Furhat LingAI Language-Teaching Bot: an Analysis Study on How Motivational Aspects Affect Language Learning Facilitation

Kayra Özdemir [redacted]
Vrije Universiteit Amsterdam [redacted]

Fatih Akçaoğlu [redacted]
Vrije Universiteit Amsterdam [redacted]

Walid Azarkan [redacted]
Vrije Universiteit Amsterdam [redacted]

Sami Bouga [redacted]
Vrije Universiteit Amsterdam [redacted]

Abstract— **This paper explores the use of artificial intelligence to enhance language learning through a tool called *LingAI*, developed on the *Furhat* platform. The models provide introductory Dutch lessons to international students, aged between 18-22, utilizing two interaction models: a friendly model and a strict model (sample size: 6). This difference in models aims to evaluate the impact of these models on learning outcomes, particularly focusing on whether motivational reinforcement enhances language acquisition. The experiment conducted assesses the students' performance through tests administered after interactions with each model. Results indicated no significant difference in memorization between the two models, with average scores of 6.83 for the strict model and 6.67 for the friendly model. Despite limitations such as small sample size and confounding variables, the study suggests that both interaction styles are equally effective in this context. The study also provides insight on how further research on the topic could be made, what key aspects need to be considered, and how AI's role in language learning could affect optimization.**

Keywords— artificial intelligence, human-computer interaction, language learning, robots, psychology, human memorization, Dutch, educational technology, interaction models, AI learning tools.

I. Introduction

Learning different languages is becoming more and more important in the increasingly globalizing world, yet old-fashioned methods (i.e., human tutoring, textbooks) often lack the personal involvement needed for effective learning. Recent developments in AI (def. artificial intelligence) offer new approaches to language learning, providing customized and interactive experiences. This paper uses a language learning tool called *LingAI*, developed using the *Furhat* robotic platform, providing an introductory lecture on Dutch to international students. The *LingAI* agent is designed to facilitate language learning through two different interaction models: a friendly model and a strict model. This paper aims to assess the outcomes of these two interaction models, determining whether motivational reinforcements truly enhance learning. These models differ in their approach and interaction style, potentially catering to different learning preferences and influencing the effectiveness of language acquisition. The implementation of AI methods in language learning caves a path to major improvements of the traditional teaching methods. This research does not only compare two interactions styles, but also seeks to offer a wider perspective on how AI models can be optimized to teach people different languages.

II. Literature review

The human memory span is a limited space, able to only hold finite amounts of information in a temporary and restrictive space. It is therefore important for the stimuli/information presented to an individual to be as efficient and involved as possible. Any extra amount of information limits the potential of the short-term memory, therefore reducing the total learning yield [1]. According to Atkinson & Shiffrin [1], information and stimuli received is sectioned depending on the 'type' of stimulus; that is, auditory stimuli are processed altogether, yet separate from other types of stimuli (visual, gustatory etc.) and therefore a motivational reinforcement is further efficient if presented distinctively than the desired teaching [2]. Hence it can be deduced that motivational reinforcement may hinder maximum efficiency, if not presented in the right way. Extra tools may help facilitate learning however it depends on the individual, and the tool itself [3]. Recently, there have been many emergences of teaching-assistive bots ranging from English, to German, to Japanese; furthermore, even in fields such as psychology (*Freudbot*). These models all have a common goal: facilitating learning - *LingAI* is no different [4]. The tool/model used for this paper is of type **b**, "language tutoring systems" that focuses around on-demand resources according to Rebolledo Font de la Vall & Araya [3]. The limitations of such AI, as noted in [3], may include lack of human interaction, limited ability to recognize errors, etc. These two aspects were the focus of challenges during the developmental stage of *LingAI*. According to Hardan, strategies that could be implemented for language learning in AI split into two categories, and the two split into further three sub-categories each. The strategies used for *LingAI* are part of Direct II – namely cognitive strategies –

such as practice, input-output; and Indirect II & III – namely affective and social strategies – such as encouragement, anxiety lowering, asking questions, cooperation [5]. Engwall & Lopes identified the optimal scope magnitude for language-teaching AI assistants to be not so large, mentioning that such models should start off small and not overwhelm the user in the human-computer interaction that takes place. Engwall & Lopes suggested approximately 10-minute interactions, rather than longer sessions that are opted to when the tutor is a human individual [6].

III. Methods

The experiment consists of two different versions of the *Furhat* based *LingAI* model: the friendly model, and the strict model (later referred to as the "efficient" model by the subjects contributing to the pilot study). The two versions differ in their completion time, due to the different attitudes they hold; yet the purpose of both models is the same. The friendly model, as the name suggests, proposes a more friendly attitude towards the subject: asking for their name, making small talk in the introduction, giving affirmative, friendly confirmations upon correct answers in oral tests; or motivational consolation upon incorrect answers. Due to this friendly behavior, the interaction takes a slightly longer completion time for the first model. The second version of the bot, however, proposes a "strict" manner. There are no negative remarks on the subject, yet the bot only operates with the same teachings of the friendly version. In essence, they aim to obtain the subject with the same (level) of information on the Dutch language.

A. Study Protocol

The participants are selected to be international students living in Amsterdam, aged between 18-22, with no prior experience on the Dutch language. As the sample group is selected within students already living in a Dutch-speaking country, a regulation is done by the questionnaire presented to the subjects beforehand, where they are asked about their experience on Dutch. If the answer given is **A2 or above**, the participations are excluded from the experimentation. If the participant passes the pre-assessment, they are presented with the consent form. Six participants were accepted to the experimentation of the eight that has taken the pre-assessment.

The experimentation is done in a noise-isolated room without other human presence. The visual *Furhat* robot is used in the interaction to replicate the original physical version of the *Furhat* robot. The subjects are presented first with either version of the bot, selected at random. The interaction follows as: *Furhat* welcomes the user, asks them what their name is (only for the friendly version), and goes on with the lesson. In the session, *LingAI* presents some Dutch words, some phrases, how certain letters are pronounced in Dutch (i.e., g, j), and verbally does a practice quiz – where the user is asked to repeat the words or phrases they've just learned. For the friendly version, the interaction also involves a small-talk section in the beginning and motivational reinforcement/appraisal upon wrong/right answers, respectively. The strict version doesn't apply any specific callout upon right or wrong answers, instead just lets the user know "correct" or "incorrect". After an hour, the subject is presented with the first test, where they are instructed to answer questions about what is taught in the interaction. (See material & apparatus in the *.zip* file) After 24 hours, the participant is presented with the remaining model of the bot; and yet again presented with the test after an hour. The interactions take approximately 5 minutes for the strict version and 6 minutes for the friendly version. The post-interaction tests again take place in a quiet atmosphere to eliminate any distractive stimuli. The tests are multiple choice questions, going over the learning acquisitions from the interaction sessions. Upon completion of the latter test, the experimentation phase for that subject is fulfilled.

B. Study Design

The methodology used for this study is an observational method, testing the results of the participants in terms of performance via assessments after the interactions. The experiment is a within-subject design, as subjects are presented with both versions of the model and compared within their own performances.

Counterbalancing measures need to be taken in the study, as the question results asked to the user at the end of the interactions may be influenced by the relevance effect [1]. In order to have an accurate evaluation on whether permanent learning has taken place or not -and it isn't just the relevance effect- the tests are implemented an hour after the interaction has taken place. The second counterbalancing measure is taken to eliminate learning and asymmetric skill transfer effects [2], where users are presented with the second version of the bot 48 hours afterwards. In [3], the authors mention the need for (approximate) equality between dependent variables to have proper results. This is implemented in the experiment conducted: as participants can't be presented with the same words/learnings taught by the AI, they are presented with words and learnings that are on the same level of difficulty (i.e., beginner, A1 level Dutch).

The independent variable of the study is the model version (strict/friendly), and the dependent variable is therefore the user memorization (i.e., performance). Some random variables in the experiment may be listed as: gender, culture background, age (despite the specific age group), etc. The control variables in the study are the computer, the atmosphere of the experimentation, the assessments, etc. Some notable confounding variables may be the exposure to Dutch by some students, as the sample group is selected as international students living in Amsterdam. The pre-assessment endeavors to eliminate this. Another confounding variable may be the similarity between the English words and the Dutch words presented, yet either way the performance varies, and this confounding variable is tried to be hindered.

C. Material & Apparatus

Necessary material for the experiment consists of a computer with the *Furhat* application downloaded and the *LingAI* simulation models installed, the pre-assessment form, the consent form, post-interaction tests. The pre-assessment form, and post-interaction test is digital material, within the workspace of a Google Forms document. The consent form is as a form of a *pdf* file. The *LingAI* simulation runs in a browser, and the

browser selected for this study is Mozilla Firefox - although this does not change the outcome. The computer used for the experimentation phase is a MacBook Air (M1). Depending on user preference, a set of headphones or earphones are also presented, however this does not affect the outcome of the study either.

D. *Data Collection & Measurement*

It is crucial to recruit enough participants to ensure the accuracy and reliability of the study's findings, henceforth the minimum subject amount is set to **five**. The data collected in the study is the user performance depending on the version of the *LingAI* model. The performance of the users will determine the general preference on the bot, as the learning acquisitions are chosen to be the same level of difficulty in Dutch (A1). Both intuitively and empirically, a better performance shows further efficiency of the bot, as it can be deduced that the subject had obtained more gain from that specific version of the model.

All 6 participants engaged in an interaction with *LingAI*, and the order of which model was presented was selected at random as a measure of counterbalancing. After an hour from each interaction, participants were provided with short quizzes in a digital form, each consisting of 8 questions. Both interactions with the social robot and both quizzes went smoothly without any interruptions or problems. All of the questions in the quizzes were about the words and phrases mentioned in the previous interactions – copies of the quizzes can be found in the study *.zip* folder.

Null hypothesis (H_0): There is no difference in user memorization between the two versions of the bot.

Alternative hypothesis (H_1): There is a difference in user memorization between the two versions of the bot.

IV. Results

This study had aimed to evaluate the effectiveness of two versions of the social bot *LingAI* in enhancing user memorization. 6 participants interacted with both versions of the bot, and their vocabulary/phrase memorization was assessed through two quizzes. The quiz scores, along with the number of participants who scored points, for both versions can be observed in figures 1 & 2.

Participant Scores

#	Participant Name	Strict Version Score	Friendly Version Score	Prior Experience Level
1	Participant A	7	7	NO KNOWLEDGE
2	Participant B	7	8	EXPOSURE ONLY
3	Participant C	8	8	A1
4	Participant D	7	7	EXPOSURE ONLY
5	Participant E	5	6	EXPOSURE ONLY
6	Participant F	7	3	NO KNOWLEDGE

fig. 1 - participant scores data

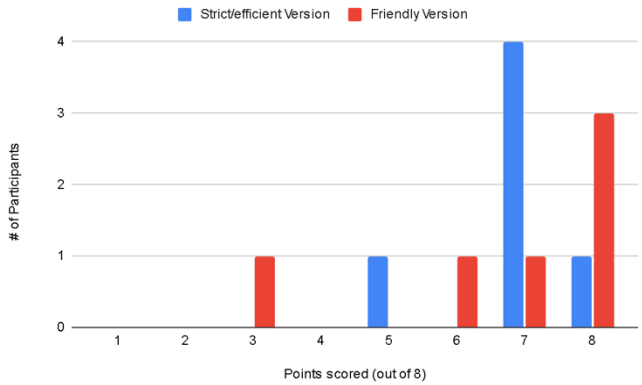

fig. 2 - participant both quiz results visual comparison

Participants achieved average scores of 6.83 and 6.67 in the Strict/efficient and friendly versions of the bot, respectively. Despite the fact that participants' scores exhibit a high degree of proximity to each other, a paired t-test was still conducted to statistically reject the null hypothesis. The paired T-test is appropriate for this analysis because it has close relation with the within-subject design.

Given the small sample size (n = 6), the calculated p-value was compared to the significance level (a = 0.05). Calculated result of the t-test yielded a p-value of 0.858, suggesting that the null hypothesis could not be rejected.

V. Discussion

Based on the obtained p-value and average score numbers, it is undeniably clear that there isn't any apparent difference between the two versions of the social bot *LingAI*. An outlier in the experimentation was Participant C, who scored the maximum out of both assessments. Nonetheless, as the score is equal for both tests, this did not change the results. There are several inadequacies in the study that may have affected the results to some level of extent: Uncontrolled, random, or confounding variables, low sample size, and somewhat deficient methodology. Every international student comes from a distinct and unique background where they are born and raised differently from one to another, and this may have led to an inaccuracy in the post-interaction assessments. This leads students to have their own distinct personalities and preferences on education, making it hard to draw conclusions. Low participant numbers limit the generalizability of findings and increases vulnerability to biases. Lastly, deficient methodology, such as a single interaction where participants interact with the social robot only once may not provide sufficient exposure to let participants engage with the versions of *LingAI*. To sum up, the analysis shows that there doesn't seem to be a significant difference between the two versions of the *LingAI* social bot. Although, it's important to consider the limitations of the study. Things like not controlling for all the variables, small number of participants, and the under-representing methodology consequently could have affected the results.

Some of the possible fixes to these limitations could be increasing sample size, broadening up the scope of the post-assessment, adding newer & more distinct features to the models, or perhaps even conducting the results assessment via a different method - such as testing solely on UX and performing a more open-ended questionnaire at the end to measure the more explicit outcomes of the study, i.e., which model users prefer without taking into consideration their performance of learning.

VI. Conclusion

Concluding, the objective of the analysis was to determine if there existed a statistically significant disparity in memorization performance between the two bot versions. The average scores obtained in the quizzes for the strict and friendly versions were quite similar, leading to the inability to reject the null hypothesis, inferring that there is no dissimilarity in user memorization between the two bot versions. All in all, the findings revealed no notable distinction between the friendly and strict versions of the bot.

In conclusion, although this study did not discover a significant distinction in the effectiveness of the friendly and strict versions of *LingAI*, it emphasizes the significance of robust research designs and the consideration of various influential factors in the collision of the fields human learning and human-computer interaction. Subsequent studies should strive to refine these aspects to gain a better understanding of the intricacies of user interaction and learning outcomes within the context of interactive bots. Furthermore, future work could also implement longitudinal studies, as perhaps such short interactions may not be sufficient for significant difference. Further improvements could notably be comparison among human tutors and AI tutors, measurements of differences between types of motivational reinforcements, or extended scopes of samples for evaluation of cultural background effects.

Statement of Contribution

Author K. Ö.: *Methodology* section, with subsections *study protocol*, *study design*, *materials & apparatus*, *data collection & measurement*; four sources on the *literature review* section (namely [1], [2], [4], and [6]) as well as writing & refinement, refinement of *introduction*, refinement of the *discussion* section, refinement of *conclusion* section, implemented *future work* options (in *conclusion*), *abstract*, *title & keywords* (along with Author F. A.), conduction of the *experimentation phase*, finalization on the formatting of the paper, implementation/formatting of *references* section, feedback implementation.

Author F. A.: Revision of *methodology* section (namely *study protocol*, *study design*, *data collection & measurement*), *results* & *discussion* section, implementation of tables, revision of *conclusion* section, creation & programming of the two models, *consent form*, preparation of materials & apparatus, *title & keywords* (along with Author K. Ö.), *pre-experiment questionnaire*, both *post-interaction assessments*, finalization of the *.zip* file, revision of *references* section, feedback implementation.

Author W. A.: *Introduction* section, three sources on the *literature review* section (namely [3] and [5]), revision of the *conclusion* section, video recording of *Lo-Fi prototype*.

Author S. B.: *Conclusion* section, revision of the *introduction* section, revision of the citations on the *literature review* section, video recording of *Lo-Fi prototype*.

References

[1] M. S. Gazzaniga, E. A. Phelps, and E. Berkman, *Psychological Science*, 7th ed. Rotterdam, Netherlands: Erasmus Universiteit Rotterdam, 2022, pp. 765-766.

[2] R. C. Atkinson and R. M. Shiffrin, "The Control of Short-Term Memory," *Scientific American*, vol. 225, no. 2, pp. 82-91, 1971. [Online]. Available: https://doi.org/10.1038/scientificamerican0871-82

[3] R. Rebolledo Font de la Vall and F. Gonzalez Araya, "Exploring the Benefits and Challenges of AI-Learning Tools," Universidad de Playa Ancha, Chile, 2023. [Online]. Available: https://www.researchgate.net/profile/Fabian-Gonzalez-Araya/publication/366957798_Exploring_the_Benefits_and_Challenges_of_AI-Language_Learning_Tools/links/63bb0cce097c7832ca9ee063/Exploring-the-Benefits-and-Challenges-of-AI-Language-Learning-Tools.pdf

[4] N. Haristiani, "Artificial Intelligence (AI) Chatbot as Language Learning Medium: An Inquiry," *Journal of Physics: Conference Series*, vol. 1387, no. 1, p. 012020, 2019. [Online]. Available: https://doi.org/10.1088/1742-6596/1387/1/012020.

[5] A. A. Hardan, "Language Learning Strategies: A General Overview," University of Anbar, Ramadi, Iraq, Dec. 10, 2013. [Online]. Available: https://doi.org/10.1016/j.sbspro.2013.12.194.

[6] O. Engwall and J. Lopes, "Interaction and collaboration in robot-assisted language learning for adults," *Computer Assisted Language Learning*, vol. 35, no. 5–6, pp. 1273–1309, 2022. [Online]. Available: https://doi.org/10.1080/09588221.2020.1799821.