Polygloss - A conversational agent for language practice

Etiene da Cruz Dalcol
Queen Mary University
London, UK
dalcol@etiene.net

Massimo Poesio
Queen Mary University
London, UK
m.poesio@qmul.ac.uk

Abstract

This paper explores the impact on language proficiency of comprehensible output applied in computer assisted language learning (CALL). Targeting speakers of intermediate level, we adapted a visually-grounded dialogue task, optimizing for language acquisition. The task was implemented as a mobile application where learners are organized in pairs and write short texts to play an image-guessing game, producing samples in a wide variety of languages. Following a framework for CALL evaluation, we conducted an analysis of the game and players' gains through time, including the measure of pre-trained XLM-r cross-lingual transformers' acceptability score of the samples. The results confirm the intended fit for intermediate speakers as well as reveal possible benefits for other levels. This research provides a successful case study of a multilingual CALL design where users have the autonomy to generate output creatively.

1 Introduction

Reaching high proficiency levels and being successful at interacting with others is the ultimate goal of many adult intermediate learners of a second language. There are, however, many obstacles along this journey, related to strategies chosen by self-directed learners, accessibility of learning materials, and the influences of the natural plateau found at the higher end of the learning curve (Ritter and Schooler, 2001).

Once a learner has reached an intermediate proficiency, they have learned the most frequent words. It can then be a struggle to jump over to the next stage because only a small number of words are very frequent and the frequency quickly drops for the following words, creating an extremely long-tailed curve. Sparsity is even more of an issue when we consider that one of the features of advanced speech to be acquired are collocations. Nevertheless, when learners can understand 80-95% of the words, they can infer a lot of words through the context, causing many students to abandon active study and focus on passive consumption of foreign media. However, there is not enough repetition of the advanced vocabulary that the student needs to learn for it to become part of the productive vocabulary (Nation and Hunston, 2013). This manifests as a much higher receptive vocabulary than a productive vocabulary, the "I can understand but I cannot speak" phase.

Notwithstanding this consensus that conversational practice is essential to go beyond this phase, most commercial language learning apps do not support conversational practice and are usually only available for the most popular languages.

In this paper we propose Polygloss, a game to provide conversational practice to intermediate level learners. While not intended to tackle all the skills necessary to overcome the language learning plateau, Polygloss draws on principles from critical pedagogy (Freire, 1972) to tackle an often neglected skill, *creative production*. We investigate and highlight its importance to overall language proficiency. At the same time, we want to do that by providing a free tool that is sufficiently general and does not sideline learners of less spoken languages.

2 Background

2.1 Comprehensible Output and Languaging

As a counterpoint to the input hypothesis (Krashen and Terrell, 1983), which argues that being exposed to vast amounts of input alone is neces-

Etiene da Cruz Dalcol and Massimo Poesio 2020. Polygloss – A conversational agent for language practice. *Proceedings of the 9th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2020)*. Linköping Electronic Conference Proceedings 175: 21–36.

sary for language acquisition[1], without denying the role of input, Swain (1985), argues that comprehensible output is also necessary. The main function of *Comprehensible Output* is allowing the students to notice their gaps when they realize what they cannot say, testing hypothesis on interlocutors, and improving fluency by gaining self-confidence. While her early work is more focused on "pushed" output, where the teacher encourages students to produce language, her later work goes deeper into interaction. Influenced by Vygotsky's sociocultural theory of mind and the *Zone of Proximal Development* (ZPD) (Vygotsky, 1978), she adopted a new term, *Languaging*, to describe the "shaping and organizing of higher mental processes through language use" (Swain, 2006).

2.2 Critical Pedagogy in the future of CALL

While ZPD is informed by Piaget's theory of children being autonomous learners, it is still founded on the mediation between a student and a more knowledgeable peer or teacher. By contrast, our work was founded on Freire's method for adult literacy (Freire, 1972), a cornerstone work for the field of critical pedagogy. Freire does not place the participation of the teacher as a superior or even as a fundamental part of the learning process, but argues instead for a learning methodology centered on the student's development of agency for the purpose of reshaping social structures of power. The process starts with a search for *charged words* during an informal chat with the students using images to facilitate the discussion (see Fig. 1). The elicited vocabulary is then used to generate debate themes that allow the students to talk about their day-to-day, explore their identities and argue their beliefs. Freire's work has influenced much socially-informed work in second language acquisition (Saft et al., 2001; Anya, 2016; Benson, 2013).

Within Computer Assisted Language Learning, Benson (2013) re-frames Warschauer's stages of CALL history (Warschauer, 1996) under the perspective of user control. He notes that intelligent CALL (iCALL), powered by artificial intelligence, is often regarded as the future of CALL. However, it can still stripe users of autonomy as designers of such systems can view autonomy as undesired or even problematic. He suggests the interesting innovations will focus on self-directed learning and the development of autonomy.

[1]While Krashen makes a distinction, in this paper we use the terms *learning* and *acquisition* interchangeably.

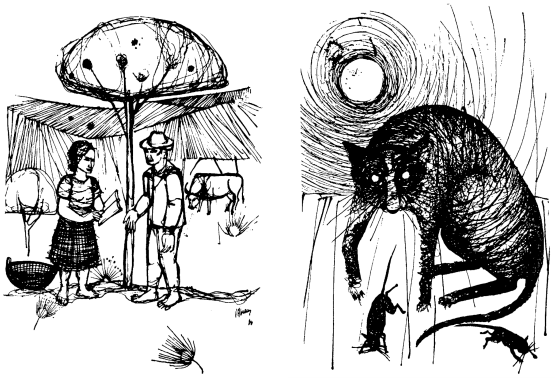

Figure 1: Two illustrations by Vicente de Abreu used in Paulo Freire's curriculum (Freire, 1967)

3 Related work

In a recent overview of the sub-field of dialogue-based CALL, Bibauw et al. (2019) review the field focusing on work where an automated system is one of the interlocutors. Although most work on computer mediated communication (CMC) is focused on written text technologies such as Wikis and Email, and employed qualitative but not quantitative methods (Macaro et al., 2012), there are nevertheless a number of relevant research papers and commercial applications we would like to mention.

WUFUN (Ma and Kelly, 2006) and TESU (Liu et al., 2014) are vocabulary trainers focused on communicative competence that go through an end-to-end analysis from theory to quantitative evaluation on the users productive vocabulary. Spanish Without Walls (Blake, 2005) is a learning program that employs a CMC application for audio and text, highlighting the importance of such tools in the context of distance learning. These applications were dedicated to teaching a single language, but MagicWord (Hatier et al., 2019), offers a multilingual word game, initially developed for Italian, French and English. Revita (Katinskaia et al., 2017) is a system with automated fill-the-gap exercises for stimulating active vocabulary. While it is proposed for endangered languages, it faces various challenges related to its multilingualism such as the lack of corpora. CALL-SLT (Rayner et al., 2010) is a system that uses a textual or pictorial representation of an interlingua to prompt users' speech in four supported second languages. Despite facing challenges like limited

vocabulary, its recognition and feedback steps are done by an underlying automated agent which was well-received by the players.

In the field of commercial mobile applications for language learning, we inspected many and perceived them as belonging to distinct groups, according to their approach: those with a tutoring approach such as Duolingo, Memrise, Busuu, Babbel, Rosetta Stone, Ling and Mango Languages; those focused on vocabulary games such as Drops, Clozemaster, LyricsTraining and Lingvist; those focused on providing comprehensible input such as LingQ, FluentU, Beelinguapp and Yabla; those focused on conversations with other learners and natives such as HelloTalk and Tandem; and chatbots such as Andy.

With the exception of the apps in the last two groups, they usually do not allow the user to create their own authentic outputs, expecting fixed responses deemed correct for exercises such as translating, sentence restructuring or fill-the-gap. Nevertheless, the Andy chatbot, while technically giving the user freedom, is limited to English and often fails to process the text provided by users. HelloTalk and Tandem are essentially chatting apps with added useful features such as correction tools. This divides the domain in applications where you either have no interaction with other interlocutors at all, or applications for advanced communication with full conversations, unguided and unstructured.

4 Polygloss, a conversational agent for language practice

The Polygloss application is an adaptation of the PhotoBook task (Haber et al., 2019) for the domain of Computer Assisted Language Learning (CALL). The PhotoBook task was created to study how people build and accumulate common ground through crowd-sourced visually-grounded dialogue. It ran on Amazon Mechanical Turk and consisted of displaying 6 images to the participants and letting them talk to each other until they figured out which images they had in common.

Our application draws on the design of this task and Freire's methodology (Freire, 1972), still using images to give users something to talk about, but making adjustments and simplifications to encourage language acquisition. The main difference is that, while PhotoBook collects data exclusively in English, Polygloss lets users pick any language they would like to learn. Another important difference is that our experiment did not run on Amazon Mechanical Turk, but rather as a free downloadable mobile application for the Android operating system, in order to capture users that are actively learning a language.

4.1 System architecture

Because of its turn-taking characteristic, mobile was picked as a more appropriate distribution platform for the application since it offers easy functions for notifying users. The programming language and development framework used were Dart and Flutter[2], instead of Java, due to their capability to compile for multiple operating systems. So far, the application can only run on phones running the Android operating system, but it could also be published to the iOS App Store in the future.

The game was built with a "serverless" architecture. The only code written for the game was the application installed in the users' phones, which acts a client application. It connects directly to the database that stores the game history and settings, the user authentication service, the image storage service, and the analytics service, provided by Google Cloud Platform[3], through HTTP requests to their API, as seen on Fig. 2. We do not have our own back-end, which significantly reduces maintenance work.

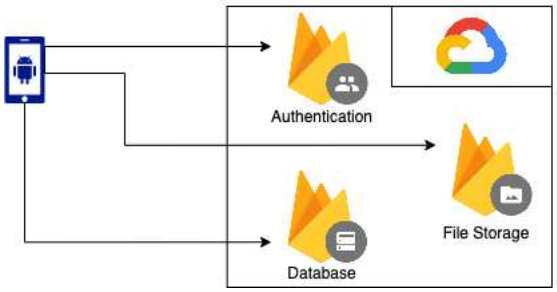

Figure 2: Architecture of API requests to Cloud Services

The game was then published to the Android Play Store, which allows an easy distribution of the software and its updates to the users, plus a set of useful development tools such as staged deployments, and collection of application feedback, statistics and errors.

[2]https://flutter.dev/
[3]https://cloud.google.com/

4.2 Design of the application

Since a match can be in any language, the user-pairing possibility is sparser: having two users willing to play in the same language at the same time is a rare event. In order to mitigate that, the matches have a shorter duration, using small texts instead of long dialogues, in comparison with the PhotoBook task, and are played asynchronously, i.e. in turns. The users get a notification on their mobile phone once they have a response or an invitation for a match and it is their turn. That means that Polygloss does not collect a dataset with the same utility as PhotoBook, since the samples produced are not full dialogues, but it still collects visually grounded short texts which could be useful for various goals such as image labelling or enriching word embeddings with more context. It also collects proportionally more learner language, which can be used to obtain insights into second language acquisition or to improve applications that fail to work with non-native speakers. Since the language option is open, there is also possibility for collecting native samples from various languages or dialects which are under resourced, such as South Tyrolean German.

The images used in Polygloss were sourced from a catalogue of illustrations[4], for which a license was purchased. In order to add an educational component, the images were manually curated to be simple, displaying usually one object or action, and were divided into categories such as "Hobbies", "Animals", "Emotions", defining the lessons, each containing between 10 and 60 images. There are 104 lessons in total, which increase in difficulty, for which the user has to collect "stars" to unlock and progress through, as seen on Fig. 3. The theme and order of the lessons was chosen based on common topics engaged by educational materials. The materials consulted were 7 different textbooks and websites destined to A1 - B2 students of German, Spanish, Greek and Brazilian Portuguese.

When the player finishes completing their profile, they are given a new match suggestion. For the new match, the app has to make three decisions: in which language will the next match be played, what lesson is selected, and who will be the opponent. The language is chosen among the ones that the player has declared in their profile, their native or mastered languages included. There is an 80% chance that the language selected will be a language that the user is interested in learning, and a 20% chance that it is a language that they already speak. The application then analyses the player's history of played lessons to decide what lesson will be started. Lessons where the player already has three stars or that require more stars to be unlocked than the player has currently accumulated are discarded. One lesson is then chosen randomly from the remaining ones. A series of queries and filtering is done to select the opponent. First they are filtered on basis of the language chosen. If the initial player is a learner, the selected opponent can be either another learner or a speaker. If the original player is a speaker, other speakers are discarded as opponents. Subsequent filtering is done to retain more active players and exclude players who have been blocked by the user.

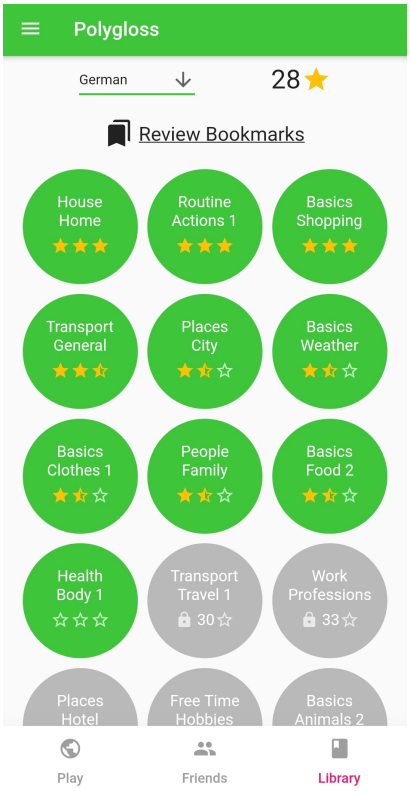

Figure 3: Screen showing Library tab containing the lesson tree

Once the match is started, the initial player is shown 4 random images picked from the collection of images in the lesson and one of them is selected. Then they are assigned with the task of helping their opponent guess which image is selected by writing a short text. One reason why

[4]https://www.flaticon.com

we display 4 images, instead of 6 used in PhotoBook, is that we felt it was still enough variety to offer context while having a more appropriate fit to the typical screen size mobile phones. After the initial player finishes their turn, a notification is sent to the opponent and then they can respond to the match and select the correct image, using the language toolbar if they wish, or not, as shown on Fig. 4. If they pick the correct one, they are awarded points which count towards their number of stars. The rounds are then reversed, and it is time for the opponent to be shown a selected image and help the initial player with a short text. In this way, both players have the opportunity to practice creative language output and are receiving input. Other features of Polygloss aimed for helping language students are present in the toolbar: tools to work with the text from their partners like Copy, Translate and Bookmark; and the possibility to give corrective feedback, allowing players to negotiate meaning and modify their round's text after the feedback received.

Figure 4: Screen showing second round of match, containing text, toolbar and image selection

5 Evaluating Polygloss for language learning

Chapelle (2001) states that CALL applications should be evaluated at three levels: the CALL software itself, the teacher-planned CALL activities, and the learners' performance during the CALL activities. She highlights 6 criteria of task appropriateness for language learning to be evaluated through a combination of judgemental and empirical methods: language learning potential, learner fit, meaning focus, authenticity, positive impact and practicality. We will discuss our results regarding these criteria throughout the next sections.

In order to investigate the efficacy of the system for language practicing, we used different automated techniques for scoring the text produced by the players, as well as conducted a user survey to inquire the players experiences during the interaction with the system.

To understand how the player is progressing over time, it is necessary to measure more than how often they succeed in the image guessing game, as improvements in conducting the task could mean simply that one got better in playing the game. It could mean getting accustomed with the user interface, observing what gives more points, or even cheating.

When applications like Duolingo, offer explicit or implicit knowledge following a certain curriculum, it is possible to test players against that knowledge and see how well they perform. Duolingo does so via checkpoint quizzes tagged with something they call "communicative component", which marks what a question tests. Doing so allows them to measure the players' evolution with very few questions.

However, during unstructured creative practice, the space of possibilities is too vast. Truly observing their language progress over time involves measuring their proficiency at different points in time. Without a progressive curriculum with which to reduce the scope of the test, we would have to conduct a thorough exam in each measuring point. That would be too labor intensive for us to prepare in the timeframe of this research, especially given the multilingual characteristic of the application. It would also overly disrupt the usual gameplay, adding long interruptions for the player. One alternative to circumvent this issue is to analyse intrinsic characteristics of the text produced by the players within the game itself, which is the ap-

proach we take in this study.

5.1 Text Quality

There are different ways in which the quality of an utterance by a non-native speaker could be measured. The Common European Framework of Reference for Languages (CEFR) states that communicative language competence can be divided into the following components: linguistic, sociolinguistic and pragmatic, each with several sub-components. The CAF framework of language proficiency (Housen et al., 2012) highlights *Complexity*, *Accuracy*, and *Fluency*. These main competences are further divided into several subcomponents and although organized in different structures, several sub-components have equivalents in both frameworks.

Given our study collects very short written texts on limited interactions, it is not possible to evaluate some of these dimensions of competence, such as phonological competence, and others would be extremely difficult to measure, such as sociocultural competence. However, in both language frameworks, each component or sub-component has its contribution to the overall language proficiency. Therefore, we will focus on a limited number of metrics, with the understanding that they contribute to the general communicative competence of the players.

Metrics of syntactic complexity have been used to indicate syntactic competence of the learners (Bhat and Yoon, 2014). As seen on Table 1, we selected 2 of such metrics to include in our study: mean length of text and mean depth of parse-tree of text. Additionally, we are using word-embeddings acceptability score as a third metric.

Word embeddings can be understood as a general class of techniques to represent the meaning of lexical units through dense vectors of real numbers. They are built with statistical or neural methods based on the co-occurrence of the units in a very large corpus of text. There is a vast range of methods used to build them, and variations on what is the base lexical unit: from character-level to whole document embeddings. These vectors have been used for language modelling (Mikolov et al., 2013), which means they are sensitive to collocations, a feature of advanced speech. One metric obtained from such models is the acceptability score, the ability of the model to predict a sample. In theory, this metric could be used to capture lexical competence and accuracy as samples with many words unrelated to each other, or containing orthographic mistakes, awkward word orders and other errors, would manifest as a lower score. At the time of writing, we were not aware of any other studies using this method specifically for measuring proficiency of language learners, but there is extensive research on how such models capture grammaticality (Lau et al., 2016) and they have been previously used to judge grammar acceptability (Warstadt et al., 2018). One caveat of this metric is that the use of extremely rare words could also result in a lower score. However, we suspect this limitation would be less significant in the context of learner language considering learners will often be using very common words and the not so frequent words they need to learn are still common enough to be well captured by such models.

Before evaluating the players over time, first we investigated how the metric themselves were behaving by dividing the sentences into groups according to the players self-declared proficiency and observing if there were any anomalies.

We used Jupyter notebooks[5] and Python 3.7 to measure the first metric, adding a Natural Language Processing library for Python called Spacy[6] for the second metric, and, for the last metric, a Machine Learning library called PyTorch[7] and XLM-RoBERTa (Lample and Conneau, 2019), a generic cross lingual sentence encoder pre-trained on 2.5T of data in 100 languages. To parse a sentence with Spacy, it is necessary to download a package for each language, which is why we restricted our dataset to the five most used ones.

5.2 User Survey

We prepared a questionnaire, sent to the players' email addresses, containing various questions regarding their interaction with the application. The main goal of the questionnaire was to tap into the perceived language learning usefulness and benefits of Polygloss. We also included questions related to its interface, user experience and entertainment value. Finally, in order to explore possible future improvements, we also had an open text field for any extra feedback or suggestions the players might have. The full questionnaire can be found on Appendix A.

[5] https://jupyter.org/
[6] https://spacy.io/
[7] https://pytorch.org/

	Metric	Evaluation
I	Text length	Syntactic Complexity
II	Depth of parse-tree	Syntactic Complexity
III	XLM-r acceptability score	Lexical Competence and Accuracy

Table 1: Metrics for text evaluation

6 Results

6.1 Polygloss in use

321 language learners from various backgrounds downloaded the application, created a profile, and played a match. During profile creation, they were asked to self-declare the proficiency level for all of the languages they speak, or are interested in learning, in 4 different levels: beginner, intermediate, advanced, and native. These levels were chosen because it was not expected from all of the players to be familiar with the CEFR (Common European Framework of Reference for Languages) scale of language proficiency (Council of Europe, 2001). Moreover, not all languages that could be declared are commonly measured according to this scale. Finally, learners' self-assessment of language skills accuracy can be considered significant (Liu and Brantmeier, 2019). We assume there is also little motivation for users to lie or exaggerate in the scenario of playing our game, unlike a scenario where, for example, one is applying for a job position that requires specific language skills when they are in dire need of a paid occupation.

The players' profiles declared over 80 different languages with various degrees of knowledge. The most popular languages were English, Spanish, French, Portuguese and German.

In a period of approximately six months, the players played 5460 matches, of which 1977 were played to completion, creating over 7000 samples of sentences or very short texts in over 40 languages and dialects. The top played languages were English, Spanish, French, German, Portuguese, Greek, Italian, Russian, Japanese and Dutch, in this order. Other languages had minor numbers, and there were very interesting samples collected, such as 35 samples in Esperanto, an artificially constructed language, and 97 samples in a dialect of German from the South Tyrol region of Italy. Because of practical reasons related to the libraries used in the evaluation, which we will discuss in the next section, we have limited our dataset to 5276 samples created in the 5 most used languages, as seen on Table 2. The full number of samples collected can be seen on Appendix B.

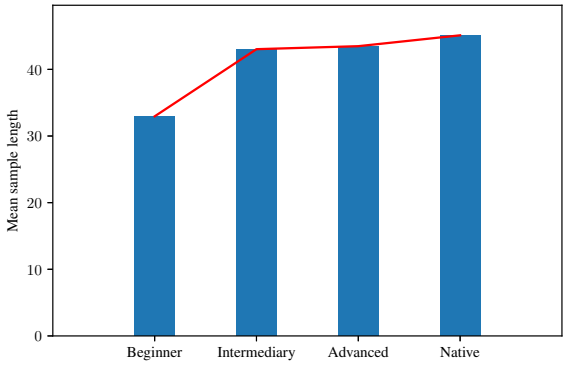

Figure 5: Mean length of sample per user group

6.2 Text quality among proficiency groups

6.2.1 Text length

During the initial study of the metrics, a difference of 10 characters was found between the mean length of samples produced by beginner and intermediary speakers, as seen on Fig. 5. A very small change was found between intermediary and advanced speakers, and only a slight difference of 2 characters was found in the mean between advanced and native speakers. Throughout the rest of this study, a Welch's unequal variances t-test is used to determine if differences among two groups are significant. In this case, apart from beginner ($M = 32.95$, $SD = 18.8$) versus intermediary ($M = 43.03$, $SD = 28.57$, $t(2864) = -11.38$, $p < 0.001$), they were not. Tests between the intermediary group and the advanced group ($M = 43.48$, $SD = 24.05$, $t(3130) = -0.47$, $p = 0.63$), or the advanced group versus native ($M = 45.11$, $SD = 30.12$, $t(2353) = -1.37$, $p = 0.17$) found no significant difference of player performance in text length.

One could argue that text length could vary according to the language, and the results could be different once breaking down. Indeed, a com-

Language	Beginner	Intermediary	Advanced	Native	Total
English	90	178	621	280	1173
Spanish	304	464	278	108	1168
German	309	454	90	173	1048
French	138	428	436	31	1035
Portuguese	349	152	31	307	852
Total	1190	1676	1456	899	5276

Table 2: Number of samples in selected languages

parison of similar levels in different languages showed they are very different, for example, beginner Spanish players (M = 38.14, SD = 22.56) wrote longer texts than beginner English players (M = 26.31, SD = 11.06, t(392) = 6.76, $p < 0.001$). After comparing each pairing of adjacent levels within each language, the full breakdown can be found on Appendix C, a wide variety of patterns emerged. Only in Portuguese could all levels be reasonably distinguished from each other, but even then, the group of intermediate speakers (M = 38.65, SD = 20.99) performed significantly better than the advanced speakers (M = 31.61, SD = 13.44), t(181) = 2.38, $p < 0.05$).

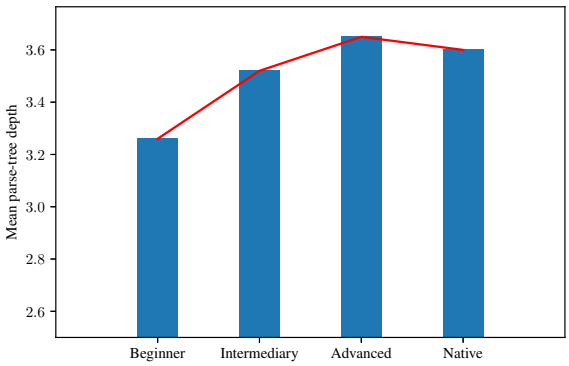

Figure 6: Mean depth of sample's parse-tree per user group

6.2.2 Depth of parse-tree

For the second metric, again the biggest difference among the proficiency groups in selected samples is between beginner and intermediary speakers, consisting of 0.26 levels in the depth of the parse tree of the samples, as seen on Fig. 6. One could make the same argument regarding differences between languages here. After comparing similar levels between languages, a consistent behaviour was not found, beginners did not vary between most language pairs, but subsequent levels often varied. Within each language sampled, in none of them it was possible to determine significant gaps between every level.

6.2.3 XLM-r acceptability score

For the last metric, before analysing the groups, a preliminary test with some example sentences, shown on Table 3, was done to observe if the scores seemed acceptable. It did not behave as expected in all instances. The example in Spanish where we compared a correct sentence and a sentence containing a grammatical error showed the error sentence as having a higher score than the correct one. The correct Spanish sentence also obtained a much lower score than the other correct examples. It is not possible to inspect the reason in detail, but most of the examples seemed to obtain reasonable results.

Once we scored and averaged the samples in all groups, differently from the previous metrics, the biggest interval was between the intermediary and the advanced speakers, being 4.32%, as seen on Fig. 7. Given the multilingual nature of this metric, we expected no significant differences once further breaking down the groups by language, and indeed, at beginner and intermediate levels no significant difference was found between any of the language pairs. However, at subsequent levels some differences emerged, especially with advanced speakers of English, who performed better than most other groups. Overall, the difference in performance between beginner (M = 81.31, SD = 32.27) and intermediate (M = 84.08, SD = 29.72, t(2864)= -2.34, $p < 0.05$) players was significant, the difference between intermediate and advanced (M = 88.4, SD = 25.83, t(3130) = -4.34, $p < 0.001$) speakers as well, and no significant difference was found between the advanced and the native players (M= 89.62, SD = 24.19, t(2353) = -1.15, p =

Sentence	Score
This is a good sentence	99.49
This is a sentence good	4.81
C'est une bonne phrase	99.87
C'est une bone phrase	72.89
Das ist ein guter Satz	99.8
Das ist ein guter satz	98.37
Esta é uma boa frase	99.86
Esta são uma boa frase	32.85
Esta es una buena frase	83.79
Esta es un bueno frase	85.76
This is a good sentence	99.49
This is shorter	99.42
This is also a good sentence but longer	99.69
This is a wet tissue	99.85
This is a wet sentence	81.81
This is a potato sentence	30.53

Table 3: Acceptability scores on XLM-r encoders of example sentences

0.24).

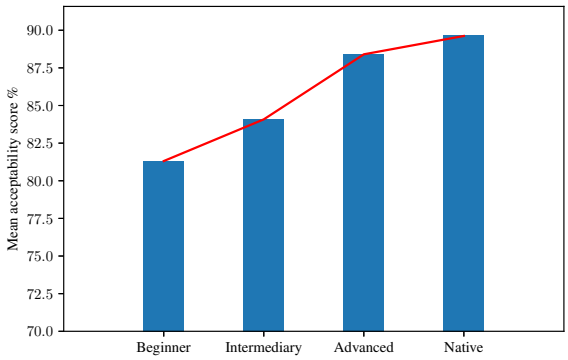

Figure 7: Mean XLM-r acceptability score per user group

6.3 Text quality over time of application usage

The average number of rounds played per player per language was 36. We used this number to divide each group of samples into two further groups, those created until the 36^{th} round, and samples produced after that.

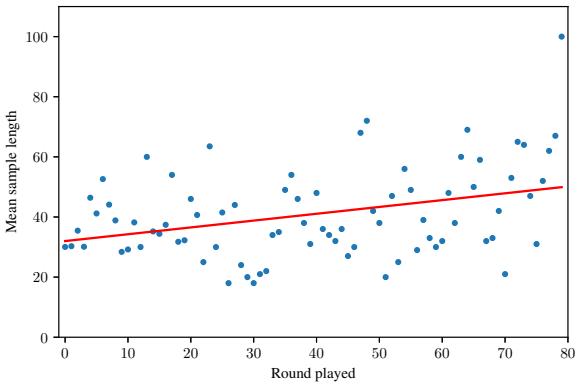

Figure 8: Mean length of beginner Spanish speakers samples across rounds played

Group	Round $\leq$ 36 Mean (SD)	Round $>$ 36 Mean (SD)	p-value
Beginner ES	35.8 (22.59)	48.71 (19.14)	< 0.001
Beginner PT	28.84 (12.22)	35.31 (18.82)	< 0.001
Intermediate ES	36.59 (18.15)	42.54 (16.21)	< 0.001
Intermediate DE	37.3 (23.33)	45.0 (21.59)	< 0.001
Advanced FR	42.68 (21.14)	56.14 (23.89)	< 0.001

Table 4: Sample length by user group until and after 36 rounds

6.3.1 Text length

Once separating the samples further down by language, many groups did not have enough samples for confident results. For example, there were no samples at all produced by English or French beginner speakers after the 36^{th} round and altering the threshold for breaking the groups into before and after to the mean of rounds played for that group did not alter the outcome. For certain groups it was possible to observe significant improvements, as seen on Table 4. A positive trend for Spanish beginner players can be seen on Fig. 8.

6.3.2 Depth of parse-tree

For the second metric, as seen on Table 5, in a breakdown per group level, none presented significant progress, and in a further breakdown per language, only beginner Spanish and advanced French players presented significant improvement.

6.3.3 XLM-r acceptability score

For the third metric, the trend seen in the plot in Fig. 9 shows an overall improvement in average acceptability score. However, as seen on Table 6, beginner and advanced players did not present

Group	Round ≤ 36 Mean (SD)	Round > 36 Mean (SD)	p-value
Beginner All	3.26 (0.89)	3.28 (0.79)	0.64
Intermediate All	3.5 (0.97)	3.58 (0.99)	0.12
Advanced All	3.65 (1.04)	3.64 (1.05)	0.92
Beginner ES	3.26 (0.91)	3.85 (1.0)	< 0.001
Advanced FR	3.35 (0.82)	3.65 (0.95)	< 0.001

Table 5: Depth of parse-tree of sample by user group until and after 36 rounds

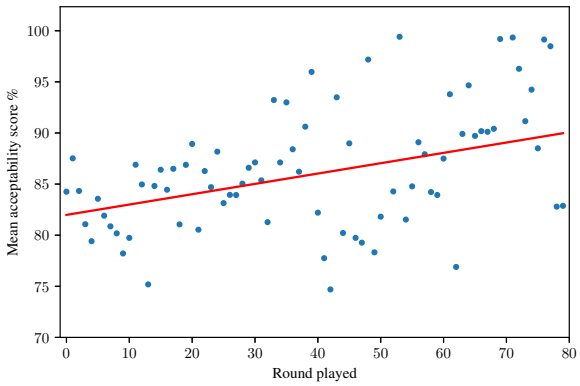

Figure 9: Mean XLM-r acceptability score of non-native speaker samples across rounds played

a significant progress. Meanwhile, intermediary players obtained an average gain of 4.6% in their XLM-r scores on later rounds.

6.4 User Survey

We sent the user questionnaire to all 321 active players, obtaining 61 responses. Many of our questions were formatted as a 1-5 scale where 1-2 is a negative response, 3 is considered neutral, and 4-5 are positive and very positive responses. Based on this, 83.6% respondents indicated that they would recommend Polygloss to a friend learning a language, 77% that it is easy to play, 83% that the game instructions are clear, 75.4% said playing Polygloss is a practical activity for learning, 73.6% that it is useful for learning a language, and 88.6% that it is fun to play. We further divided questions related to usefulness, and the results can be seen on Table 7.

Group	Round ≤ 36 Mean (SD)	Round > 36 Mean (SD)	p-value
Beginner	81.07 (32.41)	82.17(31.73)	0.62
Intermediate	82.53 (31.0)	87.13(26.78)	< 0.01
Advanced	88.09 (25.78)	89.03(25.92)	0.51

Table 6: Mean XLM-r acceptability score across time by user group (%)

In the open fields, the more complimented areas of the game were interacting with other people and the ability to be creative. The most criticized aspects were how difficult the game is for absolute beginners in a language, and the points system, for, despite being intuitive, not giving a sense of progress in the language. Common feature requests were: more ways to track progress, such as streaks, audio matches, and sentence and word examples.

7 Discussion

The division of the mean sample length metric by players' proficiency group suggests that although beginners write shorter texts in comparison with other groups, there is not a significant difference among the other groups. Once breaking the groups down into each language, it is even more difficult to tell. This is an obstacle for comparing progress among them, making it difficult to evaluate learner fit. It is still, however, a positive surprise to see a considerable progress for beginner Spanish and Portuguese and advanced French groups, given they were not part of the intended audience.

The depth of the parse-tree metric behaved differently from our expectations as it presented an odd peak at advanced players, above native speakers. This could mean that this is a bad metric, but one possible explanation for this behavior is that advanced players could be making more effort to write elaborate sentences in order to practice, while native speakers do not bother. Either way, this metric also does not help with evaluating learner fit, especially when breaking the groups down once more for each language. Across time, we understand the lack of significant progress in many groups as a sign that the game is not giving enough incentive to write more elaborately. This is also backed by our user survey, where users point grammar as the aspect with which they thought Polygloss is least helpful.

Indeed, the game does not draw explicit attention to form, which is one of the factors necessary for what Chapelle (2001) calls of *language learning potential*, which further explains these results. If a player writes a sentence containing a grammar mistake, the system does not provide a correction before progressing to the next

	Useful (%)	Neutral (%)	Not useful (%)
Expressing yourself better	62.75	23.53	13.72
Being more confident	55.77	23.08	21.15
Learning real-world sentences	46.15	34.62	19.23
Becoming more proficient	45.76	32.20	22.04
Learning new words	45.26	24.59	31.15
Learning spelling	40.00	40.00	20.00
Learning grammar	13.33	31.67	55.00

Table 7: How do you think Polygloss was useful with...?

round. However, when Chapelle says *language learning potential* in a CALL system is characterized by its difference from being simply an opportunity for language use, we would need to assume that simply using language does not lead to learning gains or we would need to restrict ourselves to a very narrow definition of *use*. In practice, there are many benefits brought by collaborative aspects that arise from language usage (Swain, 2006). Chapelle does include characteristics such as interactional modification and modification of output, which are, in essence, processes that derive from collaborative use. Although this could sound unclear, she does go on to elaborate that the exact meaning of this criterion will evolve as second language acquisition research continues to develop. Given that, it seems that our system does implement then, a partial attention to form, as it allows players to send each other feedback and modify their output. However, like on a real-world interaction, not all mistakes elicit feedback. In Polygloss, only 4.5% of the samples studied received some feedback and, in fact, the user survey also received mentions of it not being enough. One possible explanation for this is that players might be correcting others only when mistakes damage comprehensibility and are an obstacle to the task at hand. Nevertheless, some subgroups like beginner Spanish and Advanced French did show improvements in parse-tree depth across number of rounds played, and intermediate samples showed improvements in the XLM-r scores, which also captures some form.

Results from the XLM-r acceptability score metric showed it to be best suited metric for evaluating learner fit. Given we had no record of it being used in this way before, we are satisfied with how it performed. We understand that grouping languages together also facilitated interpreting the results, given our number of samples. Even if there is not a clear difference in score between advanced and native groups, the difference is clear among the other groups. One factor that could have impacted this is that we did not separate the samples from advanced speakers who were actively learning the language from the ones from players who registered it as a language they spoke, but were using the game to learn another language. For example, one could speak Portuguese at native level, speak advanced English, and be currently learning French, which they also self-evaluated at advanced level. Indeed, only 22% of the players who evaluated their English as advanced had English listed in the languages they wanted to learn, compared with 62% of the advanced French and 68% of the advanced German player groups.

The improvement observed for intermediate level players over number of rounds played is further backed by the user questionnaire, where users indicated that the game is too difficult for beginners. This result validates our intended proposal, since intermediary level speakers were our target audience for this game.

Authenticity, as Chapelle (2001) explains, is the level of correspondence between a language learning task and a task the learner can encounter in the real world, outside the learning environment. The user survey shows good results in this area, as 63.7% of players thought the game helps expressing yourself better and 46.15% thought it helps learning real-world sentences, which are important for authenticity and pragmatics. We observed usage of authenticity when players produced sentences like the one below, where they use the image provided to successfully practice discussing current world events, such as the Covid-19 pandemic, even if the image does not necessarily draw attention to the subject.

The opportunity to make such outputs is allowed by the flexibility to write creatively pro-

”Sie sollten ihre Hände waschen”
(They should wash their hands)

vided by the game's design informed on critical pedagogy. This is also particularly convenient because it does not require frequents updates to the game's content to introduce current discussion topics.

8 Future work

Even though the results are positive and the application was perceived as fun to play, practical and useful by the majority of the players, there are many avenues for future work. The first one is to modify the game to draw more attention to form, add more interaction and collaborative features, encourage players to use the feedback feature more often, and reevaluate the performance on syntactic complexity metrics. Another possible route is to implement word tips and sentence examples and reevaluate the performance of learners on lexical competence metrics. This could be done using the data collected from other players on previous matches and the users own accumulated vocabulary to expand on topics that are interesting to them. Lastly, one other possibility is to allow the matches to be played with audio, and conduct a fluency and phonology based analysis.

9 Conclusion

It is hard to find appropriate learning materials for learners looking to overcome the intermediate plateau. At this stage, it is important to employ a mix of techniques, not abandon active study and produce language using your own words. Our proposed visually-grounded task has proven to be an effective way of doing that. We developed a learning game made available in a practical way as a mobile application, playable at any time of the day, and, given the existence of available partners, sufficiently generic to be playable in any chosen language. Even though more attention could be drawn to form, it draws sufficient attention to meaning, offering creative freedom and opportunity for authenticity. It provides positive impact beyond meaning and form as players feel it helps them express themselves better. Both the quantitative and qualitative results in this study confirm the intended fit of this task for intermediary level language learners and reveal a possible fit for other groups that could be explored in future research. In addition to this primary contribution, a second contribution is the serviceable use of transformers' acceptability score as an evaluation metric. Finally, we would like to join Benson (2013) in his call to have autonomy as an explicit goal in CALL, and highlight the importance of socially informed design for the development of successful language learning applications.

Acknowledgments

This research was supported in part by the DALI project, ERC Grant 695662, in part by the EPSRC CDT in Intelligent Games and Game Intelligence (IGGI), EP/L015846/1.

References

Uju Anya. 2016. *Racialized Identities in Second Language Learning: Speaking Blackness in Brazil*. Routledge Advances in Second Language Studies. Taylor & Francis.

Phil Benson. 2013. *Teaching and Researching: Autonomy in Language Learning*. Applied Linguistics in Action. Taylor & Francis.

Suma Bhat and Su-Youn Yoon. 2014. Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication*, 67.

Serge Bibauw, Thomas François, and Piet Desmet. 2019. Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based call. *Computer Assisted Language Learning*, pages 1–51.

Robert Blake. 2005. Bimodal CMC: The glue of language learning at a distance. *CALICO Journal*, 22:497–511.

Carol A. Chapelle. 2001. *Computer Applications in Second Language Acquisition*. Cambridge Applied Linguistics. Cambridge University Press.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Paulo Freire. 1967. *Educação como prática da liberdade*. Série Ecumenismo e humanismo. Paz e Terra.

Paulo Freire. 1972. *Pedagogia do oprimido*. Paz e Terra.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The photobook dataset: Building common ground through visually-grounded dialogue.

Sylvain Hatier, Arnaud Bey, and Mathieu Loiseau. 2019. Formalism for a language agnostic language learning game and productive grid generation. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, Turku, Finland. LiU Electronic Press.

Alexis Housen, Folkert Kuiken, and Ineke Vedder. 2012. *Complexity, accuracy and fluency: Definitions, measurement and research*, Language Learning amp; Language Teaching, pages 1–20. John Benjamins Publishing Company, Netherlands.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2017. Revita: a system for language learning and supporting endangered languages. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 27–35, Gothenburg, Sweden. LiU Electronic Press.

Stephen Krashen and Tracy Terrell. 1983. *The Natural Approach: Language Acquisition in the Classroom*. Language Teaching Methodology Series. Pergamon Press.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Jey Lau, Alexander Clark, and Shalom Lappin. 2016. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41.

Hsueh Liu, Karey Lan, and John Jenkins. 2014. Technology-enhanced strategy use for second language vocabulary acquisition. *English Teaching Learning*, 38:105–132.

Huan Liu and Cindy Brantmeier. 2019. "I know english": Self-assessment of foreign language reading and writing abilities among young chinese learners of english. *System*, 80:60–72.

Qing Ma and Peter Kelly. 2006. Computer assisted vocabulary learning: Design and evaluation. *Computer Assisted Language Learning*, 19.

Ernesto Macaro, Zoe Handley, and Catherine Walter. 2012. A systematic review of call in english as a second language: Focus on primary and secondary education. *Language Teaching*, 45.

Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Paul Nation and Susan Hunston. 2013. *Learning Vocabulary in Another Language*, 2 edition. Cambridge Applied Linguistics. Cambridge University Press.

Emmanuel Rayner, Pierrette Bouillon, Nikolaos Tsourakis, Johanna Gerlach, Yukie Nakao, and Claudia Baur. 2010. *A Multilingual CALL Game Based on Speech Translation*, Proceedings of LREC. ID: unige:14926.

Frank E. Ritter and Lael J. Schooler. 2001. The learning curve. *International Encyclopedia of the Social Behavioral Sciences*, pages 8602–8605.

Scott Saft, Yumiko Ohara, and Graham Crookes. 2001. Toward a feminist critical pedagogy in a beginning japanese-as-a-foreign-language class. *Japanese Language and Literature*, 35:105–133.

Merrill Swain. 1985. Communicative competence: Some roles of comprehensible input and comprehensible output in its development. *Input in second language acquisition*, 15:165–179.

Merrill Swain. 2006. *Languaging, agency and collaboration in advanced second language proficiency*, pages 95–108. London: Continuum.

Lev S Vygotsky. 1978. Mind in society (m. cole, v. john-steiner, s. scribner, & e. souberman, eds.).

Mark Warschauer. 1996. Computer assisted language learning: an introduction. *Fotos S. (ed.) Multimedia language teaching, Tokyo: Logos International*, pages 3–20.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *CoRR*, abs/1805.12471.

Appendices

A User Survey

1. Would you recommend Polygloss to a friend learning a language?
2. Have you been playing Polygloss recently?
3. If you answered "no" to the previous question: Why? Is there anything that would have made you play it more?
4. Do you think Polygloss is easy to play?
5. Do you think the instructions and the tasks you need to do in the game are clear?
6. Do you think playing Polygloss is a practical way to advance your language progress?
7. Do you think Polygloss is useful for learning a language?
8. Do you think Polygloss is fun to play?

9. How do you think polygloss was useful with...? [Learning new words] [Being more communicative] [Being more fluent] [Learning grammar] [Learning spelling] [Learning idiomatic expressions] [Becoming more proficient in the language] [Expressing yourself better] [Being more confident in the language] [Learning sentences you can use in the real world]

10. Is there a feature you would like to see in Polygloss?

11. What would you love to see more often in language learning app?

12. What do you think of Polygloss' interface?

13. How could Polygloss be even better? Do you have any questions about Poygloss?

14. Is there any additional feedback on Polygloss, ideas, anything else you would like to say?

B All samples collected, grouped by language

Language	Beginner	Intermediate	Advanced	Native	Total
English	90	178	621	280	1173
Spanish	304	464	278	108	1168
German	309	454	90	173	1048
French	138	428	436	31	1035
Portuguese	349	152	31	307	852
Greek	259	33	0	129	421
Italian	93	157	54	30	340
Russian	97	41	6	2	148
Japanese	44	19	64	0	128
Dutch	92	18	8	1	120
Swedish	86	10	0	1	97
South Tyrolean	54	0	0	43	97
Polish	21	20	0	19	71
Hungarian	22	37	0	2	61
Indonesian	24	0	0	22	46
Finnish	28	6	0	0	39
Arabic	19	16	0	0	35
Esperanto	26	9	0	0	35
Persian	13	13	0	1	32
Norwegian	17	0	0	1	26
Korean	22	3	0	0	25
Danish	19	0	5	0	24
Mandarin	5	4	4	0	19
Turkish	6	7	0	5	18
Catalan	4	4	4	1	13
Vietnamese	5	0	0	0	9
Croatian	8	0	0	0	8
Javanese	4	0	0	4	8
Hebrew	0	3	0	4	8
Romanian	1	6	0	0	7
Estonian	3	4	0	0	7
Ukrainian	5	0	1	0	6
Slovak	4	0	0	2	6
Afrikaans	2	4	0	0	6
Georgian	5	0	0	0	6
Thai	2	0	0	0	5
Czech	1	0	0	3	4
Bulgarian	4	0	0	0	4
Latin	1	2	0	0	3
Hindi	1	0	0	0	3
Sprok	2	0	1	0	3
Irish	2	0	0	0	2
Scottish Gaelic	1	0	1	0	2
Breton	1	0	0	0	1
Tagalog	1	0	0	0	1
Icelandic	0	0	0	0	1
Serbian	1	0	0	0	1
Total	2195	2092	1604	1169	7060

C Text length performance comparison of proficiency levels broken down by language

Language	Group 1 Mean (SD)	Group 2 Mean (SD)	Welch's unequal variances t-test
English	Beginner 26.31 (11.12)	Intermediary 48.21 (37.54)	t(266) = -7.18, $p < 0.001$
	Intermediary 48.21 (37.54)	Advanced 39.87 (22.41)	t(797) = 2.82, $p < 0.01$
	Advanced 39.87 (22.41)	Native 39.55 (20.51)	t(899) = 0.21, p = 0.8346
Spanish	Beginner 38.14 (22.6)	Intermediary 39.08 (17.63)	t(766) = -0.61, p = 0.5412
	Intermediary 39.08 (17.63)	Advanced 44.38 (25.66)	t(740) = -3.05, $p < 0.01$
	Advanced 44.38 (25.66)	Native 36.31 (23.44)	t(384) = 2.96, $p < 0.01$
German	Beginner 29.49 (12.46)	Intermediary 41.7 (22.7)	t(761) = -9.54, $p < 0.001$
	Intermediary 41.7 (22.7)	Advanced 42.16 (29.16)	t(542) = -0.14, p = 0.8879
	Advanced 42.16 (29.16)	Native 52.97 (32.78)	t(261) = -2.73, $p < 0.01$
French	Beginner 35.62 (26.73)	Intermediary 48.14 (38.95)	t(564) = -4.24, $p < 0.001$
	Intermediary 48.14 (38.95)	Advanced 49.17 (23.52)	t(862) = -0.47, p = 0.6413
	Advanced 49.17 (23.52)	Native 52.32 (36.02)	t(465) = -0.48, p = 0.6339
Portuguese	Beginner 32.14 (16.28)	Intermediary 38.65 (20.99)	t(499) = -3.41, $p < 0.001$
	Intermediary 38.65 (20.99)	Advanced 31.61 (13.44)	t(181) = 2.38, p = 0.0202
	Advanced 31.61 (13.44)	Native 48.12 (35.42)	t(336) = -5.24, $p < 0.001$