Design and empirical validation of effectiveness of LANGA, an online game-based platform for second language learning

Francesco Usai, Kiera G.R. O’Neil, and Aaron J. Newman

Abstract—Computer and smartphone-based applications for second language (L2) learning have become popular tools, being integrated in many classroom-based courses and adopted by the public at large. Yet, despite a significant body of research suggests that individuals differ in their ability to learn L2, it is still unclear what factors predict successful L2 acquisition and how L2 teaching software can be designed to adapt to individuals’ strengths and weaknesses. Here, we describe the architecture of LANGA, an online game-based platform under development for L2 teaching and research, and present a demonstrative proof-of-concept study using the platform. LANGA is designed to be both an effective and engaging product from the consumer perspective, and a tool that can be used by researchers to easily implement, deploy and test different training modalities for L2 teaching. Furthermore, key features of LANGA include easy configuration of training via modular design; emphasis on gamified teaching methods; and the use of automated speech recognition to provide learners feedback on verbal production. A first prototype of LANGA was tested in a small-scale, proof-of-concept study. Changes in proficiency from pre- to post-training were measured using recall and recognition tests, while event-related brain potentials (ERPs) were used to assess changes in brain activity related to lexical access over the course of learning. The results provided initial validation of the platform: participants were able to learn a large proportion of the words taught, and retained the novel words in a two-weeks follow-up. Future directions on the development of the platform are discussed.

Index Terms—game-based L2 acquisition, CALL, EEG, ERP

1 INTRODUCTION

The advent of personal computers, smartphones and tablets, has provided opportunities to make educational content more accessible and effective. In the context of second-language (L2) teaching, Computer-Assisted Language Learning (CALL) has emerged as a field of research devoted to exploring how existing or novel technologies can be used to make L2 teaching more effective [1]. Currently, the most popular CALL applications include online courses, distance and blended learning [2], immersive scenarios such as those provided by video games that were either designed explicitly for L2 teaching, or repurposed from games developed for other purposes [3]. Nowadays, there is an enormous interest in digital platforms for L2 teaching not only from academic institutions (which are progressively integrating these platforms into traditional programs [4]) but, thanks to the large-scale diffusion of portable devices, L2 apps are now extremely popular among the public at large. For example as of March, 2016 the L2 app Duolingo reported having 100 million users, and was the most-downloaded educational app for the iOS platform [5].

Many L2 apps feature recent technological advances such as animation, and some use game mechanics to make interaction fun and engaging. However, these products differ widely in terms of teaching philosophy, as reflected in how the material is delivered and how the learner interacts with the software. Despite the growing popularity of these applications there are still many unresolved issues concerning their theoretical foundations and, most importantly, their efficacy – issues shared with related domains, such as educational apps more broadly, cognitive training, and rehabilitation (e.g. [6], [7], [8], [9]). Products in the education market are not always designed and validated using scientific evidence; further, there are no accepted standards for empirical evidence of efficacy as there are for drugs or medical devices [10], [11], [12]. In most cases there are no empirical studies available to back up the claims made by the developers of educational apps [6]. Thus even when research – either independent or conducted in-house by developers [13], [14] – is available, it may not be sufficient to fully address questions of efficacy.

In this regard, Grgurovic and colleagues [15] conducted a meta-analysis to assess if use of technology for L2 teaching was associated with improvements in learning outcomes, and whether individual differences (such as baseline proficiency levels), types of technologies used, or training methods (i.e. intensity, types of instructions given etc.) had an influence on learning outcomes. Results indicated that use of technology, either alone or integrated in the classroom context, was at least as effective as traditional classroom-based learning.

Nonetheless, the authors reported the fact that some of the studies presented issues in design, as in some cases it was not possible to discern whether group differences in learning outcomes where due to training effects or to baseline differences in proficiency. The latter aspect is of particular relevance in this context, since there is...
mounting evidence that individuals’ ability to learn a new language is influenced by a multitude of factors [17], [18]. For example, people receiving the same training and amount of exposure to a given language might differ significantly in how fast they learn, how long they retain the material, and how well they respond to a specific type of training. Cognitive neuroscience research broadly suggests that individual differences (IDs) in socioeconomic status, level of education, age, lifestyle, motivation, engagement with the task [17], and cognitive skills (e.g., working memory span, executive attention, verbal memory, etc.) predict the extent to which people improve as a result of training in many domains. One of the ultimate goals of research in second language acquisition is to develop accurate models of how IDs can be used to predict optimal ways to learn a second language customized for each individual in terms of training strategies, intensity, types of content etc. This would allow us to move from the “one-size fits all” approach generally employed in technology-based approaches, to a personalized one that optimally adapts to the unique strengths and weaknesses of each individual. Such an approach requires large-scale studies (hundreds or thousands of subjects) involving people with diverse backgrounds and cognitive skills — two conditions that are difficult to meet by traditional lab-based research. Web-based and mobile applications represent an ideal solution to this problem since they can reach virtually anyone with an internet connection, remove physical barriers due to transportation, and speed up the research process.

These considerations have inspired our approach to the design of LANGA (LANguage GAming), a language learning platform founded on five principles:

- Use of compelling video-games that make training interesting and engaging
- Use of an advanced speech recognition engine (SRE) to provide real-time feedback about pronunciation
- Serve as a powerful research tool that, operating through the web, can reach a vast and heterogeneous population that cannot be accessed with traditional lab-based research. Web-based and mobile applications represent an ideal solution to this problem since they can reach virtually anyone with an internet connection, remove physical barriers due to transportation, and speed up the research process.

As described in detail in the following section, the design principles outlined above align with the definition of mini-games for L2 teaching provided by Cornillie and Desmet in their recent review [16]. In this work, the authors define L2 teaching mini-games as tools explicitly designed to improve mastery of L2 vocabulary through explicit and fast-paced practice, wherein learning is guided by the provision of immediate feedback on the accuracy of word production. Importantly, in contrast with traditional classroom-based L2 learning, software-based training offers the opportunity to record individual’s performance, allowing teachers to track their progress more accurately and to make timely decisions about interventions that might optimize learning. Nonetheless, as the authors point out, issues related to design and test of efficacy of mini-games through empirical studies are still relatively unexplored compared to other technology-based L2 teaching applications, a gap that we aim contributing to resolve with through LANGA. LANGA is the result of a collaboration between our laboratory and a private company, Copernicus Studios Inc. (Halifax, NS, Canada). We have thus combined the strengths of our lab in empirical behavioural and cognitive neuroscientific work on language processing with the company’s strengths in animation and game design. LANGA is currently in the pre-release, alpha stage, with anticipated public beta in early 2018. Our goal is to produce a language learning software platform that is effective and enjoyable for learners, but that also houses a flexible “back end” that will allow teachers and researchers to customize the learning process for groups of people, and empirically test hypotheses concerning language learning to generate better understanding of the mechanisms leading to successful L2 acquisition. In the remainder of this paper we discuss the design principles and architecture of LANGA, and provide details of a proof-of-concept study designed to demonstrate the feasibility of using LANGA as a research platform.

2 The Architecture

In this section we describe the high-level organization of LANGA, highlighting features that underlie its utility as a tool for CALL research. LANGA is implemented as a learning management system (LMS) — a platform that delivers training and tracks user progress. The LMS application provides the functionality for associated components (implemented as applications) to communicate, store, and receive data online. Applications access the underlying database through a representational state transfer application programming interface (REST API), which contains the commonly used collections in LMS apps (words, activities, lessons, and curricula), and an interface to generate collections to store additional data as needed by the app. In addition to words, activities, lessons, and curricula, the LMS comprises a user database, an authoring tool, and the SRE. In what follows, we refer to the two general classes of LANGA users: learners and designers, with the latter comprising developers, teachers, and researchers.

2.1 Words

The LANGA LMS houses a database of training items, or “words”. An instance of a word links one or more media items, including pictures (in common image formats such as JPEG or GIF), sound files (in WAV format), written forms (as Unicode text), phonetic encodings (implementing International Phonetic Association symbols as Unicode text), and grammatical properties (e.g., part of speech; grammatical gender) — all keyed by a unique ID code (which could be the English word for the item, an arbitrary code, or anything else representable as text). Critically, in LANGA a “word” can actually be more than one word; these teaching units can be phrases or even sentences, with
Figure 1. Illustration of the four games used to teach second-language vocabulary. Teaching game (top-left); Missing game (top-right); Double game (bottom-left); Crosswords (bottom-right)

the only caveat being that multi-word “words” are still treated as single units (so there would need to be separate “words” for dog and the dog if both forms were going to be taught). Each media item (other than pictures) is associated with a “language” tag, indicating what language it is used with. In this way, LANGA is highly flexible and modular with respect to implementing new training languages: if one wishes to teach a language not already present in LANGA, one can simply upload WAV files, text, and phonetic encodings for items already in the item database, and these will be associated with the same pictures. Additional items can also be added by uploading new pictures along with the other necessary media.

2.2 Activities

In the LANGA LMS, “activities” comprise games, tutorials, and assessments. These provide a modular structure that can be used for teaching L2 content, “meta-teaching” (how to play games, or other informational content), and also for assessing L2 knowledge or administering other tests (e.g., if a researcher wanted to measure a learner’s working memory capacity). LANGA has been designed to be flexible enough to accommodate games written in several languages/environments, including Flash/ActionScript, HTML5, and Unity. In designing curricula, researchers and teachers have access to a range of activities, although these are currently limited to those designed and built by Copernicus Studios; in the future it may be possible for designers to implement their own games using a standardized API. The “casual learning” design principle of LANGA is to employ game mechanics that have already been established to appeal to a large number of people, such as those implemented in popular casual gaming apps. For example, one current game implements a “match-3” game in which learners are presented with an array of tiles with pictures of vocabulary items, and tiles need to be manipulated to place three identical tiles in a row. All games employ pictures to represent semantic concepts (i.e., the vocabulary items to be learned), and combine these with spoken and written examples in the target language being learned. Examples of currently-implemented games are shown in Figure 1.

Figure 2. Timeline of the general training plan (top panel) and of the daily plan (bottom panel)

2.3 Lessons

Activities can be combined into lessons, which are intended to encapsulate a single learning session; i.e., a set of activities lasting 5-30 min that a learner would work through in one sitting. In addition to referencing a set of activities to include, a lesson can contain prerequisites that help control which lessons are available to a learner at any point in time. This can be useful if the curriculum designer wishes to allow the learner to choose from a set of lessons (rather than progressing through a set of lessons in a single, linear order), but at the same time ensure that the learner can only choose from lessons appropriate to their level, or other criteria the curriculum designer chooses.

2.4 Curricula

A curriculum brings together a set of lessons, and is defined by the designer as a coherent set of activities, engaged in over a number of lessons, to achieve particular learning goals. As such, curricula are the level at which a set of words/activities/lessons are assigned to a learner or group of learners. See Figure 2, top panel, for an example of the curriculum we used in one experiment [6]). The ability to customize curricula allows for systematic manipulation of training parameters to enable empirical testing of the efficacy of different variables in L2 teaching. Different users can be assigned different curricula, enabling A/B testing or more complex between-subjects research designs. The architecture will also allow users to select different training modules (e.g., different themes, topics, or sets of words) within a curriculum assigned to them, enabling user-specific customization and user-driven learning, while still allowing systematic empirical testing of the curriculum.

2.5 Speech Recognition Engine (SRE)

One of LANGA’s central design principles is to support the development of L2 production skills, as well as receptive (comprehension) skills. As such, automated speech recognition is fundamental to LANGA, and an API has been developed to enable integration of the SRE with the Unity game-building platform. In games that use speech recognition, LANGA’s SRE matches a word or sentence spoken by the learner against its “grammar” (set of possible recognition targets; see below), which
consists of phonetic encodings of (minimally) each of the words being taught in the lesson. The SRE returns the word that it matches the input to in the grammar (i.e., its “guess” as to which word was spoken out of the possible options), as well as a confidence score for that guess. If the SRE’s guess matches the target word and the confidence score is over a threshold level (determined empirically through pilot testing), the learner receives feedback that their pronunciation was correct; otherwise they receive an error signal. In early stages of development we employed a commercially-available SRE, and tested several freely-available SREs. However, Copernicus Studios is currently developing a dedicated SRE for LANGA. This is because other available SREs evidenced significant performance issues, and lacked the ability to customize features important for L2 teaching. For example, SREs are generally designed to be robust to different speakers, including non-native accents. However, in L2 training an SRE should ideally provide the learner with feedback on how “native-like” their pronunciation is.

2.6 The Grammar

Within the context of LANGA, the term “grammar” is used in the sense it is used in speech recognition technology. That is, the grammar specifies the set of possible words that the SRE should match input against. At a minimum, this is the words or sentences currently being taught. Grammars can be configured in the LMS via the authoring tool, by selecting training items from those available in the item database. Assigning items automatically populates the grammar with the list of words to be trained, and allows the user to either use the default SRE confidence threshold for correct responses, or to customize this (which may be necessary for some items, or to adjust the overall difficulty level). Games specify the minimum number of training items required for a grammar (based on the design of the game), however additional items can be specified in the grammar to act as additional possible SRE targets, but which are not taught as part of the game. In this way, the performance of the SRE can be adjusted to meet certain learning goals. For example, at early stages of learning, approximate pronunciations may be acceptable. However, as the learner progresses, they must make more fine-grained distinctions in the sounds of L2 words, including between “minimal pairs” – words with distinct meanings that differ only in a single phoneme (e.g., in English late vs. rate – a distinction that is challenging for native speakers of many Asian languages because these two sounds comprise a single phoneme in those languages, and so sound the same to Asian speakers). If the set of items in the grammar are phonetically highly distinctive from each other, then the odds are that the SRE will match a poorly-pronounced version of a word to its correct target; if, however, phonetically similar “foils” are added to the grammar, then the SRE becomes more sensitive to pronunciation errors if these result in a token that is more like one of the foils. Thus phonetically similar items can be added to the grammar not as learning items, but to tune the sensitivity of the SRE according to phonetic parameters that the curriculum designer/researcher wishes learners to focus on and/or improve.

2.7 The Authoring Tool

The authoring tool is a graphical Web interface that serves two primary functions for designers: user management and content management. The user management interface allows authorized users (designers) to add or modify user accounts, display a list of registered users, to assign users to curricula, to inspect the data for a particular user (e.g., curriculum they have been assigned to; amount of use; performance; progress through the curriculum), and to contact users via email.

The content management interface allows designers to manage learning content and build curricula. Designers can assemble curricula from existing lessons/activities/words that they are authorized to access, or build content from scratch. This could comprise anything from adding new words, building new activities from existing words, building lessons from pre-existing activities, etc. When a designer assigns a word to a game, this automatically populates the game with relevant item information, including the media presented to the learner (picture, sound file(s), written form), a listing of that word in the SRE grammar, a phonetic transcription of the item (for use by the SRE), and an SRE confidence score threshold for that item. Using this modular structure, designers with no programming experience can readily implement quite elaborate curricula, manipulating the type and quantity of material taught, how it is taught, the review schedule, etc. This also makes it easy to do between-group comparison studies, such as teaching different languages using the same curriculum, or comparing the efficacy of different curricula for teaching a particular language.

2.8 User Database

User accounts can be created by authorized users (designers) or, once LANGA is publicly available, by individuals through a sign-up page. The database has been designed with user security in mind, with users being associated with one or more “groups” that control who has access to that user’s data. In this way, a researcher could create a set of accounts for a particular study under a group for that study, and restrict access so that other designers would not be able to access that group’s data. Learners’ accounts do not allow access to other users’ information. LANGA’s user management system conforms to the United States Federal Trade Commission Children’s Online Privacy Protection Rule (COPPA), and so even children under age 13 can use LANGA (with parental consent) [19]. The SQL-based user database is designed to associate user information with user IDs. It has been designed to scale to large numbers (many thousands) of users and to maintain several categories of information. These include demographic information provided by the user (e.g., age, sex, years of education); learning activity (including dates and durations of use, performance in each training session, items exposed to during training session, etc.); results of assessments (including assessments of how well the user has learned the L2 at various stages, as well as tests of other things, like short-term or working memory attention, that may be relevant to L2 learning outcomes); and also EEG neuroimaging data (which could be obtained.
in a laboratory, or using a commercially available consumer grade EEG headset, and uploaded to the database in the open-source EEGLAB file format).

Because LANGA uses industry-standard protocols, including REST APIs and SQL for data storage, access to the data is easily done using standard SQL queries. These can also be executed through other software; for example, the widely-used, open source R statistical package has libraries for querying SQL databases and importing the desired data. To this end, the user database is designed to readily allow interrogation of the data by researchers, and to scale easily to thousands of users. In this way, LANGA has to potential to leverage Big Data analytics at such time as a large number of users are actively engaged and learning on the platform.

3 PROOF-OF-CONCEPT STUDY USING THE LANGA PLATFORM

Although LANGA has not yet been publicly released, we have conducted an initial proof-of-concept study to establish basic usability and efficacy of the platform and initial teaching approaches. Here, we present the results of one such study to demonstrate LANGA’s utility as a research and teaching platform. It must be emphasized, though, that these are intended to demonstrate the basic functionality and some of the research potential of this platform, rather than representing the type of large-scale research that the final version will enable.

An early question was whether there were differences between how well words were learned via two different teaching approaches. One strategy, referred as “rote” training, is a paired-associates task typically used both in classroom settings or by other computer-based programs. In this task each word to be learned (nouns and verbs) is represented by a picture paired with its spoken form (see Figure 1, left panel). Learners are prompted to repeat the word aloud and give feedback by the SRE. The other strategy, “inferential” training, consists of pairing a picture depicting an actor performing an action on an object with the corresponding Spanish written and spoken three-word sentence with a subject-verb-object (SVO) structure (e.g., La bruja abate el basurero – “The witch throws the garbage can”). As with the rote training, the learner is asked to repeat the L2 words (in this case the sentence rather than a single word). Activities present only 4 novel words at a time, so across trials the meaning of each word in the sentence can be inferred based on the picture, the SVO syntax, and recollection of prior trials. In a series of training items each picture/sentence varies from the last only in one word (i.e., one of the nouns or the verb). This aids the learner in making inferences as to word meanings, and “bootstrapping” later learning based on newly-acquired knowledge about the language. Our motivation to examine the contrast between rote and inferential teaching strategies was based on prior work suggesting that inferential learning may be more effective than rote learning, in the long run [20, 21, 22, 23, 24, 25, 26, 27, 28, 29].

Participants followed the 10-day curriculum detailed in Figure 2, and their learning was assessed using both behavioral and EEG neuroimaging methods. To assess learning, we employed two well-established behavioral tasks — naming and forced-choice recognition — to assess learning outcomes, and we also looked at changes in brain activity using event-related brain potentials (ERPs). Naming was chosen as a “gold standard” assessment of vocabulary learning because it requires the learner to retrieve the appropriate L2 word from memory and produce it accurately, based only on a picture representing the semantic content. The recognition task was included as an additional measure because it provides additional sensitivity to words that the learner may be familiar with upon hearing them, even if they are unable to retrieve the L2 form of the word based only on the picture. ERPs were used as a complementary measure because they can reveal the underlying cognitive processes employed in a task, which are not necessarily clear from behavioral data alone. ERPs are derived from EEG (ongoing recordings of brain electrical activity from electrodes placed on the scalp), by taking segments of EEG time-locked to the onset of events of interest and averaged across all events of a particular category. In this study the dependent measure was the N400 ERP component, which is a negative-going wave maximal over the vertex of the scalp that peaks at approximately 400 ms after hearing or reading a content word such as a noun or verb [30]. The N400 is thought to reflect neurocognitive processes involved in recognizing a word and accessing its meaning [31, 30, 32]. For example, if a picture is presented followed by a spoken word, the N400 is larger when the word does not match the picture, relative to when the word matches the picture (e.g., larger in response to the word dog after seeing a picture of a bed, relative to seeing a bed and hearing the word bed). In previous work, N400 amplitude has been shown to be correlated with L2 proficiency [32], but as well L2 learners can show N400 effects even prior to their showing behavioral effects of learning [33]. Thus the N400 can be used as an index of learning that may be more sensitive in the early stages than behavioral tests (and independent of the effects of a learner’s confidence, which may affect their behavioral responses), and in particular to reflect processing of words at the lexical-semantic level.

3.1 Methods

3.1.1 Subjects

Twelve subjects (all right-handed; four males; mean age = 23.5 years, SD = 3.57) took part to the experiment and received small financial compensation. All participants had normal hearing and normal or corrected-to-normal vision, and no known neurodevelopmental, neurological or psychiatric disorders. Participants were all native English learners and reported little or no prior exposure to Spanish; all participants were given a pre-test for knowledge of the Spanish words to be taught in the study and anyone who correctly named more than 3 items was excluded from the study. Nine participants reported speaking a second language fluently (four French; one Urdu; one Polish; one American Sign Language); four reported knowledge of a third language (French, German, Punjabi and Korean); two reported knowledge of a fourth language (Italian and Arabic) and finally, only one participant reported knowledge of a fifth language (French). All the participants
gave informed consent according to the Declaration of Helsinki; all procedures were approved by the Dalhousie University Research Ethics Board.

3.1.2 Stimuli
Participants were taught 72 Spanish words (48 nouns and 24 verbs). Half the words (24 nouns and 12 verbs) were taught using the rote method and the other half via the inferential method; words were divided into two sets so that half of the participants learned one set via the rote method, and the other half of participants learned that set via the inferential method, and vice-versa. As mentioned before, Spanish-English and Spanish-French cognates (e.g., el león, the lion), were excluded from the curricula. Both transitive (e.g., sacude, “to shake”) and intransitive verbs (e.g., salta, to “jump”) were included in the training material. Spoken versions of training words were recorded from a male native Spanish speaker. Each noun was preceded by a definite determiner, (el for masculine nouns and la for feminine nouns; with 26 masculine and 22 feminine nouns in total). Verbs were produced in the third person singular, present tense. Pictures used during training were drawn by a professional artist and all followed the same “cartoon” style. For testing, a different set of pictures were used in order to ensure that learners were associating words with semantic concepts, rather than simply with the particular visual stimuli used during training.

3.1.3 Procedure
Spanish proficiency was assessed three times: prior to learning, the day after the last training session was completed, and approximately 10 days after the last training session. Proficiency was assessed using two tests: a naming test and a forced-choice recognition task; EEG was recorded during the latter task. In both tests, the order of items was randomized, and items taught via the rote and inferential methods were intermixed. All testing was conducted in a room insulated from external acoustic and electrical noise. Subjects were asked to sit on a chair positioned in front of a computer monitor for all testing.

During the naming task, participants wore a set of headphones with microphone and held a gamepad with both hands, so that they could press a right or left button with respective fingers. The task consisted of 72 trials, one per word in the training set. During each trial, a picture corresponding to one of the words. Pictures were initially surrounded by a red line, indicating that the recording function was off. If the participant believed they knew the corresponding Spanish word, they were instructed to press the one button on the gamepad to activate the recording function. There was no constraint on how much time subjects could take before deciding to say the word. Once the proper button was pressed, a transition from red to green line surrounding the picture signalled that the recording function was on, and from that moment subjects had three seconds to say the word. Recorded files were later scored for accuracy. If they did not know the word, participants pressed an alternate button to advance directly to the next trial.

During the forced-choice recognition task (during which EEG was also recorded; see below), participants saw a series of 144 trials. On each trial, a picture was shown and after 1 s a spoken word from the training set was played. On half the trials, the word matched the picture; on the other half, the word was a different word that had been taught using the same approach (rote or inferential). Upon hearing the word, participants had to indicate via button press whether the word matched the picture or not.

Data were analysed using Linear Mixed-Effects modelling (LME) separately for naming task, match-mismatch task and EEG [32]. The fixed effects part of the models included stage (three levels: pre, post and follow-up) and training strategy (two levels: rote and inferential) as predictors; as well random intercepts for each subject were included. In the case of behavioral results, the dependent variables consisted of accuracy in the naming and match-mismatch tasks (two levels: correct and incorrect). In the analysis of ERPs the dependent variable was the difference between mismatch and match ERPs, calculated as the mean amplitude on each trial, in the time window from 500-500 ms post-stimulus onset from the Cz electrode (located at the vertex of the head, where the N400 is typically maximal [32]).

3.1.4 EEG recordings
EEG was recorded during the forced-choice recognition task. We used a Brain Products (Gilching, Germany) V-Amp/ActiCap 16 channel EEG system, with a sampling rate of 512 Hz and online 100 Hz lowpass filtering; electrode impedance was kept below 20 kΩ during the recording. Electrodes were placed at the following locations according to the International 10-10 System: Fp1, Fp2, F7, F3, Fz, F4, F8, C3, Cz, C4, TP9, P3, Pz, P4, TP10, Oz. EEG data were processed offline using the EEGLAB [34] running on Matlab (Mathworks; Natick, PA), version 2015a. Processing included applying a 0.1 - 30 Hz bandpass filter, with a 6 dB rolloff per octave. The continuous EEG was then epoched from 200 ms before to 1000 ms after the onset of each auditory cue, and individual trials were inspected for artifacts and any containing excessive noise were manually removed. Channels that were excessively noisy throughout the recording were also removed Independent Component Analysis (ICA, [34] [35]) was then used to remove ocular and other artifacts. Any bad channels previously removed were interpolated after ICA using spherical splines.

3.2 Results
3.2.1 Naming Task
The main effects and the interaction terms were assessed with a stepwise regression where the baseline model had only the intercept. Results showed a main effect of stage, $F(2) = 51.23, p < .0001$. There was no effect of training strategy, $F(1) = 0.04, p = .84$, nor a statistically significant interaction between stage and training strategy, $F(2) = 0.46, p = .63$. Post-hoc comparisons for the effect of stage ($p$-values reported are Bonferroni-corrected for 3 comparisons) showed a statistically significant improvement in performance from pre- to post-training, $\chi^2(1) = 78.12, p < .0001$, and from pre-training to follow-up, $\chi^2(1) = 75.55, p < .0001$. No change in performance was observed between post-training and follow-up, $\chi^2(1) = 0.022, p < .88$. 
post-training and follow-up sessions, was no significant change in N400 amplitude between the pre-training to follow-up, post-training, in the amplitude of the N400 component from pre- to post-training, because of the large number of electrical artefacts. Results showed a main effect of stage, \(F(2) = 134.21, p < .0001\). There was no effect of training strategy, \(F(1) = 1.74, p = .19\), nor a statistically significant interaction between stage and training strategy, \(F(2) = 0.15, p = .23\). Post-hoc comparisons for the effect of stage (corrected for 3 comparisons) showed a statistically significant improvement in performance from pre- to post-training, \(\chi^2(1) = 208.53, p < .0001\), and from pre-training to follow-up, \(\chi^2(1) = 193.85, p < .0001\). No change in performance was observed between post-training and follow-up, \(\chi^2(1) = 0.29, p < .60\).

### 3.2.2 Match-mismatch task

Results showed a main effect of stage, \(F(2) = 5.819, p = .005\), while neither the training strategy nor the Stage \(\times\) Training Strategy interaction were significant, \(F(1) = 0.079, p = .78\); \(F(1) = 0.316, p = .73\), respectively. Post-hoc comparisons (corrected for 3 comparisons) showed a statistically significant increase in the amplitude of the N400 component from pre- to post-training, \(\chi^2(1) = 10.59, p < .003\) and also from pre-training to follow-up, \(\chi^2(1) = 6.30, p < .036\). There was no significant change in N400 amplitude between the post-training and follow-up sessions, \(\chi^2(1) = 0.55, p = .99\).

### 3.2.3 ERP

Data from one subject was excluded from the ERP analysis because of the large number of electrical artefacts. Results from the stepwise regression showed a main effect of stage, \(F(2) = 5.819, p = .005\), while neither the training strategy nor the Stage \(\times\) Training Strategy interaction were significant, \(F(1) = 0.079, p = .78\); \(F(1) = 0.316, p = .73\), respectively. Post-hoc comparisons (corrected for 3 comparisons) showed a statistically significant increase in the amplitude of the N400 component from pre- to post-training, \(\chi^2(1) = 10.59, p < .003\) and also from pre-training to follow-up, \(\chi^2(1) = 6.30, p < .036\). There was no significant change in N400 amplitude between the post-training and follow-up sessions, \(\chi^2(1) = 0.55, p = .99\).

### 4 Discussion

In this paper we presented the philosophy and design of LANGA, a game-based platform for L2 learning, and presented preliminary evidence for its effectiveness. The goal of this small-scale study was to demonstrate how LANGA's research-oriented, modular design can support an evidence-based approach to developing L2 apps, and also act as a platform for L2 acquisition research. Results from our proof-of-concept study showed significant gains in L2 proficiency. Learners with no prior knowledge of Spanish vocabulary were able to correctly name approximately 75% of the words included in the curriculum, and to perform sound/meaning associations in a match-mismatch task with accuracy close to 100%. Behavioral improvements where paralleled by the emergence of distinct neural responses to semantically congruous and incongruous picture-sounds pairings, reflecting successful consolidation of the learned material. Interestingly, both behavioral and neural measures of gains were maintained when tested during the follow-up session. We did not find any differences between the rote and inferential training strategies.

Our goal in presenting these results here is not to provide definitive data on the efficacy of the teaching methods employed, nor can we conclude on the basis of this relatively small sample of learners that there are no differences in efficacy between rote and inferential teaching approaches. Rather, this study is presented to demonstrate the empirical, evidence-based approach embodied in — and facilitated by — LANGA’s design. Indeed, it is important to emphasize that LANGA is a platform that enables the deployment and testing of many different approaches to L2 teaching, rather than a particular approach to L2 teaching.

This study represents one of the first to systematically evaluate the efficacy of an L2 app, while controlling amount, intensity and type of training. In this regard, our study shares some strengths and limitations with the studies of other L2 apps [13], [14]. On the one hand, it represents a systematic, empirical assessment of the effectiveness of an L2 app, a much needed advancement in the domain of L2 apps; on the other hand, the early-stage of development required that we focus the investigation on a restricted and relatively homogeneous sample of participants, limiting the ability to generalize our conclusions to the larger population. These are practical limitations of conducting research in a laboratory setting. For this reason, LANGA has been designed as a suite of games that can be delivered via web browser or mobile device, with all user activity tracked in a central database. This enables scaling to thousands or more users, which in turn will provide a rich data set that will enable leveraging “Big Data” analytic methods. While this in itself is not unique among L2 apps, LANGA’s design emphasis on research, with the ability of individual teachers or researchers to customize learning content and assess results, puts it in a unique position. To exploit this potential, LANGA features an embedded authoring tool that allows researchers to easily modify and include new content, implement training strategies and design new curricula to be tested. This level of flexibility will reveal its full potential once the application will be fully developed.

In conclusion, in this paper we have described the core
concepts behind the development of LANGA. Building on results from this proof-of-concept study, future studies will test a beta version of LANGA featuring our in-house SRE and a fully functional database, with a larger number of learners. In the future we plan on making LANGA available to other researchers and teachers to enable a range of uses.

ACKNOWLEDGMENTS

The authors wish to thank Copernicus Studios, Inc. for their assistance on this project, including the donation of software and financial contributions to the costs of the research described here. This research was supported by funding from that National Sciences and Engineering Research Council of Canada (NSERC) awarded to AJN. FU and KO were supported by the RADIANT NSERC CREATE training program.

REFERENCES


Francesco Usai BSc (2011), MSc (2013), is a PhD student in Psychology & Neuroscience at Dalhousie University. His research interests include neural mechanisms of language acquisition and recovery from language disorders.
Kiera G.R. O’Neil  Kiera O’Neil, BSc (Hons) (2013), MSc (2015), is a PhD candidate in the Psychology & Neuroscience program at Dalhousie University. Her research interests include bilingualism, adult second language learning, and neuroplasticity in blindness.

Aaron J. Newman BSc (hons) (1996), MSc (1998), PhD (2002) is a Professor in the Departments of Psychology & Neuroscience, Pediatrics, Psychiatry, and Surgery at Dalhousie University. He is a past Canada Research Chair (Tier II) and has published over 30 peer-reviewed articles and more than 100 peer-reviewed abstracts. His research focuses on the use of multi-modal neuroimaging to study neuroplasticity of language.