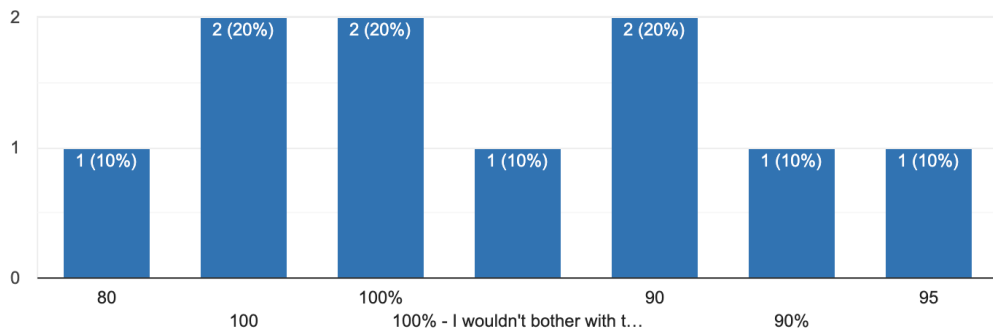
10 responses



Relative to a manual dictionary look up, to what percentage was the automatic translation helpful?



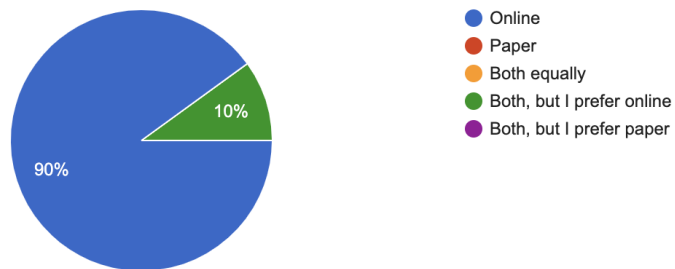
10 responses



When looking up a word in a foreign language, in general are you more likely to use online tools or paper tools?



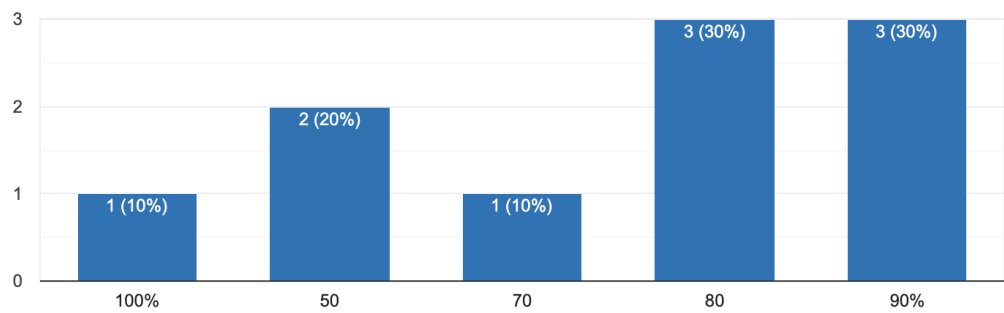
10 responses



To what percentage did using the system's built-in machine translation tools make your reading faster than if you were to look up a word somewhere else?



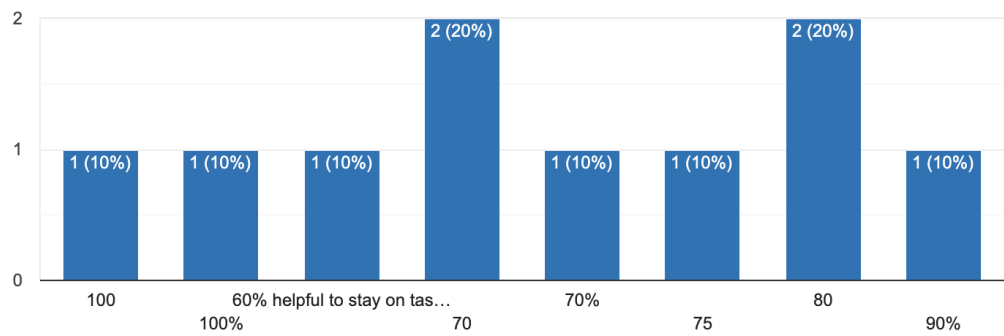
10 responses



To what percentage do you think continued use of this system could make it easier to read in a foreign language?



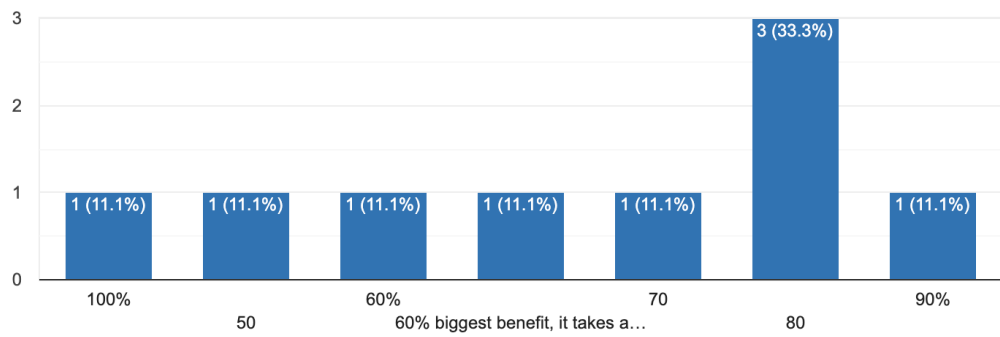
10 responses



To what percentage do you think continued use of this system could make it easier to learn a foreign language overall?



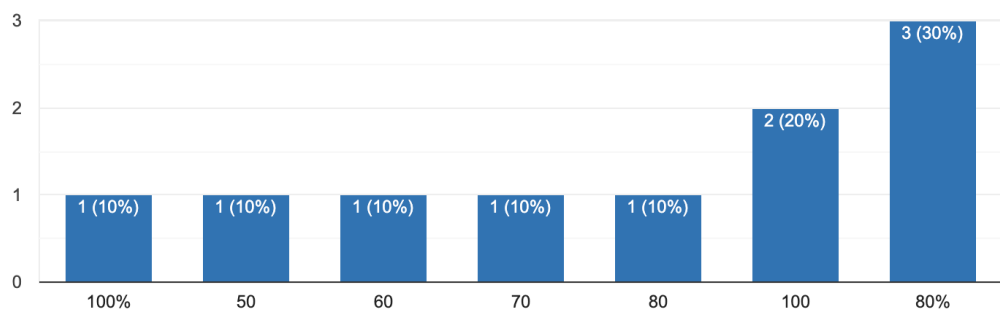
9 responses



To what percentage would you use this system regularly if it were available?



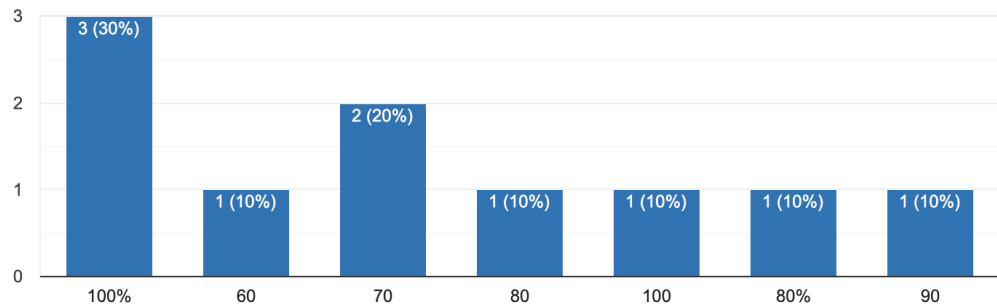
10 responses



To what percentage do you think continued use of this system could enhance your enjoyment of learning to read in a foreign language?



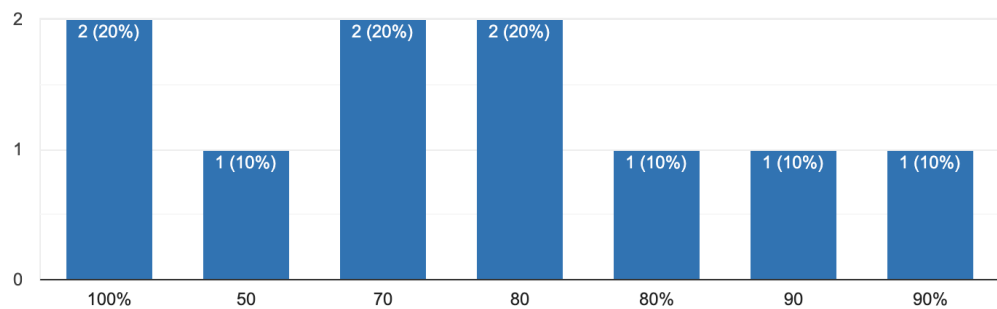
10 responses



To what percentage do you think continued use of this system could improve your confidence in learning foreign language?



10 responses



Do you have any feedback regarding how the system affected your language learning?

3 responses

Highlighted a German sentence & words were slightly mixed up, but still got the meaning of the sentence. Would definitely be good for language learning.

It definitely speeds up reading so is great.

Very useful - the instant feedback is excellent, neat presentation, very nice.