Integration of linguistic tools into a language learning mobile application

Joseph Carnec, BAI

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

Science (MCS)

Supervisor: Fergal Shevlin

April 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.

Joseph Carnec

April 19, 2022
Permission to Lend and/or Copy

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

Joseph Carnec

April 19, 2022
Integration of linguistic tools into a language learning mobile application

Joseph Carnec, Master of Science in Computer Science
University of Dublin, Trinity College, 2022

Supervisor: Fergal Shevlin

This dissertation is an exploration of how linguistic assistance tools such as Machine Translation and crowd-sourced linguistic resources can be used to provide informal, and personalised language learning opportunities on a mobile devices. We present an application which allows language learners to specify and discover personalised reading material which fits their language learning needs. The presented application additionally provides structured revision of words and phrases encountered during foreign language reading through intentional vocabulary learning exercises which teach words and phrases in context. We present a novel method of "syntax highlighting" translation output with the aim of facilitating the use of Machine Translation for foreign language learning. Lastly we evaluate the presented application under the scope of personalised mobile language learning and mobile language learning aimed at migrants. We conclude that the described system addresses a number of the needs and facilitates a number of the informal language learning practices of migrants which have been identified in previous studies.
Acknowledgments

I would like to thank to my supervisor, Fergal Shevlin, for his continued support, advice and mentoring throughout the course of this dissertation.

Finally, I’d like to thanks my parents, thank you.

Joseph Carnec

University of Dublin, Trinity College
April 2022
## Contents

**Abstract** iii

**Acknowledgments** iv

**Chapter 1 Overview of the dissertation**
- 1.1 Aims and Overview ................................................. 1
- 1.2 Overview ........................................................... 2

**Chapter 2 Introduction**
- 2.1 Motivation: The differences of language learning by necessity and by choice 8
- 2.2 The use of linguistic assistant tools to overcome communication challenges, but also to provide language learning opportunities ............................................. 10
- 2.3 Integrating linguistic assistant tools into a language learning system to meet the needs of migrants ................................................................. 12
- 2.3.1 Identified needs and practices and how the implemented application attempts to address these ................................................................. 13
- 2.4 Overview of application and discussion of implementation .......................... 15
- 2.4.1 Choosing L1 and L2 and the limitations of the open-source machine translation models ................................................................. 19
- 2.4.2 Defining subscriptions by RSS feed ........................................... 21
- 2.4.3 Defining single words using translators and dictionaries .................. 22
- 2.4.4 Constructing dictionary definitions from the Wiktionary data dump ........ 23
- 2.4.5 Translating multi-word units to inspect language chunks .................. 24
- 2.4.6 The potential of "Part Of Speech highlighting" natural text .......... 25
- 2.4.7 Optimizing "Part Of Speech tagging" text on the web ...................... 25
- 2.4.8 Adding words and phrases to the "learn list" and highlighting instances of them on the web ................................................................. 26
- 2.4.9 A need for mobile extensions to facilitate augmenting the web on mobile devices ................................................................. 27
4.4 Varying forms of idiomaticity ........................................... 64
4.5 MWEs: Idiosyncratic language use which needs to be considered .... 66
4.6 Considerations to be made in order to teach idiomatic language chunks .... 67

Chapter 5  Facilitating vocabulary acquisition of words and phrases within informal language learning on mobile devices. 72
5.1 Incidental language learning ................................................ 72
5.2 The need for a tool which allow for repetitive encounters with words in clear contexts: ...................................................... 73
  5.2.1 The frequency affect: .................................................. 74
  5.2.2 The context affect: ..................................................... 75
  5.2.3 The involvement factor: an advantage of interacting with text ..... 77
  5.2.4 Using flashcards to provide frequent exposure to context: ....... 81
5.3 Annotating Parts of Speech in translation outputs: Aiding beginner learners engage with text ............................................. 82
  5.3.1 Providing a rich context in which to infer meaning and to expose beginners to a language’s syntax. .................................. 83
  5.3.2 POS highlighting to make it easier to read translation output and to simulate translation alignment learning ...................... 84

Chapter 6  Conclusions and future work 86
6.1 Conclusions ................................................................. 86
6.2 Future Work .............................................................. 88

Appendices 98
List of Tables
List of Figures

2.1 Architecture of the application deployment ........................................... 18
2.2 The user can specify L1 and L2 ................................................................. 19
2.3 The language being selected ................................................................... 19
2.4 This updates all other relevant information within the app, such as sub-
subscriptions and the 'learn-list'. The subscriptions can be viewed from the
menu drawer .................................................................................................. 19
2.5 Subscriptions can be specified by topic (Interest), key-word or website ..... 21
2.6 Subscribing to a feed allows users to access articles ................................. 21
2.7 The user can select a subscription and browse that feed, or as seen here
browse by all subscriptions ........................................................................ 21
2.8 Select words to translate them. If the selected word is present on Wiki-
tionary, a definition is available to user, this indicated by a green outline
of the 'translation box' .................................................................................. 22
2.9 Clicking on the green underlined box bring up the Wiktionary dictionary
definition ....................................................................................................... 22
2.10 Some Wiktionary entries provide a sound clip of a native speaker uttering
the word. ........................................................................................................ 22
2.11 More than word can be highlighted, in this case the multi-word token "a
traves de" is a lexical entry in Wiktionary and as such the green outline
indicates that pressing the green box will provide a dictionary definition
for the phrase ................................................................................................. 24
2.12 An arbitrary amount of text can be selected, here a sentence was selected,
the translation is shown, the red outline indicates that now dictionary entry
exists for the selected text ............................................................................ 24
2.13 The highlighter icon can be pressed to turn on part of speech highlight-
ing, this makes the output more readable and provides information about
syntax in L1 in relation to L2. Additionally we can see that it facilitates
comparison of phrases .................................................................................. 24
2.14 Words and phrases added to the ‘learn-list’ provide an ability to see bring attention to words the user is trying to learn. Words which have a dictionary definition are highlighted in yellow, otherwise they are highlighted in green.

2.15 Words which have been added to the word list can be viewed, along with their “score” a tally of how the user did on flash card exercises for each entry.

2.16 Phrases can be viewed separately, from this screen users can delete words which they are satisfied they know.

2.17 Flash card exercises can be done by phrase or words, they come in sets of 10 randomly with words that have been answered worst coming more frequently. As we can see here, the word can appear with different inflections.

2.18 In the case of phrases the translation of the context from which the word was first added is used as the reference for the “meaning” of the phrase.

2.19 Here we see that although “a traves de” is multi word, it has a dictionary definition in Wiktionary, as such it appears as a flash card in the words section. Here we see that examples fetched for the flash card exercise can deal with multiple examples and in different inflexions.

2.20 In the case of words, the dictionary definition is presented such that the user can evaluate how well he has knows the word, the utterance can be listened to.

2.21 The flash card exercise presents sentences containing the word or phrase, removing one word from the sentence and presents four alternatives for the user to choose.

2.22 To help the user can view the translation if he chooses to, or listen to a native speaker utter the sentence.

3.1 The ANT application presenting a defined set of news broadcasts by level.

3.2 The application allows users to view annotations created by other peers while reading articles selected by teachers.

4.1 Two example sentences with idiomatic word combination annotated Subscripts and text coloring indicate strong multiword groupings; superscripts and underlining indicate weak groupings. Boxes indicate gaps. Schneider et al. [2014].

4.2 “In short” is translated to “Bref”.

4.3 “Call him up” is translated to “l’appeler”.

4.4 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

4.5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

x
Chapter 1

Overview of the dissertation

1.1 Aims and Overview

This project aims to explore how Machine Translation and crowd-sourced linguistic resources can provide language learning opportunity’s on a mobile device. For this exploration we chose to evaluate the potential affordances of such a system based on the needs of migrant language learners. Specifically we identify that migrant language learners use online linguistic tools such as online machine translation and dictionaries to overcome communication challenges and to engage with foreign languages throughout their day providing: ”educationally valuable activities across everyday environments that are not designed or arranged for learning” (Kukulska-Hulme Agnes 2019).

As such the aim of this project is to explore how linguistic tools can be integrated into a mobile language learning application to facilitate some of the language learning practices and address some the language learning needs of migrant language learners.

We will try and explore how:

- By using Machine Translation and crowd-sourced lexical resources, we can facilitate and improve the informal language learning practice of reading in a foreign language, resulting in better incidental vocabulary acquisition.

- Providing ways for language learners to read personalised content on the web allows for this practice to be engaging and purposeful, in line with the definition of personalized mobile learning outlined in (Kukulska-Hulme 2016).

- Mobile assisted language learning exercises which are present in popular language learning applications can be improved by:
– Allowing these exercises to be generated automatically based only on words and phrases which the user has specified they want to learn.
– Teaching these words and phrases in context, generated automatically using crowd sourced lexical resources.
– Harnessing these crowd sourced lexical resources can augment these exercises automatically by allowing for the user to listen to native speakers utter sentences and words in these exercises.

• By translating multi word units rather simply defining single words Machine Translation allows for language learning and vocabulary acquisition to occur at the phrase level, which is important for effective and idiomatic language learning (Godwin-Jones 2015; Contreras Kallens and Christiansen 2022).

• The merits of ”personalised mobile language learning” Kukulska-Hulme (2016) pertain to the fact that it allows for informal and personalised learning outside of a classroom context which can fit the complex needs of language learners who face communication challenges daily. We will thus investigate how the personalised and informal language learning approach of our application can fit some of the needs of migrant language learners with respect to previous investigations into their unmet needs and common practices. We will also discuss some important needs which are not met by personalised learning technologies.

1.2 Overview

In Chapter 2 We provide an outline of the language learning practices and needs of migrants. We then underline how this dissertation proposes to address some of these needs and facilitate some these practices by embracing online machine translation and crowd sourced linguistic tools such as Tatoeba. The chapter concludes by presenting the mobile application implemented as part of this project which combines these solutions in an effort to address some of the mobile language learning needs of migrants and facilitate some of their informal learning practices. During this overview of the application we present a number of implementation details which are relevant to mobile applications which are trying situate language learning on the web.

• In section 2.1 we describe how migrant language learners have specific language learning needs.

• In section 2.2 We discuss the focus on the potential of mobile learning in assisting migrants with language learning, we then argue that despite the widespread use of
linguistic assistants tools such as machine translation and dictionaries in 'informal' language practices such as reading books, using social media, reading web articles and watching videos, there has been very little research on how these tools can be integrated into a mobile language learning environment

- In section 2.3 we describe a number of findings regarding the unmet language learning needs and common language learning practices of migrants, for each of these finding we describe how this dissertation will explore ways to address those needs and facilitate those practices.

- In section 3.4 we present the features implemented as part of this dissertation, we additionally give an overview how this application provides a "language learning pipeline" to define, discover, engage with, and learn authentic language. During this section we discuss challenges and considerations which effect the scope of this implementation

In chapter 3 we provide background on the nature of mobile assisted language learning and why there has been some focus on the applications of mobile learning to assist migrants with their language learning needs as well as to overcome general communication challenges. Within the context of mobile learning we then outline the need for personalised learning opportunities, and the opportunity which exists in providing ways for language learners to engage with authentic language use in an informal way. We outline affordances of mobile devices for informal and personalised language learning. We then give an overview of previous literature on machine translation and how it can be used in the context of informal language learning to allow language learners to engage with authentic material. We describe the importance of taking into consideration the needs of target users, namely migrants, to create language learning applications which best fit the wide variety of goals and motivations which language have to learn languages. We then describe two previous implementations of applications which provide personalised and informal language learning opportunities. To identify the needs of migrant language learners we then describe the findings of two studies which performed semi-structured interviews regarding the tacit and latent needs of migrant learners and discuss their findings in relation to the aims of this study. Finally we discuss how the application presented in this paper addresses those needs and practices.

- In section 3.1 we analyse informal language learning and we argue that as mobile learning is inherently outside the classroom and self driven, there is a need to provide personalised learning which motivates users to learn in a flexible and positive way. We present evidence that gamification used by popular apps takes away from
learning. We propose that integrating language learning into daily digital media consumption through the use of translation and other linguistic resources is a possible solution to providing engaging and personalised language learning. Lastly we outline recent technological developments which makes this approach possible.

- In section 3.2 we give an overview of previous literature on the use of machine translation for language learning. We outline that despite the fact that online machine translation has been seen as a disruption to classroom contexts, it can provide the ability for language learners to engage in incidental and intentional vocabulary acquisition with content which is highly engaging, highly specific to their needs and at a lower level of proficiency.

- In section 3.3 we describe the advantages of mobile devices for informal language learning. We identify three main features of mobile phones which have an applicability to informal language learning, these are described below along with a discussion of how these features can be pertinent to migrant learners.

- In section 3.4 we underline the importance of user driven design for language learning tools, particularly for migrants. We then outline the importance of taking into account previous implementations of language learning apps and how they integrate these into their language learning practices. We then describe how participatory design strategies can be used to create an application which is fit for purpose.

- In section 3.5 we present two previous implementations of language learning applications which feature personalised, informal learning strategies. We illustrate how the findings of these implementations relate to the aims of this dissertation.
  - We discuss The Audio News Trainer, an app which provides level aware news broadcasts to users.
  - We discuss Hsu et al. (2013), an application where the manual selection of reading material by teachers is combined with collaborative annotation of text for reading comprehension support.

- In section 3.6 we present two studies which investigate the needs of migrants. This serves as a way to explore the tacit and latent needs of migrants based on previous research. We provide a discussion of these finding and relate them to the aims of this study.

In chapter 4 we try and address some of the unmet language learning needs of migrants identified in previous research, these are the need tools which facilitate fluent sounding and idiomatic language learning, such as the use of expressions and phrases in the
right socio-cultural context. We introduce the view of language as idiomatic and statistically frequent word chunks, and how these conventionalised multi word linguistic units are semantically, lexically, statistically and syntactically idiosyncratic. We argue that for a language learning system to be effective and provide a high degree of proficiency and fluency, language learners need to be able to notice these phrases and word chunks, evaluate their meaning and encounter them repeatedly in different contexts. We then provide a number of considerations which need to be taken into account for a language learning system to provide the tools for language learners to interpret and use these word combinations correctly such as seeing them repeatedly in different contexts.

- In section 4.1 we introduce the view that natural natural language use does not only pertain to the combination of lexical units with grammar, but instead is formed of statically frequent collocations and expressions with arbitrary meanings.

- In section 4.2 We describe Multi Word Expressions (MWEs), statistically frequent multi word units which behave idiosyncratically. We provide examples of MWEs within natural language use and outline why they can only be learned by example in context.

- In section 4.3 We outline the importance of learning MWEs for language fluency, and state the importance elaborating on the characteristics of MWEs to design tools for language learners to learn these idiomatic multi word units.

- In section 4.4 we outline the ways in which multiword expressions differ from natural language use.

- In section 4.5 we argue why a digital language learning environment should take into account MWEs and as such allow ”learners of the language notice these patterns and replicate them” (Scholz, 2017).

- In section 4.6 we outline a number of considerations which need to be taken into account to allow language learners to learn language to high degree of proficiency by repeatedly exposing themselves to these idiomatic language chunks. We provide examples of how the application prototype addresses these concerns.

In Chapter 5 in order to facilitate context specific vocabulary acquisition which is one of the primary concerns of migrant language learners (Epp, 2017), we set out to outline how the application provides numerous strategies to benefit vocabulary acquisition while reading. We first introduce intentional and incidental vocabulary acquisition, and how strategies which increase the frequency with which words are encountered and the quality of those contexts improve incidental learning retention rates. We then describe how
intentional exercises can further improve retention rates of words encountered. Finally we present a novel approach to presenting translation output and we argue that it allows for more deep processing of the output by simulating a word mapping, allowing learners to engage with text at lower levels of proficiency and evaluate more words and phrases with respect to the L1 language.

- In section 5.1 we introduce the concepts of formal learning, informal learning, and incidental learning, presenting arguments which point to the importance of incidental learning within language acquisition and situates the terms within the context of mobile assisted language learning (MALL).

- In section 5.2 we present three ways of facilitating incidental vocabulary acquisition through reading comprehension for language learners. We identify three factors which affect the rate of retention for incidental language learning; these are the frequency of encounter, the quality of context in which these words are encountered and the involvement with which the meaning of words are evaluated. We observe that this is of particular interest in the context of digital reading where browsing and skim reading is common.

- In section 5.3 we present the effect of word exposure frequency to incidental vocabulary acquisition and how a mobile learning app may increase this rate of exposure for a set of vocabulary which a language learner is trying to learn. We also explore how highlighting words may alert the reader’s attention to that word so that they may focus on the occurrence of that word in its context, providing a deeper processing of the text and thus perhaps increasing the rate of word retention.

- In section 5.4 we presents how part of speech tags can be used to show syntax highlighting of natural text which we describe as “part of speech highlighting”. We argue that this can be used to expose beginner language learners to the syntax of a language in the early stages of language learning and to enrich the context in which words can be inferred, allowing for the inferall of words at a lower level of proficiency. This section additionally presents a novel method of presenting translation output which allows easier reading of translations and invites learners to construct their own word mappings. We propose that this may have some pedagogical merits by inviting the user to perform deep processing of words in reference to translation in a known language and we draw parallels to the act of observing and creating word mappings of historical languages.

In chapter 6 we outline the main conclusions from this projects, challenges faced and future work that could be taken on to try and integrate linguistic tools such as machine
translation into language learning applications in order to facilitate the language learning practices and address the language learning needs of migrants.
Chapter 2

Introduction

This chapter introduces mobile assisted language learning as well as the unmet learning needs and common practices of migrant language learners in a mobile context. It then presents the ways in which the report proposes to address these needs through the embracing of Machine Translation and crowd-sourced lexical resources. The chapter concludes by presenting the mobile application implemented which combines these solutions in an effort to address the language learning needs and facilitate the language learning practices of migrants.

2.1 Motivation: The differences of language learning by necessity and by choice

With mobile phones being ever present in our lives, the opportunity to use these technologies to learn and hone new skills has created the expanding field of Mobile Learning. The Mobile Learning market is massive and expected to grow in the coming years with increased smartphone ownership throughout developing economies (gminsights.com 2022). Within this field, mobile assisted language learning (MALL) is an area which has taken hold. A remarkable number of people are using their phones to learn languages. In 2021 Duolingo, the most popular language learning app, had 40 million active users (businessofapps 2022).

In an ever globalized world, language learning is an activity which is practiced worldwide and increasingly necessary for people to bridge language barriers in their professional and social lives. A 2018 report on the future for English in the EU by the British Council states that, with increased movements of people for employment and education, there is a rising demand for language learning which is “flexible, personalized, purpose specific and
In general the increased mobility of people across the globe means that more and more people need to learn and use a foreign language (L2) for professional or personal reasons (British-Council 2018).

For both voluntary and involuntary migrants, whether it be for education, work or due to displacement, it is imperative to become integrated into the host society by learning a new language and culture (Kukulska-HulmeAgnes 2019). They must learn language in order to overcome communication challenges in their daily lives. As migrants are often having to deal with a foreign language while living in a host country, such challenges include receptive language skills such as reading and aural comprehension and it has been shown that they are also concerned with correct pronunciation, having sufficient vocabulary knowledge, and the ability to understand and the use of idiomatic phrases in a way which is correct to certain socio-cultural contexts. In this way the motivations and aims of language learning for migrants may be markedly different to that of language learners who partake in language learning as part of formal schooling or for pleasure. These language learners may not be partaking in a form of formal language learning such as schooling (Kukulska-HulmeAgnes 2019), and their goals and motivations can be markedly different to that of a student whose motivation is often informed by a curriculum based approach.

“Millions of migrants, both those who have chosen to move and those who are displaced from their country of origin, find themselves having to learn the languages of their host societies within a short space of time, for work, everyday life, education or training” (Kukulska-Hulme 2019).

This statement underpins the primary differences between student language learners or people who practice language learning as a hobby and those who are learning foreign languages to socially and/or professionally integrate into a society.

Migrants require language learning opportunities which are:

- **Effective and engaging:** Because integrating into their new host society is important to their social and professional lives, they need to learn languages in a timely manner. Additionally their motivation to learn the L2 is higher. This creates a need for language learning where progress can be made quickly and where one can invest a lot of time within these learning environments without getting frustrated or tired. This is where the ubiquitous nature of mobile devices in our daily lives can offer the opportunity to sustain ”educationally valuable activities across everyday environments that are not designed or arranged for learning” (Kukulska-HulmeAgnes 2019).
• **Purposeful and Specific to their current or future situation:** they may need to learn a language to integrate socially within a community, for professional reasons or for educational reasons. Additionally, language learning may occur “on their way to a destination, once they have arrived, and sometimes for many years after that.” These various motivations create the need for language learning opportunities which are flexible to the needs of the learner, allowing for language use in a certain context and at different levels of proficiency.

• **Idiomatic:** “the extent to which a learner’s language resembles that of a native speaker” Because the purpose of language learning in this case is to socially or professionally integrate within a host country, there is a need for language learning which provides the opportunity to reach a high level of idiomaticity (Kukulska-HulmeAgnes, 2019; Epp, 2017). This creates a need for language learning which enables a high degree of proficiency, and fluency as well as an ability to use idiomatic phrases and expressions in the right cultural context (?). This need to learn with the goal of language use also means that migrants have a focus on pronunciation and conversation.

### 2.2 The use of linguistic assistant tools to overcome communication challenges, but also to provide language learning opportunities

The practicality of Mobile assisted language learning in terms of flexibility regarding time and space (Lai and Zheng, 2018) means that there has been a focus in recent years on the potential benefits of mobile learning for migrants. Because of their mobility and ubiquity, these tools can provide ways to address and overcome “real world communication challenges” outside of educational contexts. Recent literature on mobile devices have stress the fact that they provide the opportunity for learning to take place throughout the day, and integrate themselves within different activities, as such they provide a wide array of learning opportunities which are relevant to migrant learners may be in constant contact with a foreign language.

Because migrants may not be able to have access to formal language learning settings such schooling or evening classes, language learning may be entirely self initiated and self driven, and is often informal in nature as they are constantly having to interact with the foreign language (Epp, 2017). Because of this translation is popular with language learn-
ers as it can support them in unpredictable contexts and provides them with the tools to expand their exposure to the target language in ways which are engaging and specific to their needs.

“In the dawning era of intelligent assistants on smartphones and other smart devices, it is reasonable to assume that various forms of mobile and smart assistance will be incorporated into formal language learning practices.” \textit{Kukulska-HulmeAgnes} (2019)

It has been shown that language learners use these intelligent assistant tools such as online translation services, dictionaries, online thesaurus in their self-initiated learning. For instance \textit{Bradley} (2015) investigated how students use mobile technology out of school to learn English and found that they use online translation tools and dictionaries when reading and using the internet. A investigation into the language learning practices of migrants concluded similar findings, reporting the use of dictionaries to support activities which included reading and watching television \textit{(Epp} 2017\textit{)}.

Yet a review of the literature shows that there has been very little exploration of how these “intelligent assistant tools” such as dictionaries and online translation tools can be incorporated into more formal language learning practices such as the use of mobile learning applications \textit{Epp} (2017) \textit{Kukulska-HulmeAgnes} (2019) \textit{Nino} (2020). Discussions of translation in the context of informal language learning is extremely limited and a review of topics such as the usefulness of inline translations for language learning brings up a very limited number of relevant results. \textit{Epp} (2017) mentions two examples which provide targeted support for text comprehension both of which are outside the context of foreign language learning. \textit{Veras et al.} (2014) explores a tool which allows learners to substitute synonyms within text which they have trouble understanding and \textit{Munteanu et al.} (2013) which explores text to speech to develop literacy. Much of the literature on machine translation focuses on its impact on written production exercises and its potential disruptions to classroom environments \textit{(Groves and Mundt} 2015\textit{Nino} 2020\textit{Ducar and Schocket} 2018\textit{)}. One study, \textit{Hsu et al.} (2013) presents an application which uses collaborative annotation of documents by peers on manually personalised reading material to provide reading comprehension support, but this method relies on manual annotation of material from classmates rather than automatic tools such as machine translation. As such we would argue that their is a need for investigation in how linguistic assistance tools such as dictionaries and machine translation can be used in the context of mobile assisted language learning. Given the highlighted affordances of mobile learning for the language learning needs of migrants, such as the fact that their ubiquitous nature allows
for the potential of "educationally valuable activities across everyday environments that are not designed or arranged for learning" we choose to focus our analysis of integrating these tools to address the unmet needs and facilitate the personal practices of migrants.

2.3 Integrating linguistic assistant tools into a language learning system to meet the needs of migrants.

Given their popularity amongst migrants of any kind to overcome communication challenges and for the purpose of informal language learning, this paper will explore how linguistic assistance tools such as translations and lexicons can be integrated into a language learning environment in the form of a mobile learning application. A language learning application was created with the goal of addressing the language learning needs and facilitating the personal practices of migrants. We outline a number of these identified needs and practices and discuss how the implemented application takes these into consideration.

The needs of migrants have been explored within the literature focusing on development tools for migrants to learn languages and to socially integrate Kukulska-Hulme Agnes (2019). Research has shown that they turn to mobile applications to supplement their language needs as formal language institutions often do not fulfill their needs and may be incompatible with their life situation. Kukulska-Hulme Agnes (2019); Epp (2017) this literature highlights their is a lack of focus on migrants as “mobile language learning tools has predominantly focused on foreign language learners at the post-secondary level Kukulska-Hulme Agnes (2019); Epp (2017); Burston (2014) rather than those who must use a language to survive” creating a need to investigate migrant’s insights and study habits.

To try and address this inadequate understanding of how migrant’s use mobile tools Epp (2017) conducted interviews with 18 migrants. The following list presents some their findings regarding migrant’s language learning practices and how the app addresses these. A more in depth discussion of this can be found in section 3.6
2.3.1 Identified needs and practices and how the implemented application attempts to address these.

Their informal learning activities relied on strategies that would develop their receptive language abilities. Migrants thought to improve their reading and used materials of an appropriate level they thought out “children’s books, Internet articles, and community focused newspapers.”

- This paper will present how the use of article feeds, defined by key-word topics and website URLs allows language learners to define and discover the digital spaces, articles and videos which they use in their informal language activities such that it is engaging, separated by topic and suitable to their level.

They supported this activity using dictionaries and stating their frustration with statements such as “It’s all hard”, “There’s so much to learn” and “It’s just too hard to comprehend it all.”

- This paper will explore how the use of machine translation, crowd sourced lexical tools such as dictionaries and “sentence dictionaries” like Tatoeba can be used in a mobile application to facilitate the reading comprehension of authentic texts such as “internet articles and community focused newspapers at any level”. In this respect the value of this system relative to existing linguistic resources such as google translate will be in the ease of use and total integration of translation capabilities, lexical resources within a mobile context as well as enriching the the translation output with “part of speech highlighting” allowing for an easier to read output such that language learners can start reading comprehension at lower levels of proficiency and with a wider variety of text. This output additionally provides the ability to easily evaluate how L2 phrases are translated to L1.

Migrants focused almost exclusively on learning vocabulary.

- This paper will additionally present ways of facilitating vocabulary acquisition both incidentally and intentionally. Incidental vocabulary acquisition, meaning ‘picking up’ words from language use, intentionally or not is facilitated in the following ways:

The first is related to providing lexical (dictionaries) and translation tools within a digital reading context such that a language learner can discover and learn more vocabulary incidentally. These tools allow the learner to learn words and phrases from context while reading material which they can better follow and by enabling the user to look up these words and phrases using dictionaries and translators.
Additionally the app performs on screen highlighting of words which were previously unknown by the learner which will facilitate focusing on learning how these words behave in context, by bringing attention to them during skim reading. This facilitates learning words in the learn list from context while exploring digital resources.

Within the digital tools which migrants use in their informal language practice there is an unmet need “Technology enabled rehearsal” of content which they come across while engaging with authentic language use.

- The second way in which this application facilitates vocabulary acquisition is by providing a way for language learners to pick out words which they want to practice later while reading articles. From henceforth we will refer to this as the ‘learn-list’. This ‘learn-list’ is then used to automatically create intentional vocabulary building exercises based only on those words which the user has selected throughout their reading comprehension practice. These words leverage a crowd sourced “sentence dictionary” Tatoeba to integrate the use of authentic context within these exercises. This has an advantage in that it supplements incidental language learning and restricts exercises to content which the user is trying to learn.

Tatoeba is “a free collaborative online database consisting of sentences geared towards foreign language learners.” which contains 10,000,000 sentences in 409 languages. Native speakers contribute translations of these sentences as well audio recordings of the sentence. We use this database to automate the creation of language learning exercises such as a fill in the blanks exercise or a flash card exercise. By harnessing Tatoeba the application is able to automatically generate vocabulary learning exercises which can present words in different forms according to certain contexts.

Migrants focused on ways to practice pronunciation: “parroting audio materials, whether they originated from audio or video, song, language-learning-tape or text to speech engine of a smartphone were all perceived beneficial.” Epp (2017)

- This paper will present how crowd-sourced sentences and audio recordings of native speakers speaking those sentences can be integrated into context sensitive language learning exercises in the form of flash cards and fill in the blanks. These exercises thus additionally facilitate pronunciation practice using the context in which a word is presented within these exercises. The presented prototype relies on crowd-sourced authentic utterances which can be used at a word level through Wiktionary.
while reading an article or at a sentence level using the Tatoeba database while reviewing examples during the flashcard or fill in the blanks exercises. We note that implementing automatic text to speech for any arbitrary piece of text would be additionally useful but was not implemented.

Migrants are concerned with how words and idiomatic phrases are used in specific socio-cultural contexts.

- We will explore how the use of translation and crowd-sourced technologies can be combined to allow the learner to learn authentic language use at the phrase level in a context sensitive way. This is done both within informal reading practice and intentional acquisition exercises of these multi word linguistic units.

One point of focus of this paper is the importance of language learning at the phrase level, fluent sounding language is predicated on the interactions between words and a knowledge of idiomatic phrases and conventionalised ways of expressing meaning in certain contexts is vital [Contreras Kallens and Christiansen (2022)].

2.4 Overview of application and discussion of implementation

As well as exploring a number of potential individual strategies to address the identified needs of migrants as outlined above the presented language learning mobile application presented in this paper combines these strategies to form a language learning “pipeline”.

- The application provides ways for users to define and explore reading material based on their interests and learning goals. This material is then separated and accessible depending on the language learning goals of the user at that time.

- While reading these articles, the application allows the user to engage with that material, facilitating reading comprehension when needed using dictionaries and translators, allowing the learner to listen to fluent speakers utter words encountered. Within this environment the user notices language use he does not know and specifies words and phrases which he wants to learn in the future. These words are highlighted in the article to bring them to the users attention while reading and exploring content.

- The user can than go and practice these words and phrases in a variety of authentic examples with flash card and fill in the blanks exercises. Within these exercises
the user can listen to utterances of the sentence where the words occur and provide feedback as to whether or not they know such a word. This will adjust the frequency with which the words appear in flash card exercises.

Below is a description of the different functionalities of the application and how they might used as part of an informal language learning practice.

We additionally discuss some of the implementation details which are relevant to analysing the capabilities of the systems as it is currently implemented. Through the discussion of some these implementation details, we hope to explore considerations to be made when developing such an application, namely, one which relies on Machine Translation and crowd-sourced lexical resources to facilitate language learning on the Web.
Overview of application use

Subscription feed definition and content discovery

Choose language to learn (L1) and to translate into (L2)

Subscribe feeds to by website, topic or keyword

Unsubscribe to feeds

Browse articles by feed or by the aggregate of all feeds

Can subscribe / unsubscribe to websites by long press on article

Can browse website within app

Read articles

Can be any language supported by Wiktionary, Tatoeba and Machine Translation services.

All other data: subscriptions, words, phrases etc. is specific to the language being learned

Choose language to learn (L1) and to translate into (L2)

Enable disable “Part of Speech highlighting”

Highlight word or phrase to view translation

Yes

Word entry in Wiktionary?

View dictionary definition and example usages

Listen to native speaker utter word

Add word or phrase to learning-list

Practice words in different contexts with flashcard exercises and fill in the blanks

Performance of words is tracked and less performant words appear less often

Word list can be reviewed and items removed from word-list

Intentional practice of vocabulary and pronunciation through context sensitive exercises

Audio clips of native speakers uttering example sentences can be listed to
Figure 2.1: Architecture of the application deployment
2.4.1 Choosing L1 and L2 and the limitations of the open-source machine translation models.

The user can specify any language which is supported by the system as their source language, the language they already know such as their mother tongue, generally referred to as L1, and their target language, the language which they are trying to learn, generally referred to as L2. Defining L1 specifies the learn-list and subscriptions which the user has access to under that setting as can be seen in Figure 2.2. In contrast defining L2 does not change the "learn-list" or subscriptions.

These languages can be virtually any language which is supported machine translation system, lexicons as well as the online database Tatoeba. In the current implementation, because Tatoeba and Wiktionary rely on crowd sourced information, they both have vary-
ing degrees of coverage for different languages (see: Tatoeba: tat (2022) and Wiktionary: wik (2022). This forms a bottleneck for less used languages as the system will be less reliable in providing examples and user submitted sentence and word utterances.

The machine translation models used currently are open source, these are increasingly accurate and are only slightly less accurate than commercial systems such as google translate and DeepL MTS (2022). They have the advantage of being free and they can be self-hosted to fit the application purpose. Commercial translation systems are quite costly which may deter their use in technology development such as creating language learning applications.

We use the OPUS-MT Helsinki-Nlp (2022) model which allows for 1200 language translation directions, however each of these directions must be hosted individually to support that translation direction with sufficiently low latency, hence the few choices of L2 languages in 2.3. This is because these models are quite large and it takes some time for the new translation direction to "heat up" in memory. Using containerisation such as Docker and deployment services such as google cloud, flexible deployment of these models based on the use case easily implemented with limited overhead.

Additionally these models can be relatively slow depending on the hosting method and as such, application design must take into consideration the speed of these models, and the overhead of sending individual requests. Within our application implementation there were instances where the flow of data had to be redefined in order to send translation tasks in batches to improve responsiveness of the application.
2.4.2 Defining subscriptions by RSS feed

![Image: Add a subscription]

Add a subscription
- Add by Space
- Add by Keyword
- lemonde.fr
- Add by Website

![Image: Articles]

These are the articles found:
- Un mensaje que busca un respuesta desde el espacio exterior - La Tercera
- Así se vio desde el espacio el incendio en Home Depot de San José - Univision
- Las imágenes más espectaculares de Canarias desde el espacio - Diario de Avisos
- Gobierno ruso aprueba acuerdo con Venezuela para la exploración del espacio - SVI swissinfo.ch en español

![Image: Articles]

- Guerra en Ucrania: Rusia advierte a Suecia y Finlandia contra su adhesión a la OTAN
- BBC News Mundo
- Revocación de mandato: AMLO vence en el primer referendo revocatorio de México para que continúe como presidente con una participación del 17%
- BBC News Mundo
- Cuba y Ucrania | Investigación BBC: los misteriosos grupos de Facebook que apoyan a Putin
- BBC News Mundo
- "A Chávez le dieron la oportunidad de oro para radicalizarse": 20 años del fallo golpe de Estado en Venezuela
- BBC News Mundo
- Quién es Aleksandr Dvornikov, el "camarero de Siria" que según EE.UU. es el nuevo comandante ruso a cargo de la guerra en Ucrania
- BBC News Mundo
- La guerra en Ucrania: el 45º día de la invasión rusa, en imágenes
- EL PAÍS
- Rusia declara alerta terrorista en Chine y tres

Figure 2.5: Subscriptions can be specified by topic (Interest), key-word or website. Figure 2.6: Subscribing to a feed allows users to access articles. Figure 2.7: The user can select a subscription and browse that feed, or as seen here browse by all subscriptions.

The user can specify topics, websites and key words which will create article feeds. These article feeds allow the user to discover and explore content relating to topics and registers which fits his objectives and interests. As we can see in Figure 2.5, the user is subscribing to the news website "lemonde.fr". As the subscription was successful i.e the associated RSS feed contained articles, the contents of that feed are previewed as can be seen in Figure 2.6. The user can then browse a feed as in Figure 2.7.

While reading articles relating to a topic or key word, for instance football a user can subscribe to the feed of a website he has come across in that subscription by long pressing on that article. The user can view article 'feeds' relating to a single subscription as well as from the aggregate of the subscriptions. In this way, certain topics and registers are easily accessible depending on the purpose or motivation for reading.
The article feeds rely on the Google News api. Although not implemented, it has the capacity to provide local headlines based on location, which was a form of reading which was sought out by participants in Epp (2017). One thing to note is that Google News is not available in certain countries such as Spain and Russia. These feeds provide access to a maximum of 100 articles at any one time, relating to the website, topic or key word queried, as such in its current state the app does not save any articles to the user’s account, but simply provides the current ‘state’ of the RSS feed at the time of query. This implementation limits potential improvements to the app such as labeling articles with some estimated level of difficulty or by identified topics.

2.4.3 Defining single words using translators and dictionaries

![Figure 2.8: Select words to translate them. If the selected word is present on Wiktionary, a definition is available to user, this indicated by a green outline of the ‘translation box’.](image)

![Figure 2.9: Clicking on the green underlined box bring up the Wiktionary dictionary definition.](image)

![Figure 2.10: Some Wiktionary entries provide a sound clip of a native speaker uttering the word.](image)
While reading the content single words can be highlighted and a translation in the the user’s mother tongue is shown. This translation is not given any context and as such the translation chosen is relatively inaccurate. As such the user will have to try and evaluate whether the provided translation is correct. The dictionary definitions of a word is available to the user from Wiktionary if the highlighted word is an entry in the Wiktionary database, when available the translation output box is outlined in green such as in 2.8. If the user evaluates that the presented translation does not match the context of the word, the user can subsequently choose to look at a full definition of the word form as provided by lexicon.

In the Wiktionary website a word form combined with a Part Of Speech (POS) forms a dictionary entry. As can be seen in 2.9 Animo: verbal noun and animo: interjection form two different entries.

Each entry can contain different possible senses of that word form and POS and example uses for those different senses. The application additionally fetches and presents the infinitive form of the word if it exists, as can be seen in 2.10.

From the dictionary definition the user can listen to the utterance of the each dictionary entry, from the original or infinitive form of the word. This utterance is provided by a native speaker. These utterances are not necessarily present for word entries, less common words will generally not have a utterance available and less well supported languages are less likely to have word utterances for entries.

2.4.4 Constructing dictionary definitions from the Wiktionary data dump.

Because Wiktionary does not have an API, this definition was constructed from data dumps [fil (2022)] of the Wiktionary website. These data dumps were separated by language and were extracted to JSON using [attardi (2022)]. Each language supported had one corresponding large JSON file with every word entry in the Wiktionary database for that language. These were used to construct individually running NOSQL databases, where the primary key to access each definition was the word form. This was necessary in order to limit lookup time of word entries for tasks such as fetching the infinitive form of a word.
2.4.5 Translating multi-word units to inspect language chunks

Additionally multi word language chunks can be highlighted, a translation of the highlighted text is shown. This allows the translation of multi word units such as phrases like “to a degree” or “carried out” into the language learner’s mother tongue.

In the case of the user not being sure of what sense to apply from it’s definition, it allows sense disambiguation of a words definition by way of providing the translation system with more context. This is an advantage over word by word translations or simply looking at dictionary definition of words individually.
It also allows whole passages to be highlighted such that the learner can see how phrase construction translates to L1 as can be seen in 2.13. This allows beginner language learner to understand the text such that they are able to engage with reading comprehension at a low level of proficiency.

2.4.6 The potential of ”Part Of Speech highlighting” natural text

When viewing the translation of a word or phrase the words are optionally annotated with their “part of speech”, def. “a category to which a word is assigned in accordance with its syntactic functions. In English the main parts-of-speech are noun, pronoun, adjective, determiner, verb, adverb, preposition, conjunction, and interjection.”

It may give the opportunity for the user to see how a word is behaving semantically and syntactically in that sentence. This provides the opportunity to gain knowledge about the grammatical and semantic accordance of words at phrase and sentence level for L2.

As can be seen in 2.13 this also allows for large translation outputs to be much more readable, in this way the sentence is broken down into visually distinct section, in the case of Spanish to English, clear phrase mappings allow for an understanding of how multi word phrases translate to the target language.

2.4.7 Optimizing ”Part Of Speech tagging” text on the web.

For single words or phrases such as ”a traves de” in 2.11, the POS annotation of the source language output can remain accurate as the application pulls and ”POS tags” the context in which that word was found in the web-page. This allows the system to provide the POS tagger with additional context to tag the highlighted word. After annotating the POS of the context, the selected POS is the one corresponding to the instance of the word within ”POS tagged” context.

It additionally tries to do the same thing for the target language i.e translation out-
put such as "through" in [2.11] by translating the context in which the highlighted word was found. If the word forms which make up the translation of the highlighted text are present in the translation of the context, it can thus annotate the POS with a higher degree of accuracy using the same method as described above. Otherwise it reverts to POS tagging the translation output by itself.

2.4.8 Adding words and phrases to the "learn list" and highlighting instances of them on the web

Figure 2.14: Words and phrases added to the 'learn list’ provide an ability to see bring attention to words the user is trying to learn. Words which have a dictionary definition are highlighted in yellow, otherwise they are highlighted in green.

Figure 2.15: Words which have been added to the word list can be viewed, along with their "score" a tally of how the user did on flash card exercises for each entry.

Figure 2.16: Phrases can be viewed separately, from this screen users can delete words which they are satisfied they know.
A selected word or phrase can be added to the user’s "learn-list". A word is categorised as such when it has an entry in Wiktionary, else it is treated as a "phrase". This means that proper nouns and acronyms will often be treated as a phrase.

The user can view the "learn-list" by word, phrase or both. From this list, the user can remove words or phrases they no longer want to practice, or see highlighted. They can additionally inspect the definition of an entry in the list by pressing on it as seen in 2.16. From this list they can also see how well they have done in flashcard exercises for that word or phrase. The score is a tally of how well they have reported to know the word when it came up in the flashcard exercise, and this score relative to other scores impacts how frequently each word will come up in the flash card exercises, the higher the score the less frequently a word will appear in the list.

Instances of words and phrases in the "learn-list" found in subsequent texts will be underlined allowing the user to pay attention to this word instance in its context.

2.4.9 A need for mobile extensions to facilitate augmenting the web on mobile devices

Word and phrase underlining is performed on page load, and as such this feature is applied whenever the user is browsing any website through the app.

This application’s features are applicable to any browsing activity on phones, not simply articles found within the app. For example the user is able to follow any link within an article without any change in the capabilities of the system, a Google button could very well be implemented which simply opens Google within the application and browses the internet.

This resembles desktop browser extensions. Desktop extension provide functionalities which are accessible on any web page. Publishing browser extensions for mobile browser is not currently possible to everyone as on Desktop browsers, and Firefox only supports a few "white-listed" browser extension for android, while chrome supports none.

To get around this the application interacts with a browser window or "web view" is by using the react-native-web-view react-native webview (2022) package.
However as of this writing this, the support for communication between the application and the "web-view" is minimal and as such implementing complex asynchronous behaviours which are dependent on the state of the application as well as the state web-view is more difficult. Additionally their is no support such as error messages and debuggers present when developing web-view functionality as they are not designed to facilitate additional behaviour while accessing websites from mobile devices. On the other hand desktop browser extension benefits from debuggers and more sophisticated APIs to aid development. For instance we were not able to integrate jQuery in the web-view to manipulate text, and had to rely on lower level JavaScript methods such as ran (2022).

Developing mobile language learning environments which situate mobile learning on the web would therefore benefit from wider support for mobile browser extensions.

2.4.10 The complexity of communicating with a web-view

To illustrate the challenges of adding features to a mobile web-view within a mobile application which may limit the complexity with which mobile devices are able to augment language learning situated on the web, we describe how the features which interact with the web-view were implemented below:

The react-native webview (2022) package supports 'on-load' java script injections and message passing.

Message passing is a way for the application to communicate with the browser driver inside the web-view. It consists of "on-message" callback functions in the application and in the web-view. These serve as message handlers of String messages which are sent using "post-message" methods.

Because only strings can be passed between the web-view and the application, state variables such as the array of words to underline are placed into a message containers, which contain the message type and the payload. These messages are then serialised into Strings by the application and de-serialised by the web-view browser or vice versa.

Depending on the message type the payload is processed in a different way, for instance if
the message contains the array words to underline, the message type will be "underline-words", and a handler function will underline found words in yellow, the "underline-phrases" function on the other hand will underline phrases in purple.

To trigger message passing during run-time, outside of when the website is being loaded for word and phrase highlighting, the application relies on React Native Hooks to trigger message passing on state changes, and the web-view uses JavaScript event callbacks such as "on-highlight", these are initiated with the initial "on-load" JavaScript injection.

Complexity arise with challenges such as needing to limit the number of on-highlight callback messages which are sent and processed by the applications back end, or when a text needs to be un-highlighted and re-highlighted within the web-view depending on the applications state, as both the web-view and the application need to keep track of a common state.

To deal with this, all state management responsibilities are given to the application, and React Native's state management features such as "useState", "useEffect" and "useReducer" hooks facilitate the development of properly functioning behaviour.

2.4.11 An alternative approach: Scraping articles into simple text representations,

The original approach to providing linguistic assistant tools within articles was attempts by trying to 'scrape' content from articles using Mozillas' readability API [mozilla](2022). This would have facilitated feature implementations relating to the reading of articles in the application, such as underlining words and highlighting text for translations. This is for two reasons, firstly the HTML could have been rendered natively in the app without the use of a web-view, and additionally, the structure of the HTML would have been scraped and simplified to it’s core components making tasks such as finding word instances and pulling the context of words from a web page filled with arbitrary HTML structures much easier. This approach was not used as the parser struggled to provide an accurate output for certain websites, and would omit paragraphs of content by error. Another disadvantage of this approach is that website lose their visual design, and articles often contain video and various more complicated HTML elements which the readability API does not scrape.
2.4.12 Challenges of getting the context of a word in complex HTML web pages with limited support.

To illustrate the complexity of interacting with text on a web-page we will describe how the context of a word is taken from a web page in order to provide additional context to the POS parser and to provide a reference meaning for phrases by translating that context as is seen in \(2.18\) below, where the lifted context of "que incluye" is "En todo el mundo, los consumidores y las empresas se enfrentan a una escasez de productos, que incluye desde el café hasta el carbón".

We were not successful in adding jQuery functionality to the web-view on page-load and as such we had to rely on using `ran (2022)` to find the context of a word.

To find the text surrounding a highlighted word such as in a paragraph, the text of the parent element in which that word resides was tested to see if it was greater than the desired length of 40 words. If it the content of that element was not 40 words, the method would re-curse onto the highlighted word’s parents’ parents to check if that element contained text of the required length and so on... until an element in the hierarchy was reached which contained the highlighted word, and its found context was of a sufficient length. This context was then passed to the application and the instance of the word and it’s 40 surrounding words were taken to be the context of the word. This approach works for the large majority of the time but there are some cases where the word will repeat within a large context, and the application does not have the information to decide which instance of the word in this large context was highlighted. One solution to this which was not implemented is to use the context of parent elements in the hierarchy to 'drill' down to target word, by using smaller and smaller contexts to find the target word. This method was not implemented in the current prototype.
2.4.13 Practicing words and phrases using Flashcard exercises which show context

Words and phrases in the learn-list can be practiced intentionally using flash card exercises.

To provide create the stack of flashcards at the start of each round, sets of entries in the learn list are selected at random. This selection is performed such that entries with a lower ”score”, which would indicate that the user ’knows’ these entries less appear more frequently in the stack.
The flash card exercise consists of the user reading the word in multiple contexts fetched from the Tatoeba database, before looking at a representation of that words meaning, either from a dictionary definition in the case of words as is seen in 2.16 above or by the translation of the context in which the word was found in the case of phrases as is seen in 2.17.

### 2.4.14 Fetching sentences with specific phrases

The Tatoeba database has the capability of fetching sentences from a chosen word in different inflexions as can be seen in 2.17 where "que incluyen“ is an inflexion of "que incluye“. As such the examples provided can show the word in different contexts and inflexions.

It is important to note that Tatoeba has no understanding of "phrases“ per se, for example to fetch example sentences containing "que incluye“ in different forms the application must specify to query the database for a sentence which simply contains both "que" and "incluye“. This is necessary as phrases and collocations do not necessarily have constituents which appear beside each other, such as in "que no incluyen". Therefore there is not guarantee that the sentences retrieved will contain the words "que incluye“ beside each other to mean "which includes“. For instance an example sentence fetched might be "que haces, no te inclu‘i".

Thus the user is shown sentences where a phrase is present, and where the constituent words of a phrase are present but not combined in a way to form a phrase. An example in English could be the following:

A user highlights the phrase "To a degree" in an article. While practicing the phrase in a flash card exercise he will can be shown the following two examples:

"I agree with you to a degree”

"I couldn’t go to the shop to buy a newspaper as it was 30 degrees outside.”

As it stands the implementation has no way of controlling the ratio of ‘idiomatic’ and ‘non-idiomatic’ examples a phrase, difficulties implementing something like this are described in Chapter 4.
Multi word units which have lexical entries in Wiktionary do not suffer from the same issue as they can be treated as "semi fixed expressions" such as "a través de" in 2.19; these are fetched from tatoeba as fixed single words-with-spaces, but can still be retrieved with different forms of inflexion.

2.4.15 Fill in the blank exercises, improving on existing applications by providing flexibility and the ability to practice aural comprehension.

Figure 2.20: In the case of Figure 2.21: The flash Figure 2.22: To help the user words, the dictionary defini- card exercise presents sen- can view the translation if tion is presented such that tenes containing the word he chooses to, or listen to a the user can evaluate how or phrase, removing one native speaker utter the sen- well he has knows the word, word from the sentence and tence. the utterance can be listened presents four alternatives for to. the user to choose.

The user can practice seeing the word in example contexts using a the "fill in the blanks" exercise in the application. These sentences will contain one of the entries in the "learner
list” and omit one Noun, Verb, or Adjective or Adverb from the sentence at random. The game provides for options, each of which is one of similarly, a Noun, Verb, Adjective or Adverb.

The implementation of this game is heavily inspired by the Clozemaster website which additionally provides an endless stream of example sentences from Tatoeba. In fact this website is was the initial inspiration of trying to 'improve' personalised language learning exercises by restricting the material they test to the ”learn-list”.

On top of the features which https://www.clozemaster.com/ boasts the exercises is constrained to the learn-list, and allows for the sentence to be listened to, and for the translation to be viewed. This allows fill in the blank exercises to be flexible to the proficiency and goals of the learner. For instance if the learner is struggling he can choose to translate the sentence to try and guess which word is missing from the translation as seen in ??.

He can also try and listen for the word from a fluent speaker utter the words.

The availability of sentences which have a corresponding audio recording of a fluent speakers uttering the sentence is limited on Tatoeba, as such for a more robust implementation of this game, which can provide sentences for uncommon words and phrases would need to rely on automatic text to speech, this illustrates the limited scope of Tatoeba in terms of coverage of uncommon words and phrases.

This flexibility provided by Machine Translation in allowing language learners to choose the level of difficulty of exercises on a case by case basis could be reflected in the feedback weight of each answer, for example a correct answer where the translation was viewed will indicate less that the learner 'knows' the word being tested.
Chapter 3

Background Reading

3.1 Mobile language learning: an opportunity for content driven and informal learning

In this section we argue for the use informal and personalised language learning on mobile devices facilitated by the integration of digital tools such as Machine Translation.

We analyse informal language learning and the importance of personalised language learning, we argue that as mobile learning is inherently outside the classroom and self driven, there is a need to provide personalised learning which motivates users to learn in a flexible and positive way. We present evidence that gamification used by popular apps takes away from learning. We propose that integrating language learning into daily digital media consumption through the use of translation and other linguistic resources is a possible solution to providing engaging and personalised language learning. Lastly we provide recent technological developments which have made this approach possible.

There are a wide array of goals and motivations to learn languages. Some language learners practice languages as part of schooling, for pleasure or out of necessity such as for work, education or to integrate socially into a new host society. As mobile language learning is inherently self-directed [Nino (2020)] [Lai and Zheng (2018)] due to the fact it occurs outside of a classroom context in a way which is self driven, the practice of mobile language learning, be it in a formal or informal way needs to be aware of the goals and motivations which guide language learners at different levels of proficiency and from different backgrounds [Lai and Zheng (2017)] [Nino (2020)].

Self-driven language learning therefore benefits from engaging in personalized content...
found through 'content driven' Nino (2020) and 'informal' language learning.

"Mobile learning lends itself to activities involving sustained language practice over time and learning beyond classroom walls, including exploration of authentic language problems and challenges in everyday life. Learning experiences should allow learners to express and follow their interests, to work towards personal goals, and to engage in content creation and sharing, while raising awareness of strategies in listening and speaking.” Kukulska-Hulme (2016)

Informal language learning is associated with learning languages while engaging with the theses languages in the day to day, it "can happen "spontaneously, informally and even involuntarily as one simply engages with authentic language use while seeking entertainment, working or socializing Sockett (2014). In this way informal learning is highly contextual, and allows the user to direct his own learning by drawing on existing resources and by using a foreign language throughout their daily life.

According to Informal Language Learning theory, language learning occurs outside formal classroom settings, ”unconsciously and incidentally through interaction with the native speakers or exposure to authentic language input through technology”. Sim (2012). The term was defined by Coombs and Ahmed (1974) defined informal learning as ”the lifelong process by which every individual acquires and gathers knowledge, skills, attitudes and insights from exposure to the environment at home or at work through reading newspapers and books or by listening to the radio or viewing films or television.” Informal education is unorganized, unsystematic and even unintentional at times. Coombs and Ahmed (1974), Godwin-Jones (2015) points out that especially for learners of English, language learning is often a side effect of seeking entertainment online and socializing on social media. In this way we can see the potential for informal language learning within a mobile learning application. The massive amounts of multimedia content freely accessible on the web can be used as a way to provide highly engaging, highly personalized and specific language learning opportunities.

Informal and content driven language learning is less in popular language learning apps which generally rely on exercise heavy approaches García Botero et al. (2019) Even analyzed within a “curriculum-driven” Nino (2020) context it has been observed that while students maintain positive attitudes towards the most popular language learning app Duolingo, which takes a formal and exercise heavy language learning approach , it fails to maintain long engagement for these students. “Duolingo can encourage out-of-
class learning through fun activities, but interviews reveal a lack of sustained motivation, self-monitoring and self-management reflected in the low usage of the application.” García Botero et al. (2019).

More and more language learning apps are relying on gamification (e.g points, leaderboards and badges) to motivate language learning, but this has been criticised as a misuse of gamification. A fixation on gamified aspects of these apps ”take away from language learning” and ”wastes users time” Mogavi et al. (2022). As well as this, competitive aspects of gamified learning and a focus on strict routine practice generate high drop out rates and less interest in learning due to disappointment Mogavi et al. (2022). This calls for a different way of motivating users to practice self driven language learning. We argue that incorporating language learning with daily activities relating to digital devices such as reading the news and seeking entertainment is a focus which mobile applications should focus on given that these are approaches taken by migrant language learners and students, who engage in these informal language learning practices with the help of online linguistic resources such as dictionaries and translation tools Godwin-Jones (2018).

As such an investigation of how translation can facilitate self-directed and informal language learning that relies on existing resources can provide insight in how these tools can facilitate learning which is highly flexible and personalized. The impact of such an investigation is greater then ever for the following reason:

- Web 2.0 technologies provide access to quasi endless amounts of largely free multimedia content in a variety of languages. Providing language opportunities wherever online digital resources are being used, such as while seeking entertainment, socialising, getting informed, taking part in education and working.

- The ubiquity and low cost of smart phones means that technology based educational resources can impact communities that may not have had access to more expensive desktop based resources in the past Elt (2017).

- The availability of crowd sourced digital resources provide authentic utterances of words and sentences in a large range of languages. Open source translation tools such as open-nmt provide free access to translation performance only slightly below state of the art MTS (2022) allowing for free access to accurate translations for a wide variety of domains and languages.

- These Machine Translation models have experienced a recent large improvement in accuracy due to the adoption of Neural Machine Translation in 2016. This possi-
bilities created by this improvement has yet to be explored for informal language learning and the majority of the literature which have considered it in a pedagogical context have analyse it’s use within the context of a classroom Lee (2021).

3.2 The advantages of the mobile device for informal language learning

In this section we describe the advantages of mobile devices for informal language learning. Sung et al. (2015) identifies three main features of mobile phones which have an applicability to informal language learning.

- Mobility and portability. The fact that mobile devices are small and lightweight means that learning can happen anywhere and at any time, thus the mobility of mobile devices facilitate spontaneous and unplanned language learning in a way that other digital learning environments fail to do. This is perhaps the most defining characteristic of the mobile learning environment, as it inherently facilitates unstructured and unplanned language learning which diverges from formal language learning outside of a classroom context making mobile learning tools particularly relevant to learners who may not be participating in formal language learning. The ubiquitous nature of mobile devices provides language learners with an opportunity to learn languages at any place and at any time, it has been shown that mobile learning technologies are used spontaneously by language learners to fill out “pockets of time” and thus mobile learning applications should be effective in these small intervals of time in which they are being used Elt (2017).

- Context sensitivity and individuality: One of the most important affordances of informal language learning in a mobile context is the ability for language learners to seek out and engage with online communities and authentic language use which is specific to a topic or pertains to a certain register. This is because Idiomatic or “native sounding” language use is context specific: The meaning of words, the conventionalised use of words and expressions, and the range of vocabulary and phrases encountered within a piece of realized language is highly specific to things such as topic, register and context as well as dialect and region Sockett (2014) p.49. To illustrate this, three example English sentences with the same meaning but different registers is shown below:
Our technician repaired the fault on 12th June. Now it’s your turn to pay us.

Although the fault was repaired on 12th June, payment for this intervention has still not been received.

The company laid him off because he didn’t work much.

His insufficient production led to his dismissal.

If you need any help give us a call.

Should you require any assistance, please feel free to contact us ...

<table>
<thead>
<tr>
<th>Informal</th>
<th>Formal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our technician repaired the fault on 12th June. Now it’s your turn to pay us.</td>
<td>Although the fault was repaired on 12th June, payment for this intervention has still not been received.</td>
</tr>
<tr>
<td>The company laid him off because he didn’t work much.</td>
<td>His insufficient production led to his dismissal.</td>
</tr>
<tr>
<td>If you need any help give us a call.</td>
<td>Should you require any assistance, please feel free to contact us ...</td>
</tr>
</tbody>
</table>

The wide availability of different content available on mobile devices allows language learners to engage in informal language learning activities which match their language learning goals. The language learner is able to seek out digital spaces and multimedia content which matches a certain desired register. By engaging with this content they can notice and learn words and phrases which are relevant to those goals Godwin-Jones (2018). This is particularly important for language users who are aiming for fluency in their target language for personal or professional reasons. For instance, somebody trying to improve fluency when working in a foreign country would benefit highly from being in contact with more formal registers such as newspapers articles and business blog posts. On the other hand, somebody trying to assimilate socially in a new area will want to learn casual language use such as slang and informal speech, which may even be specific to a region or dialect. As we can see from the table above language use varies notably depending on context and mobile applications need to take this into consideration in their design to provide the ability for language to learn context specific idiomatic use.

• Social connectivity/interaction. Mobile devices allow users to share information, collaborate and communicate with ease. This has a lot of potential for language learners which are targeting language use as the learning outcome of their informal language learning practice. It has been shown that language learners interact with online communities to practice foreign languages Godwin-Jones (2018). These spaces include social media as well as affinity spaces’ and allow users to practice language production with native speakers on topics personal to them. This allows for language learner’s exposure to and production of topic specific content, motivated by sharing interests and information with native speakers and other language learners. Additionally the fact that virtually all smartphones are equipped with a camera and microphone means that a mix of textual and aural language use opportunities arise within these digital spaces. Sockett (2014) observes how Facebook provides different opportunities for language use, such as online chat, posting texts...
or images, and commenting on others posts.

Interestingly social media apps and online communities provide multilingual spaces and “social media feeds”. In these spaces, language learners can stumble upon and engage with a foreign language within an online space which is primarily in a language which they are fluent in. For example an English speaker can read a post from a Spanish speaking twitter account which he has followed and can read and comment on these posts. As these feeds such as on Twitter, Instagram and Facebook are to some degree customisable, by managing which accounts are followed or liked, social media and other feed based websites such as Youtube or Reddit give language learners the ability to control what proportion of their daily online content is in a target language which they are learning. Overall, interaction with these digital spaces, where language learners are free to engage with communities and explore multimedia content in that language or specific to that region provides the ability to immerse oneself in a language and culture, and provides highly engaging and personalized language learning opportunities [Godwin-Jones (2018)].

Informal learning tools such mobile learning apps have the opportunity to enhance the learning opportunities within these online spaces [Godwin-Jones (2018)]. We will explore how machine translation and the use of crowd sourced tools can augment these online spaces into language learning environments. These environments will provide the ability to learn and explore new language use at the word and phrase level and incorporate these spaces into a structured ‘input plus’ [Hulstijn et al. (2013)] revision system generate dynamic and intentional vocabulary learning exercises in the form of flash cards and fill in the blanks.

### 3.3 Machine Translation for language learning

In this section we present literature on the topic of Machine Translation in the context of language learning, highlighting that despite the fact that focus has been put on online machine translation being used as a crutch for written exercises in a classroom context, these systems can allow language learners of all levels to engage with authentic language use within digital spaces, in a way which is personal to their goals and motivations. Machine Translation has made large progress in recent years [Macken et al. (2019)], creating new possibilities for online machine translation to be used in a variety of fields such as professional translation as well as language learning. Additionally the fact that online machine translation is accessible from digital devices including mobile phones means that
it has been largely embraced by language learners \cite{nino2020}. 

Literature on the usefulness of online machine translation as a language learning tool is largely contradictory, this is arguably due to the rapid rate at which machine translation is improving. For instance the shift to Neural Machine translation systems were only adopted in 2016 and improved causing an estimated 60% less translation errors then previous statistical machine translation based systems: “The A.I. system had demonstrated overnight improvements roughly equal to the total gains the old system had accrued over its entire lifetime.” \cite{lewis-kraus2016}. This large shift in performance means that literature which pre-dates 2016 relied on statistical models, which produced lower quality output. This improvement results in more naturally sounding translations and a better ability to translate idiomatic language use \cite{ducar2018}.

The value of Machine Translation as a educational tool has a limited body of research \cite{lee2020}. Yet studies highlight the potential benefits of machine translation within the arsenal of language learners’ digital tools e.g \cite{nino2020}.

Discussions about the potential of machine translation within language learning has been limited to language learning in scholarly context \cite{lee2021}. The wide availability of Machine Translation has generally been seen as a disruptor rather than aid to language learning \cite{jolley2015}. Tools such as Google Translate have been viewed by educators as a source of academic dishonesty and as a way for language learners to complete work and engage with the language. A survey at Duke University showed that 88% of second language learners admitted to using online machine translation and while 77% of teaching faculty disapproved of it. \cite{clifford2013}

These tools are seen as disruptive to the language learning process, used as a shortcut for language production exercises as in \cite{nino2009} where a study of advanced Spanish learners reported that 75% of them used online machine translation ”to create quick drafts to build on” as part of their studies. This unhealthy dependency on machine translation can be seen as a way for student’s to easily complete work without properly engaging with the topic and thus limiting potential progress towards proficiency in their target language. Despite this perceived threat to Language Learning in a scholarly context \cite{case2015}, some research points to the fact that language educators should see machine translation as something that should be embraced rather than rejected such as in \cite{benda2014} or more recently in \cite{nino2020} where it is stated that

”From all these data, we can conclude that the use of OMT (Online Machine
Translation) technology does not seem to be a hindrance for ILL (Independent Language Learning); it actually seems to help with (mostly written) comprehension, vocabulary in context and as a quick language checker for small (written or oral) utterances”.

Overall the literature has been repeatedly reinforcing the fact that machine translation is a potential hindrance to language learning when used to circumvent written exercises. There has been very little analysis of the merits of machine translation, as highlighted in Niño [2009], “its wide availability, immediacy, and multilingualism, as well as its success in producing short lexical units and simply structured texts”. According to Briggs et al. [2018] learners’ perceptions of Machine Translation have improved considerably as output accuracy continues to rise and it has been shown that it used by language learners to engage with media in L2 Godwin-Jones [2015] Epp [2017] Bradley et al. [2017].

Machine translation can provide affordances to informal language learning applications, where language learners are given the opportunity to translate words and phrases which they do not understand so that they are given the means to discover new idiomatic language use at the multi word level. Additionally this would allow language learners at different levels of proficiency to engage with authentic material which may be outside of their comfort zone. By providing the ability to translate some words and phrases which the user does not understand, the language learner is afforded the possibility of learning more vocabulary incidentally through context.

One point to note is that within social media use, because language use on these spaces is filled with typos, emojis and nonsense words which translators are not aware of, the ability for machine translators to provide support in spaces with a large amount of written communication is limited.

In the context of translation for reading support, their is less need for accurate output. Machine translation output can produce two types of errors, fluency and adequacy errors Martindale and Carpuat [2018]. Fluency refers to lexical, grammatical and stylistics aspects of the translation. While adequacy relates to whether the meaning of the original phrase is present in the translation.

In the act of translating short phrases and sentences as a way to understand parts of authentic language use a language learner may be unsure of, there is much less importance put on the fluency of the output text then in tasks such as using MT to inform writing exercises as explored in Niño [2020] or when used as real time support conversa-
tion. For instance translating "How are you" google translate will resort to the informal “¿Cómo estás?” and as such would be incorrect in a formal settings, but it can be said that such a strict reliance fluency does not exist in the context of reading where the focus is on allowing the reader to comprehend and follow text in a way which limits frustration and allows engagement with the L2 language at any level of proficiency.

The affordance to explore more authentic multimedia content means that language learners will be able to engage with varied online spaces and registers specific to their needs. Learning vocabulary and phrases which are highly relevant to their language learning goals.

In order to provide a way to learn vocabulary intentionally in a way which is flexible to register and language learning goal, the implemented informal language learning system allows users to add word and multi word chunks of language to a personal vocabulary list. Thus formal language practice such as vocabulary building exercises can be produced from informal language practice, allowing for both intentional and incidental language learning which is specific to learner goals, highly engaging, and topic specific.

This dual approach means that language learners can remain flexible in the way they choose to learn languages based on their preferences and goals.

### 3.4 Designing digital language learning tools to fit specific needs

In this section we highlight the importance of user centred design and a close analysis of migrants learning practices and needs when providing tools which support language learning for migrants.

We underline the importance of user driven design for language learning tools, particularly for migrants. We then describe how the needs of migrants for informal language learning on mobile devices have not been addressed, and how an understanding of the language learning needs of people who use language in their day to day lives may motivate designers and technology creators to address those needs by proxy.

In order to design an application which is trying to assist language learners such as people trying to assimilate within a culture, technology designers and educator need to
consider the needs of people in different situations and from different backgrounds, how learning can happen at different levels of language proficiency, and with different language learning goals in mind. Kukulska-HulmeAgnes (2019). While popular apps may take the approach of targeting the largest possible market to maximise their potential user base, their is an opportunity to better facilitate language learning by making these digital tools personalised for a specific use case as well personalizable by the end user Elt (2017). In Elt (2017) personalization in the context of teaching is defined as ”a teaching approach which takes account of learner differences and preferences while enabling groups to reach pre-determined goals”.

In “Mobile Language Learning Innovation Inspired by Migrants” Kukulska-HulmeAgnes (2019) observes that there has been a wealth free digital and mobile resources targeted at migrants. This is because the flexibility of mobile learning presents affordances to the language learning needs of migrants: “Language teaching approaches based on established materials and classrooms cannot fully meet the requirements of the newcomers in terms of language learning related to their personal situation, and in terms of the practices of accessing language learning classes in the midst of complex life circumstances and limited means.” Kukulska-HulmeAgnes (2019).

Because informal language learning is varied, complex and highly personal, Godwin-Jones (2015) designers of future mobile applications should not lose sight of identified issues and solutions found in the production of these applications, as well as continue to learn from the informal language learning practice of migrants including the resources and apps they are embracing as well as what they value and dislike about them. “working collaboratively with migrants to facilitate a voice and a role in ongoing developments will be key to ensuring that innovations are fit for purpose and support not only language learning but also wider ambitions of equity, participation and social inclusion.” Kukulska-HulmeAgnes (2019).

Kukulska-HulmeAgnes (2019) observes that for a variety of factors, most of the existing and easily accessible mobile language learning tools were not built with refugees in mind, despite the size and the specific needs of the refugee population, yet there exists opportunity in learning from migrant’s experiences and the “technology supported language learning opportunities that migrants are embracing” and expand these conclusions to informal mobile language learning as a whole. They state:

“the motivation from educators, policy makers and technology developers to define, describe and provide resources for the learning and language practice
which can take place out of class. . . . . . might come from the realization that innovations resulting from a concern with supporting migrant learners can also benefit other mobile populations, such as students and business people, who share similar needs for more adaptable and individualized language learning.”

This conclusion outlines the value of the experiences of migrants as well as the projects which try to address their language learning needs. It also stresses the importance of including the user in the design of digital learning tools for language learning and social integration, particularly with migrant populations whose experiences and situations are unique and under-represented within traditional design considerations of commercial language learning applications. In the statement above, Kukulska-HulmeAgnes (2019) uses the term “migrants”, which is taken to mean: “People who have been forced to leave their homes and move to another country or region are a growing, educationally underserved population on a global scale. Those who migrate voluntarily can also experience difficulties obtaining necessary education or training for their specific needs” and which are a subset of “mobile populations” which additionally include “students and business people”. This group is defined to be: people who need to use foreign languages in their day to day due to short or long-term relocation and thus have similar needs for more “adaptable and individualized language learning” than Language Learners in a scholarly context. They conclude that an analysis of the informal language learning practices of migrants, including the resources and apps they are embracing and what they value and dislike about them could inform the design of new digital language learning tools which target “mobile populations”. As well as inspire innovation and provide insight into the potentials of mobile language for “mobile populations”, this could serve to incentivise creation of tools which serve migrants, allowing them to benefit more from MALL applications which traditionally do not address their needs.

As Kukulska-HulmeAgnes (2019) suggests, in order for this application to be “fit for purpose” and fit the complex and tacit (needs which cannot be put into words) and latent (needs people are not aware of) needs of migrant learners one must include them in the development process, exploring their needs and the way they embrace digital tools as part of their language learning practice. We will use a ”user centered” design approach to analyze the needs of this subset of FLU language learners. This is a participatory design strategy which focuses on users and their needs at each phase of the design process. In this design approach user interviews and group discussions can be used to identify user needs and the context in which they use a system. Information about a user’s needs is “co-constructed” Van Mechelen et al. (2017) through discussion and observation of a user’s past, current and potential experiences to provide the best possible tool. Because
of the close contact with designers, participatory design strategies are able to identify tacit needs as well as latent needs.

3.5 Previous implementations of personalised learning

A number of mobile learning applications which target personalised learning are presented in "Personalised of language learning through mobile applications" Kukulska-Hulme (2016) with the goal of "illustrating different ways of personalised learning through mobile devices" with the motivation that learning beyond the classroom is not widely explored and that current language learning applications attract large numbers of subscribers for various reasons, but many learners do not persist in learning due to the difficulty of self motivated learning. We will present two such studies in this section. We will discuss how their findings may relate to the aim of this dissertation: how linguistic tools can be integrated into a mobile language learning applications to facilitate some of the language learning practices and address some the language learning needs of migrant language learners. We discuss finding which show learners are enthusiastic about engaging in authentic level aware content, which can be improved further by allowing for that content to be personalised. We also present a study which uses manual content recommendation and manual reading support. The results showed that reading comprehension was improved when reading material was manually adapted to the learner, providing insight into the usefulness of a system which automates the same affordances.

3.5.1 Audio News Trainer: Allowing students to engage in authentic audio material, separated by level of difficulty

One of the studies described was the an Audio News Trainer Read and Kukulska-Hulme (2015), an app which was designed to promote aural comprehension and presented fixed audio recordings with three different levels of difficulty including speech speed and accents. Additionally the app had a feature which allowed users to post summaries and comment on others summaries on Facebook, to provide motivation through social engagement. The app was designed to be used independently and research was undertaken to study it’s motivation properties. The results found that students were enthusiastic to engage with the audio content and sharing there resumes, while relatively little engagement regarding summaries posted by others.
From these findings we see that providing users with authentic content which is of the appropriate level can motivate their engagement with informal learning strategies. It also shows that allowing users to share aspects of this engagement may be a motivating factor.

This study provides a case study which illustrates the motivating factors of informal language learning but fails to address the importance of personalisability within a framework which allows for informal language learning. This is noted in the discussion:

"Furthermore, just as our interest level varies during real news broadcasts on the radio or television, since certain topics affect us more or less than others, then in a future version of the app it might be possible to add a function whereby students can label the content of given recordings so that subsequently students can receive it categorised and only listen to categories that particularly appeal to them. Finally, it should be noted that students..."
also positively value the way in which they can use background knowledge on the content of a given recording as a useful scaffolding element when trying to understand the recordings.

The importance of allowing the users to define their own content with which they will practice languages informally is expressed above. The suggestion made is to allow users to define categories, such that methods to discover these recording can be used by users to engage with content of their choice.

One aspect of informal language learning which is addressed in this study is providing a way for learners to access pre-existing content by difficulty. This is important both for aural and written comprehension to facilitate intake and reduce frustration for the learner. While not implemented, our framework for discovering content through RSS feeds could additionally label that content with a level of difficulty. These algorithms exist and our effective while being language agnostic [Ribeiro et al. (2019)] and thus, such a strategy would simply require articles within the app to be analysed with such methods.

While it has not been investigated, podcast and audio recording are similarly distributed with RSS feeds and in fact was the method used in [Read and Kukulska-Hulme (2015)], as such the system for content discover presented within this application could facilitate aural comprehension in a similar manner to this study.

In the application presented, to allow users to discover articles of interests to them we take the approach of harnessing RSS feeds generated by Google News. This provides ways for learners to explore topics and websites without necessary knowledge of websites and content feeds.

3.5.2 Collaborative Reading Comprehension support: Manual content discovery and user generated word annotation for reading comprehension support

Another presented study explored collaborative reading comprehension assistance [Hsu et al. (2013)], where students are able to make comments on words to help each other and to collaboratively annotate the document. The content this reading was done one was based on articles teacher selected from a survey on the students interests, which were
level appropriate.

Implemented in 2013, we see a lot of similarities with our application, it is in essence it provides a manual framework for recommending articles from teachers based on surveys, with in situ assistance for reading comprehension taken from other students annotations in a collaborative reading assistance. As such the application presented in this study can be seen as a feasibility experiment into how a similar framework can be developed using RSS feeds to automatically define content feeds from the Web and provide in situ reading comprehension assistance using dictionaries and automatic translation.

Figure 3.2: The application allows users to view annotations created by other peers while reading articles selected by teachers.

"The experimental results show that students in both of the experimental groups who learned with the adaptive articles (i.e., the ones recommended by the learning system based on their preferences and reading proficiency levels) achieved significantly better reading comprehension in comparison with the students who read non-adaptive reading materials in the control group,"
implying that the personalized recommendation mechanism was helpful to the students in motivating them and improving their reading effectiveness.”

Post-test results showed that personalised learning material improved their motivation and reading effectiveness. By providing automated content discovery based on user specified key words and websites we provide accommodate ”adaptive articles” without the need for a teacher to manually recommend articles based on an interest survey.

One interesting aspect of this application implementation which is in common with the ANT implementation is that language learners are invited to collaborate and share their interpretation of text, or results from an inquiry with lexical resources such as dictionaries to engage with text. This aspect of ”collaborative reading” can motivate learners through social engagement.

3.6 Analysis of migrants needs and language learning practices

In this section we present the findings of semi-structured interviews conducted in Epp (2017) and Abou Khalil et al. (2019) relating to developing mobile assisted learning tools for migrants. We elaborate on the identified knowledge gaps within these interviews and re-frame these finding with respect to the application presented in this study, while critiquing the limitations of the approach taken in this paper to facilitate the practices and respond to the needs of migrants. The first study explores the latent and tacit needs of migrants, revealing that while some of their needs can be supported by the presented application, migrant learners also express needs regarding social inclusion, friendship and the need to speak with foreigners. The second focuses on the language learning practices of migrants, and ways in which migrants use technology as part of their language practice. We conclude that the findings regarding the current practices of migrants, and their need for level aware, personalised and contextual language learning is well supported by the presented application.
3.6.1 Interviews with Syrian refugees in two different contexts, the limitations of learning on the web

Abou Khalil et al. (2019) investigates the language learning needs of two specific groups, Syrian refugees in Lebanon who are trying to learn English to emigrate to Canada and Syrian refugees who are in Germany trying to better integrate into society. By analyzing and contrasting the needs of these two separate groups it is shown how language learners needs are dependant on their personal situation. Participants in Germany wanted to learn languages to better integrate into society, find work and "become a German citizen", they reported the need to meet locals, integrate into society, feel 'equal' to Germans, and improve their aural communication skill. Participants in Lebanon focused on learning English to leave Lebanon for an English speaking country, or to succeed in school and report the need to be surrounded by English and the need to have the opportunity to speak to foreigners. We will discuss these findings with respect to how the proposed applications fits these needs, and critique the limitations of informal learning on mobile devices:

It is noted that both groups were small making the findings of this study not entirely robust. The methodology and findings of this study are outlined below:

To identify the tacit latent needs of these two groups, they were engaged in exploratory sessions which investigated which tools they currently used to learn languages. This was followed by a brainstorming activities where invited to explore challenges which they faced throughout their language learning. They then brainstormed about solutions to learn particular "lexicon", e.g lexicon to find a job, or to buy food in the supermarket. Finally they asked to prototype these solutions on paper. Tacit and latent needs were identified using a general thematic analysis of the interviews. This involves familiarizing oneself with the data, searching for themes within the collected data, CODING the data by theme, defining and naming the themes and finally producing the report.

Identified tacit needs across both groups include:

- A need for time management: many migrants reported not having time to learn languages due to their busy lives.
- A need for reviewing of material: participants report forgetting words that they learn. Participants in Germany, who are trying to learn German to integrate into German society report hearing uncommon words which they do not have the opportunity to review.
A need for "social learning": Participants in Lebanon reported wanting to practice language learning with other language learners, describing the solo efforts of language learning as lonely. On the other hand participants in Germany desired practicing the language with Germans.

Focus on the affordances of mobile learning express the ability to provide practice across contexts and within small pockets of time. The need for time management is a common need in both groups, it could be argued that a focus on informal learning will provide the opportunity to learn languages across contexts and activities such as reading the news.

The participants express the need for reviewing of material. Words which are encountered are forgotten as they are not being re-used, this highlights the need for structured language practice to review material previously learned. However the limitation of the approach presented here is that it only allows for this review process to occur for vocabulary acquired during reading within the application. Further work could therefore allow learners to take note of vocabulary during their day, such as through voice and written notes, which is then reinforced with intentional learning strategies presented here. In this way, the mobility of mobile learning as a educational tool is better suited to different contexts, which aligns with identified affordances of mobile learning regarding mobility and ubiquity.

Participants in Germany expressed need to interact with locals, to practice communication and conversation as part of efforts to better integrate into society. This highlights the limitations of trying to situate learning on the web, language use is primarily for interpersonal communication and as such, learning through reading has limited scope.

Identified tacit needs of migrants in Lebanon:

- Need for motivation and discipline: Users expressed difficulty with engaging with self driven language learning, three participants expressed the need for teachers which would "make them" practice languages.

This highlights the importance of personalised learning for self-driven language acquisition in the context of a learner who is not living in a L1 speaking country. Within this context, their may be greater difficulty to motivate language learning, as learners do not have the opportunity to engage with the language as often. Approaches which situate learning as part of other tasks could help these needs by providing external motivations besides learning a language, the need to get informed, to seek entertainment, or to engage in educational content. Identified tacit needs of migrants in Germany:
• Need for Contextual Vocabulary Teaching: The participants expressed needs for specific vocabulary knowledge, and a difficulty with disambiguation of the sense of (polysemous) words in different contexts in German. "Moreover, participants stated difficulties understanding and using bureaucratic vocabulary or field-specific terms.”

Here we see the importance of situating learning within particular language use, and allowing learners to learn words in context. Vocabulary acquisition of words and grammar does not provide the tools to know the contexts in which they should be used, and as such learning should occur in context, where words and phrases are observed within language use. Participants also found it challenging to engage with bureaucratic vocabulary, highlighting the fact that register and topic has an effect on the ability for a learner to comprehend language.

Latent needs expressed throughout both groups include:

• Need for Self-Expression: All the participants expressed the latent-need for self-expression such as how to introduce themselves.

• Need for fun: All participants expressed needs for fun applications, migrants in Germany expressed a need to watch movies based on vocabulary they know: “The content we find is usually either boring or very difficult to understand”

The need for self expression was different in the case of migrants in Germany and in Lebanon. In Lebanon learners wanted to express themselves better for job interviews, and to obtain visas. On the other hand the need for self expression amongst German participants related to communicating with local Germans. Here we see that language learners have specific goals and motivations within their language use, and personalised learning would therefore benefit from topic specific learning. However the need for communication reflects a need for communication practice rather than topic specific vocabulary acquisition, again showing the limitations of reading to learn languages.

The need for fun once again reflects the need to situate learning based on content of interest, yet it can be to difficult to understand. This supports the importance of providing support for learners to naturally engage with foreign languages, such as seamless translation support as well as level specific content.

Identified latent needs of migrants in Lebanon:

• Need for Foreigners: During sessions based on generating prototypes, many ideas involved meeting and talking with foreigners to practice the language.
• Need for the presence of English: Participants in Lebanon wanted to interact constantly with English in order to learn the language, one participant stating, "I would like to have the name of everything in the supermarket in English."

The need for the presence of English highlights the affordance of a tool which support content discover and cross disciplinary practice as it can provide opportunities for language learners to engage with the language whilst not having the opportunity to do so in their environment such as the participants in Lebanon.

However, again, this opportunity to expose language learners to the L2 language is limited to digital forms of language use, whilst this study shows that migrants desire to converse with foreign speakers in person.

Identified latent needs of migrants in Germany:

• Need for German Friends: Migrants expressed the need to feel socially integrated to have friendships with locals.

• Need to feel equal: Migrants expressed the need to feel equal during conversation with Germans, this was due to the fact that they could not communicate as well as native speakers.

The need for social inclusion and cultural integration into society sets them apart from other language learners in a classroom context as stressed in the literature, and this need is not addressed specifically by the application presented. However language learning opportunities which allow language learners to improve their proficiency, use of idiomatic language and pronunciation can facilitate integration into society Kukulska-HulmeAgnes (2019).

Overall the findings of this study highlight that migrant learners have specific goals, motivations, challenges and needs. We conclude that our application addresses the need for fun, an "entertaining application", for content rehearsal, for exposure to the target language and for contextual and pragmatic vocabulary teaching. Yet we conclude that with respect to migrants, many of their needs are not met, as these needs relate to social and cultural challenges which can only be indirectly addressed through language learning. These include the need for friends, for opportunities to converse with foreigners and the need to practice self expression. One point which may lead future enquiry into the design of this application is the desire of participants in Lebanon for Social learning, a desire to learn languages with "fellow compatriots", which could be addressed within the context
of a mobile application, as is shown in the two studies explored above, by allowing users to share their progress, or compete against one another.

3.6.2 Interviews with Canadian migrants regarding language learning practices and technology use

Another study which conducted similar Participatory Design principles but in this case with 18 migrants shows that language learners which are migrating for personal reasons, or for education and work, i.e not forcibly displaced have similar language learning needs. In order to address "limited understanding of how migrants initiate mobile tool use to support their learning of the dominant language and their interactions with that language", an overview of previous analysis regarding the tools which support language learners to attend gaps in their knowledge which they "notice" while engaging in authentic language use, be that production or intake, are listed the following quote:

"A variety of tools that include mobile translation services and dictionaries can be used to fill gaps that migrants have noticed (Demmans Epp 2016b). These gaps include lacking vocabulary knowledge (Demmans Epp 2015), lacking pronunciation knowledge (Demmans Epp 2016b; 2016c), and the inability to communicate (Demmans Epp 2016a). Some specific tools further aim to provide targeted support for text comprehension or other language skills. Among them is a tool that allows learners to substitute synonyms within texts that they are having difficulty understanding (Veras et al. 2014) or practice the pronunciation of words (Munteanu et al. 2013) and fill any gaps they have noticed."

To build on top of the work listed above, Demmans conducted semi structured interviews to try and insight into the following research questions:

- 1) how they support their communication and language learning needs
- 2) which of their communication and language learning needs are not being met by current MALL tools.

They note that from an analysis of the tools which they appropriate as part of their language learning practice, we can infer ability gaps and user needs relating to their habitual language use. Key findings from semi structured interviews include:
3.6.3 Findings regarding Learner-identified knowledge gaps:

- Participants were not deeply concerned about ensuring their grammar was correct, or their pronunciation perfect, rather their primary concern is vocabulary.

- Lower level proficiency learners struggling to determine the meaning of a vocabulary item through context.

- “In contrast, higher proficiency learners (Ju, Pio, Zhen, Miao, Ana, and Davi) identified many vocabulary gaps but reported being able to use context to determine a word’s meaning.”

- “Participants emphasized gaps in their vocabulary knowledge because these gaps were seen to gatekeep participant access to information as well as their communication with others: “the words [were] too professional” (Shu) and “sometimes I just forget the word the person is speaking and I don’t get the meaning of the word” (Davi)”

- Language learners were concerned about pragmatic language use, ”They were specifically concerned about the appropriate use of vocabulary and colloquial expressions.” and that this concern was addressed by engaging with authentic media, such as TV and Music.

**Key takeaways:**

This study concludes that vocabulary acquisition is main concern of the migrant learners interviewed in this study, while other aspects of language learning were not as focused on.

Interview responses highlights the importance of providing ways to allow lower proficiency language to engage with authentic text and quickly attain the meaning of words, as they struggle to infer meaning from context.

Participant responses also emphasize the effect which register has on their ability to understand speech.

Participants wanted to know how to use language in a socio-cultural appropriate way, using vocabulary and colloquial expressions to sound idiomatic.

3.6.4 Findings regarding technology use
• Despite the fact that migrants have been identified as having to overcome communica-
tion challenges throughout their day to day lives, there was a general reliance on tools which supported receptive knowledge, and focused on studying rather than interaction such as translation during conversation.

• Language learning outside of English language courses included practicing their receptive knowledge by watching videos, reading articles.

• For reading, participants sought out materials of an appropriate linguistic level.

• "They supported this activity using dictionaries and indicated their frustration with the difficulty of learning English through statements, such as “It’s all hard” (Luis), “there’s so much to learn” (Ana), or “it’s just too hard to comprehend it all” (Miao)"

• "Participants had appropriated an array of general-purpose technologies (see Table 2) to support their language-learning activities. These activities included listening, reading, and writing. For example, participants would use L1 subtitles to verify their comprehension of video dialogue or they would use English subtitles to support their decoding of dialogue. However, none of these tech-nologies met learners’ desire for additional planning and rehearsal opportunities. ”

**Key takeaways:**

Migrants language practice relates to intake and study rather then overcoming communication challenges, this practice is informal, and in the case of reading they needed material with the right level of difficulty.

The use of dictionaries was used to overcome proficiency gaps in reading but frustra-
tion was expressed at how these tools are not sufficient to support reading.

Migrant learners wanted a way of planning and structuring their informal language prac-
tice.

**3.6.5 Discussion**

It is noted that a focus on vocabulary learning was more prevalent for lower proficiency learners, higher proficiency learners had a wider range of self identified knowledge gaps. As well as this an analysis of the tools which a language learner would benefit from based
on his current practice and use of digital tools has limited value due to the fact that wider language learning practices such as "socio-collaborative" approaches to learning, or a wider focus on language production and conversational skills may simply not be supported by any tool, digital or not, outside of a classroom setting and thus would not be identified as a potential need under this framework.

Here we see a marked advantage of the exploratory approach in the first study relating to Syrian refugees. By asking participants to envisage a "perfect future", participants can think outside the box and address needs current technology use has little impact on. Whilst in this study, a analysis of their current practices grounds the discussion in what migrants are trying to address, not what they aught to address. However one advantage of this approach is that it provides insights in ways of facilitating and improving current language learning practices.

The study concludes that linguistic tools that migrants use and are exposed to do not provide meet their specific challenges:

"ELL (English Language Learners) experiences indicate that too few tools scaffold the larger learning challenges faced by these migrants. These challenges include ELLs' ability to communicate; understand multiple registers, accents, and varieties of English; monitor their own learning; and obtain socio-emotional support."

The overall conclusion of the study was similar to Kukulska-Hulme Agnes (2019) observations on the importance of personalised learning, which is that language learning tools should provide a structure for self directed learning "Perhaps, most importantly, tools need to start addressing larger learning needs. It is no longer enough for tools to target specific knowledge components. It is time for tools to move towards developing an independent learner who can identify what she or he needs to learn and then appropriately target his or her activities towards developing those skills."

We feel that the approach presented in this paper through the application prototype implemented, addresses a number of the suggestions of this analysis. Providing an environment for content discovery, which enables topic and register specific practice does allow the language learner to, as noted in the conclusion, "identify what she or he needs to learn and then appropriately target his or her activities towards developing those skills". Within targeted and well supported reading comprehension language learners are able to engage with that language in way which limits frustration expressed in the interviews.
The app also provides the framework to learn vocabulary and colloquial expressions in a pragmatic and context sensitive way which is structures their informal language practice, provides opportunities for rehearsal.
Chapter 4

Informal language learning, dealing with and learning idiomatic language in chunks

One of the stressed advantages of informal learning is it’s ability to expose learners to idiomatic language use and as such idiomatic expressions Godwin-Jones (2018).

In an effort to explore how a learning applications can facilitate the learning of idiomatic phrases and collocations in context to allow learners to be able to produce naturally sounding language, we will elaborate attributes of these phrases, and outline considerations which need to be made for a learning application to teach language learners these phrases.

4.1 A view of language, statistically frequent word combinations

Informal language learning primarily relies on users interacting with authentic use of language, this can be input through text and video or through output and interaction in the form of communication with others.

As Godwin-Jones (2018) points out, one view of language informed by corpus linguistics and discourse analysis sees language “above all as a set of patterns and conventional word groupings. Studies examining real language exchanges show that language use is characterized by repetition, reuse, and re-purposing of chunks of language Beattie and Ellis (2017)” This points to the fact that authentic language is formed from typical/s-
statistically frequent word combinations or “collocations” as well as fixed expressions or semi-fixed expressions such as idioms. In this way formal views of how language can be produced from grammar and lexis don’t provide the tools for the generation and interpretation of fluent language production. Because natural sounding language is pervaded by patterns which extend beyond the word level, it is not enough to construct language from an understanding of grammar and vocabulary, instead language learners need to learn conventionalized phrases and collocations:

"Instead, we have offered an alternative perspective on the nature of language, as well as its acquisition and processing, which puts meaningful sequences of all lengths at its core. Acquiring language involves becoming increasingly familiar with the sequences used by a linguistic community, along with an increasing mastery of the ways they can be processed, organized, and combined Chater and Christiansen (2018)."

4.2 Irregular word combinations: Multi Word Expression

These linguistic “patterns and word grouping” contain Multi Word Expression (MWEs) which are linguistics units which are multi-word and contain a level of idiomaticity.

These multi word language units dominate authentic language use. They provide a number of challenges to language learners in that they are vital for naturally sounding language productions but they are by definition divergent from the habitual composition of its individual words. These divergent properties do not follow the rules of the language but instead are arbitrary and conventionalised according to the language. This divergence can occur in terms of meaning, (semantic) their role in the sentence structure (syntactic) the possible inflexions in which they can appear (lexical rigidity) and their likelihood to occur over alternatives (statistically).
A “level of idiomaticity” can be taken to mean that the expression somehow deviates from the habitual language properties of the component words of an MWE, whether that be, lexically, semantically, syntactically or statistically. “The idiomaticness or idiomaticity, is the syntactical, grammatical, or structural form peculiar to a language. Idiom is the realized structure of a language, as opposed to possible but unrealized structures that could have developed to serve the same semantic functions but did not.”. Correct or naturally sounding grammar and syntax within a language is often arbitrary to that language. Language which follows the conventions of that language is said to be idiomatic, unidiomatic phrases such as a “fast shower” sound wrong to fluent speakers due to convention despite the fact that the word constructions are often totally comprehensible and grammatically correct. To underline the importance of taking MWEs into consideration when designing language learning applications, we discuss how the example MWEs annotated in Fig. 2.3 could cause issue for a language learner.

- **Taken in:** This is an example of a word combination which has semantic idiomaticity, the meaning of these two words in this context is an “idiomatic one”, in this case “bring to” as opposed to literal meaning “take in”. A language learner would have to know both the idiomatic and literal meaning of the word combination and in which contexts they are used. For example the difference in meaning between “My wife had taken her ‘07 Ford Fusion in for a routine oil change” and “were these pictures taken in Germany” is cannot be inferred from the meaning of ‘taken’ and
• Oil change: This is an example of statistical idiomaticity, idiomatic uses “oil change” as opposed to “oil swap” which would sound unnatural, this is an arbitrary rule of idiomatic English which has been conventionalized and thus a language learner would have to know this rule through experience and by seeing the “collocation” in several contexts. These conventions are unpredictable and follow no obvious rule, for instance it is idiomatic to say a cold shower but not too say a cool shower.

• ‘07 Ford Fusion: This is an example of a proper name, these have arbitrary meaning in that the sense of the entire phrase does not stem from the meaning of ‘07 or Ford or Fusion. These can be multi-word and as such generally behave as a fixed noun which contains spaces. These do not generally pose problem to language learners.

Because these divergent properties do not follow the rules of the language but instead are arbitrary and conventionalised according to 'language use', their divergent properties must be learned by example and in context.

Some other examples of MWEs include: ad hoc, by and large, The Chair, kick the bucket, part of speech, in step, trip the light fantastic, foundation model, call (someone) up, take a walk, do a number on (someone), take advantage (of), pull strings, kindle excitement, fresh air, (https://people.eng.unimelb.edu.au/tbaldwin/pubs/mwe2021.pdf)

4.3 MWEs variable flexibility which need to be considered by language learners and technology creators

Native speakers are generally not aware of the arbitrary constraints placed upon idiomatic language use, as they don’t encounter alternative, unidiomatic constructions. But because these constraints are arbitrary language learners need to learn these constraints and patterns to achieve a high degree of proficiency. MWEs such as collocations restrict language use, in that they form habitual word combinations which are necessary to communicate meaning in an authentic way. Such As Sung (2003) has stated, sufficient knowledge of collocation is one of the essential parts of language proficiency, having positive role in improving language learners skills including
speaking, writing, reading and listening. Furthermore, a lack of this knowledge is an important obstacle to the attainment of native-like fluency and accuracy [Nation (2001), it can be said achieving a higher level of proficiency involves learning the idiomatic use of these single word forms [Godwin-Jones (2018)]. Because idiomaticity or “realized structure of a language” occurs at a phrase or multi-word level, we see it necessary to explore characteristics of these idiomatic multi-word units with the goal of analyzing how a mobile learning application may teach idiomatic language use enabling language learners to achieve a high degree of fluency in a language.

4.4 Varying forms of idiomaticity

For this we will borrow definitions and formal analysis of these multi-word language units, which contain idiomaticity, from Natural Language Processing literature. NLP pipelines need to correctly manage the semantic and syntactic idiomaticity of natural language. This allows machine translation systems to correctly translate expressions as shown below.

![Figure 4.2: “In short” is translated](image)

![Figure 4.3: “Call him up” is translated](image)

(MWEs) have many definitions, such as in (Ivan / Baldwin 2002) where they are defined as “idiosyncratic interpretations that cross word boundaries (or spaces)”. We will take their definition to be a sequence of words that display lexical, semantic, syntactic, pragmatic and/or statistical idiomaticity” [Baldwin and Kim (2010)].

Put plainly, they refer to any multi-word language unit which exhibits some linguistic differences to what would be expected from the composition of its individual parts. MWEs can display the following idiomaticities, or divergences from regular language use:

- **Lexical idiomaticity**, [Baldwin and Kim (2010)] define lexical idiomaticity to mean that a MWE has component words which do not habitually belong to that language’s lexicon. For example, ‘ad hoc’ or ‘de facto’.
• Syntactic idiomaticity is when the syntax of a phrase deviates from grammatical constructions expected in a given language, and so the grammatical purpose of a phrase is not deductible from typical language use. The example given in Baldwin and Kim (2010) is by and large as it is the coordination of a preposition and an adjective, which would normally be considered an incorrect grammatical production.

• Semantic idiomaticity is the property that the meaning of a MWE is not derivable from the meaning of its parts, often described as the transparency of the expression. An example is kicked the bucket, which has extremely divergent literal and idiomatic meanings. This property is not a binary one and so MWEs display a varying amount of non-compositionality, for instance working like a dog does not mean to work in the way a dog would, but has a semantic meaning that is "partially predictable" from its constituents yet it should not be interpreted literally.

• Statistical idiomaticity, Like semantic idiomaticity statistical idiomaticity is a non-binary property which places MWEs on a spectrum. A phrase which is statistically idiomatic is one which "occurs with markedly high frequency". This means either that the frequency of the phrase relative to the frequency of constituent words is high. Another possible interpretation is that its frequency of use is higher than other word constructions which signify the same thing (Manning and Schütze 1999). Phrases that display a high degree of statistical idiomaticity are sometimes described as institutionalized phrases Sag et al. (2002), they are a way to express an object or idea in a conventional way, for instance "black and white television" is more conventionalised than "White and black". This property captures the "naturalness" of a certain phrase construction. (Baldwin and Kim 2010), other examples of collocations include “telephone booth” as opposed to “telephone box”. Below is a table of English collocations paired with potential alternatives which do not sound natural:

<table>
<thead>
<tr>
<th>Natural sounding English</th>
<th>Unnatural sounding English</th>
</tr>
</thead>
<tbody>
<tr>
<td>The fast train</td>
<td>The quick train</td>
</tr>
<tr>
<td>Fast food</td>
<td>Quick food</td>
</tr>
<tr>
<td>A quick shower</td>
<td>A fast shower</td>
</tr>
<tr>
<td>A quick train</td>
<td>A fast train</td>
</tr>
</tbody>
</table>

As we can see the concept of collocations captures the intuition that within natural use, phrase constructions display arbitrary and unpredictable idiomaticity. The existence of a conventionalised collocation gives rise to alternative phrase constructions which sound extremely unnatural such as “A quick train” which has been coined an "anti-collocation" Pearce (2001).
4.5 MWEs: Idiosyncratic language use which needs to be considered

Formal definitions of multi word expressions try to describe or underpin the way in which authentic or “realized” language contains idiosyncrasies and constraints relative to regular language use at the multi word level. For instance the word idiom used to classify multi-word expressions can be defined as “the realized structure of a language, as opposed to possible but unrealized structures that could have developed to serve the same semantic functions but did not."

Because of their complex and ambiguous definition, see Savary et al. (2019) it is hard to evaluate what proportion of language they form. Jackendoff (1997) estimates that the number of MWEs which a native language learner knows is in the same order of magnitude as individual words. Another estimate is in Lewis and Conzett (2000), stating “collocation is found in up to 70 percent of everything we say, hear, read, and write”. MWEs such as collocations restrict language use, in that they form habitual word combinations which are necessary to communicate meaning in an authentic way. Sufficient knowledge of collocation is one of the essential parts of language proficiency, having positive role in improving language learners skills including speaking, writing, reading and listening Nation (2001). Furthermore, a lack of this knowledge is an important obstacle to the attainment of native-like fluency and accuracy Nation (2001).

We conclude that because of the fact that such a large proportion realized language contains idomaticities, be it semantic, syntactic or statistical, which form a constraint on and inform what is correct and authentic language production, a language learning application that intends to teach language at a high level of proficiency and “idiomaticity” (the extent to which a learner’s language resembles that of a native speaker source) needs to consider how language learners can learn these “word chunks or phraseological patterns”. Such a learning environment needs to consider how language learners can explore, interact and learn these idiomatic cities which are present at a multi word level.

This constructive view of language learning which informs the design of the presented application is informed by this quote in Scholz (2017) which was presented in Godwin-
Jones (2015) discussion of the merits of informal language learning online.

“Rather than conceptualizing language as rule based, language can be thought of as a collection of patterns that are observed through repeated use. In this sense, learners of the language notice these patterns and replicate them, as opposed to learning a concrete set of grammatical rules from a textbook. Scholz (2017).

In the next section we will explore how a Language learning tool can support intentional and incidental "vocabulary" acquisition of these language chunks, that is to say intentional and incidental learning of multi word language units.

4.6 Considerations to be made in order to teach idiomatic language chunks

To conclude this chapter, we will underline the considerations which have informed the design of the application in the goal of teaching multi-word to promote language learning which for fluency and awerness of idiomatic use of language. There are a number of considerations to be made regarding Multi Word Expression which will be explored in this section. These observation can be used to inform the design of a language learning app which allows for users to encounter, identify and re-use chunks of language as per Scholz (2017).

The following observations are made:

MWEs vary in the degree of divergence from “regular language use”, they can be partially or fully semantically non-compositional where the meaning of the whole cannot be fully inferred through composition of its constituent parts, for example “taking the train”, where the presence of train changes the meaning of the word to take, i.e the subject is not “taking” the train in a literal sense. Other examples where the verb is changed by the presence of a particular noun include, to “give a demo” and to “call up”. The degree in which these multi word chunks exhibit semantic idiomaticity varies from case to case, for example the meaning of idiomatic phrases such as “kick the bucket” or to “let the cat out of the bag” cannot be deduced in any way from the individual words. At the other end of the spectrum you have what Sag et al. (2002) coins as “Institutionalized phrases” which
have no semantic or syntactic idiomaticity. Because of this, the use of dictionaries is ineffective at providing meaning of non-computational combinations of words. On the other hand machine translation is able to provide equivalences in the L1 language and allow language learners to evaluate the non-compositional meaning of multi-word linguistic units.

MWEs range in terms of their semantic, syntactic and lexical flexibility: Because by definition these linguistic units display behaviors which are divergent from the classical rules of the language, the semantic and syntactic behavior of these multi word units cannot be learned by a classical understanding of grammar. Instead the behaviors which MWEs portray can only be learned through the repetitive encountering of these word chunks \cite{Chater and Christiansen (2018)}. We illustrate this with the following example:

A language learner which encounters “in short” once needs to see other examples of its use in multiple contexts to deduce a number of qualities relating to that phrase. These qualities are listed below:

- An understanding of when the phrase is meant literally and when it is meant idiomatically: The idiomatic meaning of “in short” roughly means “to summarize” but a literal meaning can sometimes occur such as in ”the dog was running in short grass”. A language learner would need to see multiple examples of the phrase to fully understand its meaning as well as the conditions under which its idiomatic and literal meaning are expressed. Notably this requires the learner to be able to discover and interact with both examples of idiomatic and non-idiomatic uses. Furthermore a system which is trying to present examples of fixed and semi fixed expressions would need to provide examples and counterexamples of both. Because in the case of fixed expressions, no words can be placed in between its constituent words, such as “in very short” a language learning system may want to show this by providing examples of sentences which use the idiomatic phrase and the non idiomatic phrase along with a translation. This would allow learners to identify when a phrase is meant idiomatically and when it is meant literally.

By allowing language learners to inspect phrases in context using machine translation, they can evaluate when a phrase is meant idiomatically or literally.

Flash card examples which show the phrase in idiomatic or literal instances provides ways for the learner to learn when the meaning is meant literally or idiomatically, however these examples need to be accompanied by a translation of the phrase, which is not the case in the current state of the application.
Additionally this reliance on machine translation can only work if MT systems are able to detect when a phrase is meant, idiomatically or literally themselves. This is a known point of difficulty for MT systems [Ducar and Schocket (2018)], particularly for languages with less preferment coverage. A language learner cannot learn idiomatic phrases through the inspection of translation output if that output does not detect an idiomatic instance of the phrase and instead presents and literal instance of a phrase.

- The lexical flexibility of phrase [Sag et al. (2002)]: Fixed expressions such as "in short" can be treated as words-with-spaces [Sag et al. (2002)]. Examples in English include, by and large, in short, kingdom come. These undergo no inflection and only appear in their original form. For instance language learners would need to know that it would be improper to use variations such as “in shorter”, or “in very short” to express the idiomatic meaning of the phrase “in short”. Some multi-word units such as Semi-fixed Expressions can display levels of lexical variation such as spills the beans. Semi-fixed expression can contain words in between its constituent words and can change the order with which words appear, for example “spilled the beans” and can appear as ”the beans were spilled”, or ”give a demo”, ”a demo was given” and ”give a revealing demo”.

- The cultural connotations of a phrase: A language learner would need to see examples of how this idiomaticity behaves and is used within certain contexts. Just as words are context specific, multi word units such as expressions have certain socio-cultural connotations, and are to be used in the correct context. For instance "in short" is associated with a formal written context, use of such an expression in casual conversation would seem un-idiomatic and rude. Language contains different ways of expressing meaning using phrases, for example the fixed expression “to a degree” and “more or less” are multi word synonyms which are used in different registers. Just as with words, expressions need to be discovered and language learners need to interact with these language chunks multiple times either through informal language learning or through example sentences to know which contexts these multi word units should be used.
MWE have different meaning in different contexts, for instance the phrase "take away" is translated in four different forms in the figures above. This illustrates how language learners need to see multi word expressions in different contexts, to be able to learn to disambiguate the sense of MWEs. Further more it is not enough to provide a dictionary definition each individual word, one solution to this is to show the different possible meanings of a phrase through a translation of that phrase being used in different contexts.

Particularly on the internet, the idiomatic qualities of language are ever evolving such as new ways of interpreting certain expressions and the invention of new terms and proper nouns [2002]. This means that lexical resources cannot be relied on to keep track of the lexical, syntactic and semantic particularities of specific instances of language use found throughout the world. This highlights the usefulness of machine translation, which is able to provide the meaning of idiomatic language use.

Additionally [2002] points out that Multi word expressions are inherently specific to different topics and contexts and that new fields generate new vocabulary and new multi word units with idiomatic qualities. This stresses the importance of being able to practice informal language learning within certain topics and registers. Particularly for
the specific requirements of migrant learners who need to learn idiomatic language use within varied instances of language use such as in education, in professional settings and even within variations in regional and national dialects. One way to do this is situate learning within different digital spaces which contain idiomatic language use specific to that space and the topics within it. Informal language learning situated online can provide these affordances as there is a large volume of interaction with authentic language use specific to a certain topic.

In order to encounter, identify and re-use these chunks of language or patterns [Chater and Christiansen (2018)] within this large volume of interaction mobile applications can underline instances of phrases which language learners have encountered in the past and which they are trying to learn as they have intentionally selected them for further study. As such phrases are easily noticed in situ and thus a language learner can learn the idiomatic nature of that phrase, i.e. the contexts in which that phrase is used idiomatically, its meaning in different contexts and the possible lexical inflexions which are possible for that phrase.
Chapter 5

Facilitating vocabulary acquisition of words and phrases within informal language learning on mobile devices.

This chapter serves to analyse how the presented application facilitates incidental and intentional vocabulary acquisition.

5.1 Incidental language learning

This section introduces the concepts of formal learning, informal learning, and incidental learning, presenting arguments which point to the importance of incidental learning within language acquisition and situates the terms within the context of Mobile Assisted Language Learning.

Formal Learning is a form of language learning which is organized and structured often following a curriculum. This form of learning can include textbook approaches such as writing exercises, reading comprehensions as well as intentional vocabulary acquisition exercises such as flash cards and fill in the blanks Eaton (2010). Language learning resulting from activities associated with formal learning such as language learning exercises is known as intentional learning.

Informal language learning can happen spontaneously, informally and even involuntarily as one simply engages with authentic language use while seeking entertainment, working or socializing Sockett (2014). By simply interacting and producing language, language is learned, intentionally or not. This is known in the literature as Incidental Learning Eaton.
“The term incidental learning is used, in applied linguistics, to refer to the acquisition of a word or expression without the conscious intention to commit the element to memory such as “picking up” an unknown word from listening to someone or from reading a text.” Hulstijn et al. (2013)

Incidental learning is normally associated with input based language use such as reading and listening. The meaning of words and phrases which users do not understand are obtained from context allowing for vocabulary acquisition. It is estimated that adolescent native speakers of English who have completed high school have a receptive “they are able to understand” vocabulary of at least 20,000 base words, i.e. words that have no inflected form, open and opened is not counted twice. (Nation 2001.) “The accepted view is that they cannot have learned such a large number of words solely by means of explicit vocabulary instruction. Rather, they must have learned most words in an incremental way through repeated encounters through listening and reading.” Hulstijn et al. (2013)

Huckin and Coady (1999) states that once the first one thousand words of a foreign language are learned intentionally in the classroom, most vocabulary acquisition occurs incidentally.

Digital learning tools such as MALL applications can enable informal language learning by providing a way to explore, interact and learn within digital spaces such as social networks, video games and affinity spaces and while consuming multimedia content such as articles, videos and movies.

5.2 The need for a tool which allow for repetitive encounters with words in clear contexts:

In this section we present three ways of facilitating incidental vocabulary acquisition through reading comprehension for language learners. We identify three factors which affect the rate of retention for incidental language learning; these are the frequency of encounter, the quality of context in which these words are encountered and the involvement with which the meaning of words are evaluated. We observe that this is of particular interest in the context of digital reading where browsing and skim reading is common.
5.2.1 The frequency affect:

The following section presents the effect of word exposure frequency to incidental vocabulary acquisition and how a mobile learning app may increase this rate of exposure for a set of vocabulary which a language learner is trying to learn. We also explore how highlighting words may alert the reader’s attention to that word so that they may focus on the occurrence of that word in its context, providing a deeper processing of the text and thus perhaps increasing the rate of word retention.

According to a large amount of research, incidental vocabulary acquisition depends on the frequency with which words are encountered by the learners as they read text or listen to language use. Webb and Chang (2015) outline how numerous studies have shown the positive effects of word exposure frequency on incidental word acquisition (Pigada and Schmitt 2006; Rott 2007). For instance, Pellicer-Sánchez and Schmitt (2010) study showed that words with an exposure frequency of 10 or more yielded better results than words with fewer encounters.

These studies show that a number of encounters with a word were needed to learn the spelling, grammar and meaning of a word. Waring and Takaki (2003) Additionally they show that with increased encounters, words are more likely to be retained.

This illustrates how an informal language learning environment can increase the chance of incidental word retention from reading by increasing the rate with which words which the learner is trying to learn are encountered. To allow language learners to encounter words frequently a digital language learning environment can do the following things:

- Provide engaging content in the target language which the user is motivated to read, watch or listen to, as he finds it informative, engaging, and or by allowing him to replace multi-media consumption in his daily life with content in the L2 language.

- Allow the user to engage with topic and register specific context such that words and phrases which he is trying to learn as part of a purpose specific and informal language practice are encountered more often.

- Make instances of words which the user has chosen to learn, more noticeable through highlighting and underlining, such that instances of said word will be noticed and the language learner can focus on those word instances with more intent, inferring the meaning, such that the user can retain the spelling, grammar and meaning of
the word.

This is especially important when reading on digital screens where it has recently been emphasized that accessing content digitally leads to “skimming” as opposed to reading. Ribeiro et al. (2019) When consuming content on digital screens it has been reported that much less deep reading occurs. Digital reading often involves scanning and browsing, looking for headings and links to navigate pages. These forms of reading may not allow for incidental learning, as such highlighting words that language learners are trying to learn can provide them the opportunity to focus in and engage with that particular instance of that word in its context.

- Provide more example contexts in which that word is used: By providing more example contexts in which that word is used, the learner can learn the spelling, grammar and meaning of the word in different contexts. This is done through formal learning exercises which present the word in context such as fill in the blanks exercises and flash card exercises.

5.2.2 The context affect:

This section presents how the degree with which a word’s context can hint to a word’s meaning or “quality of context” has a positive effect on the rate of word retention from incidental language learning through reading. We present how allowing for seamless access to lexical resources and translators within a digital learning environment can allow language learners to infer the meaning of words and support incidental language learning.

For the meaning of a word to be learned incidentally, one needs to be able to first notice it, and the meaning of the word needs to be guessed from context:

“readers do not always notice unfamiliar words when reading a text. If they do, guessing the meaning is not always possible” Hulstijn et al. (2013).

Importantly it has been shown that context plays an important role in the learning of incidental words, “The words that are encountered in a rich and clear context with relevant cues to the word meanings are more likely to be learned.” Teng (2019) A reader’s ability to infer the meaning of a word is thus affected by the “context” that word is encountered in. This “context” is subjective to the reader as it relates to how well he is understanding the text and the words surrounding a word which is being inferred.

A digital learning tool which is supporting incidental learning therefore must provide a way to create high quality contexts, taking a high quality context to mean:
“A rich and clear context with relevant cues to the word meanings are more likely to be learned” [Teng (2019)]

We see that for a learner to learn a word through the act of reading comprehension they must be able to infer the meaning of the word from contexts, encountering a word multiple times without any way of inferring it’s meaning results in little chance of being able to learn that word’s meaning.

For a language learner to infer the meaning of a word, he needs to have a context with words which he understands surrounding it. The need for “high quality context” highlights the importance of providing ways for learners to access content which is suited to their level.

When that is not possible we propose that giving language learners instantly accessible tools to discover the meaning of words using lexicons and translation tools such that the meaning of words surrounding them is provided improves the subjective quality of context of that word. Both because words in the direct context of that word can be defined through lexicon or translation but also because the overall sense of the passage or text is understood better by the reader.

Furthermore, providing instantaneous access to lexical tools and translation means that reading in a foreign language can occur at lower levels of proficiency. It is noted that a reader’s ability to follow a certain instance of language use is dependant on context, topic and register, and as such these tools additionally allow for language learners to practice reading comprehension in a wider variety of contexts, with less frustration and greater engagement.

The combination of providing high quality contexts in which to encounter word instances, as well as providing the means to encounter word instances with a high frequency as has been described above means that such an informal language learning environment supports incidental vocabulary acquisition in such a way as the learning is topic specific, engaging and where word retention is maximized.

In “The effects of context and word exposure frequency on incidental vocabulary acquisition and retention through reading” [Teng (2019)] explores the interaction of the two stating: “In other words, the repetition of new words in more informative contexts seems to have a greater effect on the acquisition of word meaning than the repetition of words in less informative contexts. This highlights the importance of more informative contexts in
the acquisition of word meaning, and may partly explain previous findings on difficulties faced by learners in comprehending the meaning of words despite multiple encounters (Pigada and Schmitt (2006); Teng (2019); Waring and Takaki (2003)). The contextualisation of the target words may have been insufficient.”

5.2.3 The involvement factor: an advantage of interacting with text

This section introduces the involvement hypothesis for second language vocabulary learning, and discusses the way in which a “lookup” of a word or phrase’s meaning can change the degree of involvement in evaluating a word’s meaning, which may have an effect on word retention. In the context of the application presented, we describe how translation and lexical dictionaries can be used to provide lexical information about text encountered in an article be it a word or a phrase, and describe how this environment allows for the user to listen to utterances of words present in Wiktionary and look at translation of certain words in context evaluating how they are behaving syntactically and semantically, which will increase involvement further. We then conclude that designing tools for vocabulary acquisition in an informal setting needs to consider the degree of involvement needed to evaluate the meaning of words, and that this can be a trade off with ease of use and cognitive load.

In Incidental Learning in Second Language Acquisition Hulstijn et al. (2013) presents the relevance of depth-of-processing theory, which states that the:

“the chance that some piece of information will be stored in long-term memory is not determined by the length of time that it is held in short-term memory but rather by the shallowness or depth with which it is initially processed. “

This theory underpins the involvement hypothesis for second language vocabulary learning. Under this model the retention of unknown words encountered in vocabulary learning are conditional upon the degree of involvement of a word. Under this framework, involvement consists of:

• a) a motivational component, comprising the need to determine a new word’s meaning, and

• (b) a cognitive component, comprising of a search (e.g., dictionary look up)

• c) evaluation (e.g., evaluating whether the information obtained from the dictionary applies to the verbal and nonverbal context).
Simply put, this theory illustrates how the act of recognizing that one does not understand a word’s meaning, doing a search of that word’s meaning using some resource and evaluating whether the found meaning applies to the context in question has some effect on the degree to which that word will be retained. It reflects a common intuition that acting on a thought such as taking note of something that passes through one’s mind will increase the likelihood of them remembering said thought later without having to check the note.

Hulstijn et al. (2013) quotes Hans Eysenck’s “memory performance is determined far more by the nature of the processing activities engaged in by the learner than it is by the intention to learn per se.”

Within the context of language learning, this can be seen to say that it is not one’s level of desire to retain a word which will have an impact on whether or not they will retain that word, but far more important is the way that word’s meaning is processed. This idea can be related to many memorization strategies commonly used by people, such as repeating something out loud, taking note of something or making sure to include a piece of information in conversation as to retain it. A language learner thus is engaged in one of these strategies which increase involvement and thus retention when they: encounter in context, realize that a word’s meaning cannot be understood from prior knowledge or inferred from context, do a lookup of the word meaning using some tool and evaluate whether the result fits in the words context.

This shows the advantages of providing a way for language learners to partake in “processing activities” by interacting with text. There are a number of studies which support this within a reading context such as Chen (2012) Çakmak and Ercetin (2017) which investigated the influence of dictionary use on vocabulary acquisition in reading contexts and concluded the consultation of external materials in an attempt to make up for lexical gaps in knowledge is a highly effective practice for incidental vocabulary learning.

According to Zhang et al. (2021) there have been some investigations into whether people prefer to use a monolingual or bilingual dictionary, for instance an English speaking person trying to learn Spanish could use a Spanish to Spanish dictionary or an English to English dictionary. In Atkins and Knowles (1990) an overwhelming majority preferred bilingual dictionaries, but that monolingual dictionaries were superior to bilingual ones for that they were of a higher precision due to the fact that translation equivalence between languages introduces inaccuracy in meaning. This is to say that there does not exist a
perfect translation of a word from one language into another as “languages do not simply name existing categories, they articulate their own” Culler (1976).

A dictionary is a listing of lexeme or infinitive forms of a word for a lexicon. For instance a dictionary will contain the entry for the lexeme football which will contain the inflection footballs. In the case of Wiktionary the inflected form of a word will contain redirection to the lexeme word entry football. Within the word entry can be found the definition, pronunciation and examples uses of the word. Entries generally correspond to definitions, usage examples, etymologies, and pronunciation.

**Involvement in the application**

We present how in the presented application, the lookup and evaluation of words is drastically changed relative to dictionary look up thus having an effect on the involvement of fetching the meaning of words. This will have an effect on the degree to which the word will be retained in memory.

When a language learner is performing an evaluation of some text to understand its meaning, this can be due to not understanding a single word within that chunk of text or multiple words. The user can choose to highlight either, and is provided with a translation. If the highlighted text is an entry in Wiktionary the user has the opportunity to view that dictionary definition of that word, which includes its infinitive form, its current reflexive form, examples sentences for both, and the option to listen to audio recordings of the word being uttered.

When a single word is translated a machine learning model does not have context to select an equivalent word form, as such it will simply choose the most probable translation based on the most probable word mapping from one language to another. Upon highlighting a single word the prototype presented in this paper will thus simply present the most common translation with no awareness of context. If the user evaluates the translation to be wrong based on the context, he can thus select to see a dictionary definition of the word form in a bilingual dictionary.

In the prototype presented language learners will be able to read a large amount of authentic text and when they encounter a word which they do not understand the possible sequence of “processing activities” is underlined below:
Motivational component: Noticing that there is a word/phrase which has a meaning which is not understood.
Cognitive component: highlighting the word/phrase, bringing up the translation provided
Exploration: Listen to an utterance of a the word
Exploration: Choose to add word to user-list of words
Evaluate: Part of speech
Evaluation: Evaluate whether the translation provided matches the context of the word/phrase instance
If evaluation fails:
In the case of a word: If the highlighted word does not match the context then you can choose to highlight the context of the word such that you are providing more context to allow the translation software to give a more accurate translation of that word instance.
OR
Press the lexical definition of the word form which will bring up the dictionary of the word, as well as the current inflection the word is found under.
The user must then choose which sense from the dictionary definition which matches the context.

As we can see as well as providing rich contexts and bridging a proficiency gap, the act of interacting with text in order to evaluate its meaning creates a degree of involvement which should have an effect on retention. In the case of single words present in Wiktionary this interaction also provides an opportunity to look at example sentences and listen to native-speaker utterances of the word, increasing involvement further.

One advantage of the approach taken in the application is that by providing a way for language learners to translate a phrase which may be containing a word the language learner is trying to evaluate, we are providing the affordance of seeing how that word behaves syntactically and semantically in context which will increase involvement further.

Finally, the designing of systems which allow the user to interact with text, considerations need to be made in how the digital tool can increase involvement and facilitate word retention. For instance, providing a translation of a word inline may reduce the degree of involvement but increase the ease of use and reduce the cognitive load required to do a lookup of words. In the case of the application, the reason this approach was not taken was to allow for interaction and evaluation of multi-word units such as phrases.
5.2.4 Using flashcards to provide frequent exposure to context:

This section introduces the concept of “input plus” described in Hulstijn et al. (2013) which has been used to supplement incidental vocabulary acquisition by getting learners to reprocess the meaning of a word in an “elaborate” way. It then presents how the prototype application provides a form of this input plus strategy by showing words and phrases in many example sentences within flashcard exercises. The implementation shows that Tatoeba can be used to supplement flashcard exercises with example sentences, and that these flash card exercises can be used to promote intentional vocabulary acquisition on a restricted set of words and phrases which the user has chosen to learn.

Hulstijn et al. (2013) gives an analysis of why language learners would benefit from a language learning approach which combines large amounts of input supplemented with intentional vocabulary learning exercises which “process” their meaning again after reading. Hulstijn et al. (2013) goes on to say that “Rich, elaborate processing, however, is not enough either” stating that new information will not be retained if not “frequently reactivated”.

Hulstijn et al. (2013) notes that dozens of studies e.g Peters et al. (2009) give empirical evidence that vocabulary acquisition through reading (“input only”) can be improved by getting students to process their meaning in an “elaborate way”, and to process them again in the form of exercise. (“input plus”)

As such, as well as providing a digital tool where word instances are encountered frequently and allow users process their meaning in an elaborate manner as described in the previous section, an application which gives language learners the opportunity to process those meanings again later on through some other form of “processing activity” will supplement vocabulary acquisition from reading in a foreign language.

The application presented in this paper give users the ability to add words to a list of words which they want to learn. This list is specific to their account and represents a set of words the user is currently trying to learn. The user is then invited to frequently process their meaning in an ”elaborate” way. This is done through intentional language learning flash card exercises.

These flash card exercises can be accessed frequently. They present a word or phrase in varying contexts, inviting the language learner to infer the meaning of the word from context.
Before looking at the lexical definition of the word the user cycles through instances of the word in varying contexts, starting with the instance in which the word was first added.

The provided examples include the word form(s) of the word or phrase with different inflections and in the case of phrases potentially having words removed and added in between constituents. This allows the language learner to evaluate his understanding of the word or phrase with different examples from different contexts and registers, and in different inflections.

Once the user has guessed an inferred or known meaning of the word or phrase from the provided examples he will look at the answer. In the case of a word that answers in the form of a dictionary definition and in the case of a phrase that answers in the form of the translation the context from which that phrase was pulled.

Additionally if the example sentence has an utterance provided from a native speaker the language learner can listen to that audio clip.

These exercises create many frequent opportunities for the meaning of a word to be inferred from varying degrees of context. If we are to believe that subsequent interactions with a word in different contexts could lead to some form of retention of that word’s meaning, then these exercises will be able to supplement incidental vocabulary acquisition.

5.3 Annotating Parts of Speech in translation outputs: Aiding beginner learners engage with text

This section presents how part of speech tags can be used to show syntax highlighting of natural text which we describe as “part of speech highlighting”. We argue that this can be used to expose beginner language learners to the syntax of a language in the early stages of language learning and to enrich the context in which words can be inferred, allowing for the inferall of words at a lower level of proficiency. This section additionally presents a novel method of presenting translation output which was developed based on user feedback to allow easier reading of translations and to invite learners to construct their own word mappings. We propose that this may have some pedagogical merits by inviting the user to perform deep processing of words in reference to translation in a known language and we draw parallels to the act of observing and creating word mappings of
historical languages.

In an effort to make incidental vocabulary acquisition more accessible to language learners with a lower level of proficiency, the prototype allows the user to annotate both the source text and translated text with part of speech highlighting. In a way this is a form of “syntax highlighting” which can be found in programming environments for natural languages such as English. As shown in the image, every word which is selected for translation is highlighted with the word’s color corresponding to its part of speech including nouns, pronouns, adjectives, verbs etc. To the author’s knowledge this is a novel idea in the context of a language learning application. Found examples of projects which highlight the natural language with the part of speech of words include https://english.edward.io/ and onlinepartofspeechtagger.net. These are used for education on the subject of part of speech or simply for entertainment purposes.

5.3.1 Providing a rich context in which to infer meaning and to expose beginners to a language’s syntax.

Within the concept of incidental language learning, we present two arguments regarding the potential applications and affordances of highlighting the pos of speech of text inline while reading.

We see an opportunity for this form of context enrichment or text augmentation particularly for beginner learners. For a language learner which does not recognise a word, it allows the user to see how that word is behaving semantically and syntactically in that sentence. This provides the opportunity to gain knowledge about the grammatical and semantic accordance of words at phrase and sentence level both for words which are unknown and to solidify understanding of words which are known. This could enable exposing beginner language learners to the syntactic structure of a language.

Additionally language learners may be able to infer meaning of words due to part of speech highlighting of text. This is enabled by the fact that part of speech tagging is an extremely accurate process in natural language processing. Although not implemented, this affordance of part of speech highlighting text could be used to create a richer context in which words can be inferred and reduce the level of proficiency required to infer the meaning of words while reading. For instance seeing that a word is a noun means that the user may be able to guess the meaning of that word based on context when he would not have been able to do so otherwise.
5.3.2 POS highlighting to make it easier to read translation output and to simulate translation alignment learning

Given that the prototype additionally shows the translation of the source text we identify potential affordances of “part of speech highlighting” both the input and output of a translation. We observe that the effect is similar to a word alignment output.

Translation alignment is a task within natural language processing with applications in machine translation which involves creating a mapping of parallel corpora, from which parts of the first text correspond to which parts of the second text. The resulting mapping can be one to one, but is often many-to-one, one-to-many, or many-to-many as such manual alignment is a complex task [Benda (2014)].

To the author’s knowledge there is no literature on the use of translation alignments for language learning. However, in one instance, there has been an investigation of how translation alignment can help teach historical languages such as ancient Greek and Farsi. In Translation Alignment for Historical Language Learning: a Case Study, [Benda (2014)] analyzes the pedagogical merits of the creation of manual translation alignments and inspection of translation alignments for historical language learning. The act of creating a word alignment was shown to have some pedagogical benefits by promoting a deep processing of the words in relation to the translated meaning, additionally by providing a word alignment of a source text and a translation, language learners with no prior knowledge of ancient Greek were able to explore the original text through the use of a digital tool which highlights word alignments on hover.

A language learner viewing the output of a translation is invited to create word alignment of the source text and translation, to see how the two parallel texts match.

We propose that doing so with both the translation and source text being highlighted according to their part of speech may give the following affordances:

When selecting a whole sentence or multiple words which are not understood due to a few key words, the user can more quickly find how these few key words map to the translation output.

A beginner language learner viewing the output of a translation may be able to start exploring how words and phrases map to their L1 language at a much lower level of pro-
ficiency than without any part-of-speech highlighting.

This could give the opportunity to explore the syntactic structure of a new language and how it relates to their mother tongue while reading authentic text.

It may also provide a way for the language learner to start assimilating the meaning of many words and phrases in L2 very quickly as the part of speech highlighting in the source and target text will allow for the learner to infer translation mappings easily. Because word mapping can be many-to-one or one-to-many or many-to-many this may also help to see how words map to phrases by a process of elimination.
Chapter 6

Conclusions and future work

6.1 Conclusions

This chapter will summarise conclusions made regarding findings and arguments presented in this dissertation, while proposing future directions that work can be taken in the exploration of how language learning apps can integrate linguistic tools to facilitate the needs of migrants.

We implemented and presented an application which allows for personalised reading of articles, with integration of linguistic tools, to provide reading support and facilitate the discovery, engagement with, learning of highly specific language use at the word and multi-word level.

In the Application Overview we provided an overview of how a language learning mobile application which situates language learning on the web may be implemented, describing the limitations of machine translation, part of speech parsers and crowd sourced solutions to provide linguistic assistance and structured practice of vocabulary selected while reading web content.

This dissertation has explored the language learning needs of migrants through an analysis of previous research which has identified their unmet needs and practices. We conclude that the application presented addresses some of their identified needs for "fun", review of material, time management, contextual language learning, but that mobile learning applications are limited in their ability to address needs which are specific to migrant populations, such as social inclusion, and the need to converse with locals. We identify from the interviews and from previous implementations of personalised and informal mobile learning applications that there is the potential for providing support for collabo-
rative language learning such as sharing and discussing L2 content with others on social media or within the application. Additionally we noted how allowing learners to take note of language which they want to practice through their day such as through voice and written notes could extend the affordances of structured informal learning revision to contexts outside of digital reading.

We highlight the need for personalised and informal learning of languages for migrant learners and presented two applications which emphasise this approach. These two implementations highlight how level appropriate and personalised content increases motivation and engagement. These studies also describe the potential for allowing learners to collaborate within digital spaces with summaries and by providing help with each other, we identify this as a potential future work from our study.

We identified how machine translation and the use of the sentence dictionary Tatoeba can assist language learning outside of a classroom context by providing the resources to engage with authentic text and allow for structured study of informal language learning practices. We then analyze how these tools can be integrated into a digital reading environment to encourage incidental vocabulary acquisition and intentional vocabulary practice. We show that the application facilitates a structured revision of informal language learning and provide the tools for context specific language practice.

We analysed how the context effect, the frequency effect, involvement factor affected incidental vocabulary acquisition within the application.

Due to the needs of migrants for idiomatic and effective language use we explore how the above findings can provide incidental and intentional acquisition of language at the multi word level. We elaborated on the fact that because definitions of idiomatic language use rely on arbitrary meanings based on how they are used within a language, machine translation is the only way to provide support for idiomatic word chunks and as such is necessary for learners to notice, use, and learn new phrases and collocations. As such we concluded that MT’s ability to translate such phrases is the primary bottleneck in allowing learners to evaluate idiomatic word chunks, this is important for less well covered languages.

Within a formal analysis of MWEs we described consideration which need to be made to allow for intentional practice, and incidental learning of multi-word idiomatic language chunks. These include the fact that they are context specific, have varying degrees of
flexibility, and are non compositional.

We described how the implemented system has implications which are not limited to the reading of articles, but function on web page and as such behave similarly to a Desktop extension. We highlighted difficulties in implanting such a strategy on mobile due to the lack of development support and sophisticated API within the react-native-webview.

6.2 Future Work

This section proposes a number of improvements to this dissertation and future works which could be undertaken within the topics covered.

Future work could include formally describing strategies to develop features to run within an integrated web-view within a mobile application, or exploring ways of allowing for easier developing of and publishing ”browser extensions” for android to enable modular language learning opportunities on mobile devices situated on web pages. On android potential solutions could include somehow integrating ”chromium browser” into mobile applications which support the installation of app extensions [Chr (2022)].

Facilitating development and publishing of mobile browser extensions could be a step forward in providing modular and purpose specific linguistic assistance on mobile devices, which is in essence what the implemented application is attempting to achieve by facilitating structured practice of informal language learning through reading on a device which is mobile and ”can sustain educationally valuable activities across everyday environments that are not designed or arranged for learning” [Kukulska-HulmeAgnes (2019)]

Research on how mobile learning can address the language needs of migrants found that they would benefit from opportunities for social integration, and contact with native speakers, rather than classical understandings of language practice in a formal setting [Kukulska-HulmeAgnes (2019), Epp (2017)], as such it is important to stress that the explored system is not a solution to language learning for migrants who are trying to integrate into society, but rather a possible addition to the tool belt to facilitate their learning practice and address unmet needs.

The dissertation would benefit from a further elaboration regarding efficient architecture solutions which provide language translation capabilities in any supported directions with low latency without having to run an instance of Opus-MT for every supported
translation direction. Currently ‘switching’ the translation direction of a translation introduces latency and slow translation until the docker container has a chance to ”warm up” and run at a sufficient speed.

It was described how Tatoeba can be used to create language learning exercises at the single word and multi-word level, further work could include augmenting these exercises further, such as ways of detecting idiomatic and un-idiomatic examples of a phrase thus providing control in which type is shown intentional phrase acquisition practice.

Because Tatoeba and Wiktionary relies on crowd sourced information for its dictionary definitions, sentences, translations and utterances future work could investigate other ways of providing dictionary definitions and sentence translation in a open source way. In the case of sentences possible solutions could be relying on simply fetching examples from large corpora generated on from the web Kilgarriff et al. [2010], which would allow for much more coverage, especially for poorly covered languages. Instead of relying on native speakers uttering entries in these databases, the system could rely on automatic text to speech for pronunciation practice.

The dissertation underlines the importance of personalised and level-aware authentic content in informal language practice. Their would be value in exploring ways of providing information on the difficulty of web content to allow the user to discover and read content which fits his needs.

Topics of content discovery using RSS feeds which fetch content by website, key word or topic could be explored to allow for content discovery which is specifically audio and video, allowing language learners to practice aural comprehension. Furthermore Machine Translation could be used to provide language support for multimedia web content, however this would rely on text to speech, which introduces further inaccuracy in the output translation.

Youtube provides a large amount of content with user submitted subtitling, thus conclusions regarding structured practice of vocabulary encountered during informal language practice, such as generating flash card exercises and underlining instances of words in the ”learner-list” could be extended to this context, it would simply require a way for user to highlight Youtube subtitles within the application, which was not achieved as part of this dissertation.
Literature on digital reading for English foreign language learning stresses that digital reading of web content does not allow for linear deep reading of text and as such incidental language learning is affected as it easier to skip over content that they do not understand Ribeiro et al. (2019), as such providing more linear forms of content on mobile devices such as by using a system similar to Firefox’s reader view mozilla (2022), which de-clutters articles may have an effect on the affordances of a system similar to an application presented.

Our attempts at using the reader view API mozilla (2022) available to create a de-cluttered form of article which only contains the essential text of the article failed as the parsing strategy often mis-parsed the content of the article, causing the application to omit large section of it. This would have facilitated for more intricate interaction with text from articles, as the HTML structure of the article would have been simplified and the application may not have needed to rely on a web-view to render content. This would have meant simplifying the content of articles, videos, images, banners, animations relating to the website would have been removed.

One direction of further study is related to methods of highlighting instances of words and phrases within a web page. In order for language learning applications and websites to highlight instances of words and multi-word chunks when they are used in a certain meaning, or under any form of inflection, non trivial strategies would have to be used. Due to characteristics of MWEs, this particularly in the case of phrases, Sag et al. (2002). This relates to MWE annotation, identifying idiomatic phrases in natural language use, which is a problem which is studied in NLP such as in Savary et al. (2019).

Although the implemented system supports any language, this application was designed, implemented, and evaluated with little understanding of non European languages, and as such no analysis was done regarding how the observations made in this paper extend to non-alphabetic languages such as Chinese, Japanese and Korean.

Additionally this is of particular importance when evaluating the use Part Of Speech tagging and highlighting throughout the application, where the use fullness of such features is in allowing users to read translation output to compare phrase production between L1 and L2.

The Microsoft Azure translation API supports word mappings output, it could be interesting to investigate the merits of presenting word mappings in a translation output
for language learning. Other features which could enhance translation output to allow
language learners to infer grammatical and syntactical rules when using online machine
translation is a syntax parse tree view, which is available on the stanford-nlp POS tagger
par (2020).

Some literature has stressed the need for user specific progress reports within informal
learning contexts, to provide feedback and motivate learners as they see they progress
in their informal language practice, this could be a potential addition to the framework
presented in this dissertation.
Bibliography


Appendix

...