School of Informatics

Informatics Research Review
Assessing Mobile Language Learning

[redacted]

January 2021

Abstract

Over the last decade, mobile language learning (MLL) applications have exploded in popularity. We review the nascent body of literature assessing the effectiveness of MLL apps, and report two core findings. First, most MLL apps are based on outdated educational philosophies, and lacking in complexity that they claim in their marketing. Second, research commissioned to determine the efficacy of these apps is often lacking and incomplete. We end with recommendations for the future of the field.

Date: Friday 22nd January, 2021

1 Introduction

Over the past decade, the continued proliferation of smartphones has given mobile language learning applications the opportunity to establish themselves as a popular alternative to traditional language courses. Apps like Duolingo, Babbel, and Memrise make promises like attaining levels of fluency similar to an average college course [Lesson Nine GmbH, 2021] and have been installed hundreds of millions of times from the Apple App Store and Google Play. While promises are easily made, it's hard to verify how effective mobile language learning (MLL) actually is. This paper will take a look at the numerous studies that have tried this over the past few years, and identify the gaps in the academic literature on the assessment of MLL.

Before we start, it is worth it to briefly categorize the different approaches to MLL. When surveying the landscape of available applications, it becomes clear that there are three main situations MLL can be applied to: first language acquisition in young children (e.g. apps for training reading and writing skills), second language (L2) learning in the classroom (e.g. online environments created by textbook publishers), and individual L2 learning such as the course-replacing apps listed above. These categories are sufficiently different to each warrant their own review, and this review will solely focus on the final category, individual L2 learning, for it is the category that is most isolated from other fields. That is, whether apps for first language acquisition or classroom learning are successful depends greatly on the environment in which they are used: the input a child receives from their caretakers, the capacity and resources of a teacher, et cetera. Only in individual L2 learning can we reliably try to assess the quality of an application or approach without the need to capitulate to external influences.

As we will see, however, studies into MLL do not always make such a distinction, leading to the occasional vague definition of what it means to be an MLL application. To keep the distinction clear, this review will start with an in-depth look at studies that assessed language learning applications individually, before presenting studies with more

general subjects. Throughout this, we will discover that most modern language learning apps are based in the otherwise largely abandoned educational philosophy of old-fashioned behaviourism, and our review will culminate in a discussion about the consequences of these findings. At the end, we will assess how much current research is able to say about MLL and what further research needs to be done.

2 Individual language learning apps

While there are dozens of different MLL applications, with each a slightly different mix of human curation and algorithmic automation, there are a few that are vastly more popular than the rest, so we will dedicate most of our attention to them.

Claiming over 300 million active users, *Duolingo* is one of the most popular MLL apps [Duolingo, 2021]. *Duolingo* offers a wide variety of languages and divides its courses into themed sections such as "Food" or "Animals" that are themselves divided in 1-8 lessons. When completing all the lessons in a section, the next sections are unlocked, until the student completes the course. Each lesson is built around three types of exercises: translating sentences directly, building sentences from individual words, and dictation (see Figure 1). A few of the most popular language courses have extra features, such as longer stories to read and answer questions about, and podcasts aimed at language learners. Except for the most popular courses, which are curated by *Duolingo* employees, all content on the app is written by volunteers, who build courses by creating the themed sections, writing sentences, and adding all their translations.

The MLL application *Babbel* takes a different approach. *Babbel* is a paid service that promises courses built by experts in return [Lesson Nine GmbH, 2021]. Compared to *Duolingo*, it offers a much more tailored approach to language learning, including exercises that try to simulate real-life conversations and detailed explanations of pronunciation. In addition to the linear course, the application includes a personalized "Review" section, which analyzes the student's progress and offers a new list of previously learned words

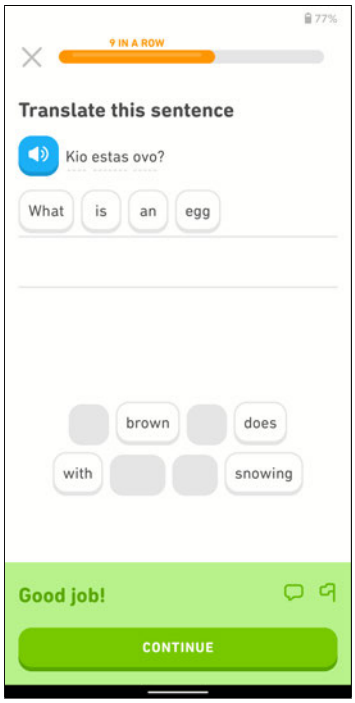

Figure 1: An exercise in *Duolingo*.

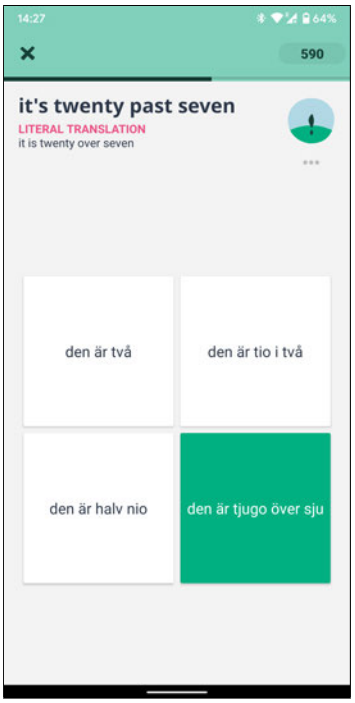

Figure 2: An exercise in *Memrise*.

to review each day. On the other end of the spectrum is an app like *Memrise*, which is almost completely based on community content and focuses on learning words and short phrases [Memrise, 2021]. Each word or phrase goes through a series of seven exercises, becoming progressively harder, each one visualized as a new stage in planting a seed that grows into a flower (Figure 2). While *Memrise* therefore offers an impressive set of vocabulary, these words are always single entities, and never combined into bigger exercises. More curated features, such as listening exercises and videos by locals, can be accessed through a paid subscription.

Next to these relatively well-known applications, a myriad of other options exists, each with their own differentiating feature. For example, *Busuu* relies on its learners to check each other's work by having them check the assignments of people that are learning their native language in-between their own exercises [Busuu Ltd, 2021], and *Lingvist* has a Course Creator mode which allows the student to enter what words they want to learn, and then picks relevant exercises from its database [Lingvist, 2021].

For most of these applications, there is no literature or rigorous assessment of their

efficacy available, though both *Duolingo* and *Babbel* have commissioned researchers to measure the effectiveness of their applications ([Vesselinov and Gregor, 2012] and [Vesselinov and Gregor, 2016], respectively). Each of these studies were of the same form. A sample of beginner-level learners was instructed to use the Spanish course of the app in question "often" for eight weeks. Before and after the study they took WebCAPE, a placement test commonly used in college language courses, to measure their performance. All participants took the Web-based Computer Adaptive Placement Exam (WebCAPE), an exam used by universities in the USA to determine what course a learner should take, before and after the eight-week study period. The point difference between the two tests was taken as the main indicator for the students' improvement rates. For example, in the *Duolingo* study, the students improved 8.1 WebCAPE points per hour of study. Given that WebCAPE requires a student to gain 270 points for them to be placed into a second semester class, rather than a beginner class, the researchers concluded that 34 hours of using Duolingo is equivalent to one semester of college Spanish. Using the same calculation, their second study concluded that *Babbel* users reached the same level after 21 hours of using the application. Both *Duolingo* and *Babbel* have prominently featured these numbers on their web pages and in their marketing.

While these studies seem to show success, they were also directly commissioned by the companies that built the courses and were only focused on one single language. Regrettably, more independent research is few and far between. That is not the only issue with these studies, however. As [Krashen, 2014] notes in a discussion of the *Duolingo* study, roughly half of the initial sample of 156 students dropped out during the program, possibly signalling dissatisfaction with the service. In addition, they point out that in university language courses, students are typically less motivated, younger, and less experienced than people volunteering to learn a language for a scientific study, possibly highly skewing the results in favour of the app. Most importantly, however, he alludes to the general idea that a test like WebCAPE is the wrong method for assessing language competency, because it is rooted in the behaviourist view of instruction, rather than the

constructivist view.

What is meant by this? For context, let us take a very brief excursion into educational philosophy. Behaviourist views of learning were pioneered by Skinner. This school is often seen as the 'traditional' approach to learning, and argues that lessons should have a rigid structure of small steps, that learners should be incentivised to learn by using stimuli like grades, and that performance can be measured on standardized tests [Beatty, 2013]. In contrast, constructivism assumes that no knowledge is standardized or objective, but rather "organized into interrelated patterns [...], constructed from all our previous experiences" [Nunan, 1993]. These patterns are called *schemata*. Whereas behaviourism assumes that a learner comes in with no background knowledge, constructivism argues that teachers should take into account what pupils already know and build on that. It encourages collaboration and struggling with ideas rather than following a set curriculum.

3 General findings

As it turns out, the vast majority of language learning apps, including the major ones named earlier, are mainly based on the behaviourist approach, as shown by [Heil et al., 2016]. They reviewed the 50 most downloaded language learning apps from the Apple App Store and Google Play Store at the time of their research and categorized them according to 149 criteria ranging from monetisation to modes of grammar instruction and corrective feedback. One of their key findings was that 84% of apps taught vocabulary as isolated words, devoid of any context.

It is striking, then, that we see that current literature has all but discarded behaviourism as a viable means of language learning. Many studies have found that interacting with language in context has significant benefits over rote learning. For example, [Mason et al., 2009] studied Japanese university students learning German. The students, all beginners, were given a list of 36 words. The story group was then told a story in which all of these words occurred, while the control group was told to simply study the

list for 20 minutes. After this, they had no further instructions to study these words, but two weeks later, the students were tested on these words again. On average, the story group got 65% correct, while the control group only came as far as 36%.

One might hypothesize there are significant differences between behaviourist approaches—more concretely, that studying words on a mobile phone is somehow different and perhaps more justified than studying words on paper. This does not seem to be the case. [Zhang et al., 2011] studied 78 Chinese university students who were practicing for the national College English Test. All studied a list of 130 words. Half received these words over SMS, five words a day, and the other half received all words on a paper list at the start of the study. After a month, both groups took a test on the words (the posttest) and a second, similar test a week later (the delayed test). While the SMS group showed a significant advantage over the paper group in the posttest, there was no significant difference on the delayed test, suggesting that any difference between the two methods only appears on the short-term.

Others seem to arrive at the same conclusion via different ways. For example, [Steel, 2012] did not instruct any group of participants to use a particular MLL application, but rather surveyed close to 600 foreign language students in general. 56% of her respondents reported using some kind of language learning app to complement their university courses, and students reported the ability to practice "anywhere and anytime" as the main benefit of using these applications. "Many" students reported using a variety of language-related applications at the same time, including dictionaries, translators, games, conjugation trainers, and flashcard apps—regrettably, Steel neither provided an exact figure nor a breakdown between different types of mobile-assisted language learning apps. Following their findings, however, they do have the grounds to hypothesize that MLL is mainly beneficial as a complement to classroom learning, rather than replacing it entirely.

4 Recommendations for future research

This leads us to something of an impasse: MLL clearly allows for some conscious learning of language and improving test scores, but many experts don't think that this is the right framework to judge language acquisition, and there seems to be no inherent benefit to using technology over paper. Two considerations flow from this.

First, there is a need for a constructivist approach to teaching language learning app efficacy. This could take many forms; an example would be a design not dissimilar to Vesselinov & Grego's design, where students are evaluated before and after using an app for an extended period of time. The change would be in the method of evaluation: for instance, the students could be made to have a conversation in the target language with a native speaker, and competency could be assessed by analysing the choice of vocabulary and grammar in these conversations.

But even this forgets one of the main pillars of constructivism, which is that learning is highly individual. When describing some basic principles for mobile language learning apps, [Stockwell and Hubbard, 2013] outline how using mobile apps can both improve accessibility, because of their flexibility and adaptiveness, and decrease accessibility, if a user has e.g. impaired visual acuity or manual dexterity. No one approach fits all, so trying to decide on whether MLL is an effective way of studying might be in vain. Does drinking tea improve studying? That wholly depends on whether you like tea or not. Perhaps the most telling are the findings by [Wang and Smith, 2013]. They built a very comprehensive system for learning English over e-mail and followed 56 Japanese college freshmen over the course of three years, assessing their performance and having them take surveys every year. Their main conclusion was that they severely underestimated the effect of the students' motivation on their research, dwarfing any other result. Students that wanted to learn the language did so, and the others dropped out, regardless of what tool they were using.

5 Conclusions

All in all, this review has two main conclusions. First, it is disheartening to see the lack of independent, peer-reviewed research into the various claims made by language learning applications, and the generally lackluster nature of the few studies that have been commissioned by these companies themselves. Obviously, MLL is a very young field, but now is the time for researchers in this field to standardize on testing frameworks, common baselines, and standards of ethics, like so many fields have done before.

Secondly, we found out that these standardized frameworks should not be based on the current pedagogical norm in these apps, as most of them subscribe to an outdated philosophy of learning. While they can be a useful tool to improve vocabulary and grammar skills for the motivated student, more specific claims, like MLL being better or worse than other tools, or specific comparisons between tools or approaches, have not yet been made effectively. In order to do that, future research will need to be both more creative in its evaluation metrics and more targeted to specific groups of learners.

References

[Beatty, 2013] Beatty, K. (2013). *Teaching & researching: Computer-assisted language learning*. Routledge.

[Busuu Ltd, 2021] Busuu Ltd (2021). Busuu. Retrieved from `https://www.busuu.com/`.

[Duolingo, 2021] Duolingo (2021). Duolingo - the world's best way to learn a language. Retrieved from `https://duolingo.com`.

[Heil et al., 2016] Heil, C., Wu, J., Lee, J., and Schmidt, T. (2016). A review of mobile language learning applications: trends, challenges, and opportunities. *The EUROCALL Review*, 24(2):32–50.

[Krashen, 2014] Krashen, S. (2014). Does Duolingo "trump" university-level language learning? *The International Journal of Foreign Language Teaching*, 9(1):13–15.

[Lesson Nine GmbH, 2021] Lesson Nine GmbH (2021). Babbel. Retrieved from `https://babbel.com`.

[Lingvist, 2021] Lingvist (2021). Lingvist. Retrieved from `https://lingvist.com/`.

[Mason et al., 2009] Mason, B., Vanata, M., Jander, K., Borsch, R., and Krashen, S. (2009). The effects and efficiency of hearing stories on vocabulary acquisition by students of german as a second foreign language in japan. *Indonesian Journal of English Language Teaching*, 5(1).

[Memrise, 2021] Memrise (2021). Memrise. Retrieved from `https://memrise.com`.

[Nunan, 1993] Nunan, D. (1993). *Introducing discourse analysis*. Penguin.

[Steel, 2012] Steel, C. H. (2012). Fitting learning into life: Language students' perspectives on benefits of using mobile apps. In Brown, M., Hartnett, M., and Stewart, T., editors, *Future challenges, sustainable futures*, pages 875–880. Proceedings ascilite.

[Stockwell and Hubbard, 2013] Stockwell, G. and Hubbard, P. (2013). Some emerging principles for mobile-assisted language learning. *The International Research Foundation for English Language Education*, pages 1–15.

[Vesselinov and Gregor, 2012] Vesselinov, R. and Gregor, J. (2012). Duolingo effectiveness study – final report. Retrieved from `http://static.duolingo.com/s3/DuolingoReport_Final.pdf`.

[Vesselinov and Gregor, 2016] Vesselinov, R. and Gregor, J. (2016). The Babbel efficacy study – final report. Retrieved from `http://comparelanguageapps.com/documentation/Babbel2016study.pdf`.

[Wang and Smith, 2013] Wang, S. and Smith, S. (2013). Reading and grammar learning through mobile phones. *Language Learning and Technology*, 17(3):117–134.

[Zhang et al., 2011] Zhang, H., Song, W., and Burston, J. (2011). Re-examining the effectiveness of vocabulary learning via mobile phones. *The Turkish Online Journal of Education Technology*, 10(3):203–214.