Development of a Chatbot and Evaluation of Its Effects on Learning and Intrinsic Motivation of a Public Secondary School's Spanish Language Learners

Julian Eymard JANUBAS[*], Josiah Jose DEYSOLONG, Hanz Lucas ESTOPIA, Karl Mykell TABBAY & Jun Rangie OBISPO
[a]*Department of Computer Science, College of Computer Studies, Xavier University – Ateneo de Cagayan, Philippines*
*julianeymardjanubas@gmail.com

Abstract: This study developed and evaluated a Messenger-based chatbot designed to enhance learning and intrinsic motivation among Spanish language learners at a public secondary school. Using the pretest-posttest design, the study involved 28 students who were assessed on their Spanish language skills and motivational levels before and after interacting with the chatbot. Descriptive and inferential statistics revealed a statistically significant improvement in academic performance, affirming the chatbot's effectiveness as an educational tool. While the paired samples t-test did not show a statistically significant increase in intrinsic motivation, the Repeated Measures ANOVA did reveal a significant effect, suggesting that the chatbot had a meaningful impact on motivation when considering changes over time. Spearman's rho correlation analyses indicated no strong relationship between post-intervention intrinsic motivation and learning outcomes. Additionally, linear regression analysis showed that pretest grades were a strong predictor of posttest grades, whereas pre-intervention intrinsic motivation did not significantly predict posttest grades. These findings suggest that while chatbots can effectively enhance language proficiency, their influence on intrinsic motivation is more complex and requires further investigation.

Keywords: Chatbot, Intrinsic Motivation, Self-Determination Theory, Spanish Language Learning, Pretest-Posttest Design, Secondary Education, AI in Education

1. Introduction

Learning a foreign language effectively requires substantial exposure and opportunities for practice, which are often limited when the language is not widely spoken within the country (Ellis, 2008). Students often encounter obstacles such as low motivation, high anxiety, and ineffective learning strategies, which can impede their language learning progress (Sparks & Ganschow, 1993). Language instructors and researchers frequently report difficulties in applying vocabulary, creating accurate vowel and consonant sounds, grammatical rules, translating Spanish phrases and sentences, and conjugating and employing irregular verbs (Sapungan, 2016). Among these tasks, applying grammatical rules is seen as the most difficult, followed by translating and conjugating irregular verbs (Sapungan, 2016). Since its inception in 2009, the Philippines' Department of Education has implemented the Special Program in Foreign Language (SPFL) to enhance foreign language learning among secondary school students (Department of Education, 2017). To address the challenges in Spanish language learning, this study developed a Messenger chatbot and explored its effects on learning and intrinsic motivation among students at a public secondary school. The integration of digital technology in educational settings offers benefits, including diverse information access, enhanced engagement, personalized learning experiences, and increased learner autonomy and motivation (Bećirović, Brdarević-Čeljo, & Delić, 2021; Palmer, 2022).

Several key research questions were investigated for this study: First, how does the use of a language learning chatbot affect the academic performance of public secondary school students enrolled in a Spanish language elective course? Second, to what extent can the use of a language-learning chatbot affect the intrinsic motivation of these students? Third, what is the relationship between intrinsic motivation levels and academic performance among students using the chatbot for Spanish language learning?

2. Review of Related Literature

Recent studies highlight the effectiveness of chatbots in enhancing educational outcomes. For instance, a study demonstrated that using ManyChat in a blended learning environment significantly improved the performance of Grade 9 students in TLE Dressmaking compared to traditional methods (Corvera, 2022). In the Philippines, a study showed that chatbots helped senior high school students learn how to conduct research in various domains using Dialogflow, promoting flexible learning (Rivera & Ureta, 2018). At Xavier University, a chatbot used in the course "The Life and Works of Rizal" was found to enhance student engagement and motivation (Chiu et al., 2023).

Research underscores the significant role of chatbots in enhancing language-learning, offering personalized, interactive tutorials that adapt to student needs (Dokukina & Gumanova, 2020). Notably, GENGOBOT, a chatbot integrated with the LINE platform, supports Japanese language grammar learning through interactive exercises and content accessible directly within the chat interface (Haristiani et al., 2019). Another study highlighted chatbots' utility in English language teaching, particularly in mobile environments, where they enhance social interaction and provide authentic language practice (Chuah & Kabilan, 2021). Similarly, Dokukina and Dumanova (2020) noted that chatbots facilitate personalized and interactive language practice, although they pointed out challenges like the lack of emotional intelligence in chatbots. Additionally, a study in Vietnam showed a Facebook chatbot significantly improved English preposition learning, evidencing chatbots' effectiveness in foreign language instruction (Nghi et al., 2019).

Chatbots significantly enhance student performance, especially in language acquisition, by offering a platform for convenient language practice that leads to improved fluency and competence (Haristiani et al., 2019). Studies using the Non-randomized One-Group Pretest and Posttest Design demonstrate that chatbots positively impact Spanish language skills, showing a notable improvement in scores post-interaction (Chiu et al., 2023). Furthermore, in language and other academic areas, students using chatbots achieve better outcomes compared to traditional methods (Dokukina & Gumanova, 2020).

3. Methodology

3.1 Research Design

The research design is similar to previous study on a chatbot (Chiu et al., 2023). It involves data gathering, chatbot conversation outline, development phase, user acceptance testing (UAT), evaluation, and usability testing. The development process follows an agile software development life cycle for incremental development.

3.2 Detailed Methodology

Researchers collaborated with a Spanish language teacher who provided lessons sourced from the "Spanish 1-4: Beginner, Elementary, Intermediate and Advanced" Udemy course (n.d.). Materials focusing on present tense verbs and adjectives, relevant to the experimental week, were integrated into the chatbot's knowledge base. Additional content from "Spanish and Essential Grammar Review" (Torres-Robles, 2003) and conjugation resources from Linguasorb (2024) were included to enhance the chatbot's instructional capabilities, with all content formatted into text-based instructions for Chatfuel's knowledge blocks.

The chatbot featured three modes: Translation, Learn Bot, and Practice Quiz Bot, focused on "Adjectives" and "Present Tense Verbs." Users could select their preferred mode, with safety measures in place for unrelated queries. Chatfuel's NLP engine processed inputs and accessed stored data blocks to provide accurate responses.

Development, which took 10 weeks, involved creating conversation flows and blocks in Chatfuel, linking the chatbot to a Facebook page, and using rule-based NLP to manage various user queries. The chatbot was integrated with ChatGPT for enhanced functionality.

Five students from the partnered secondary school participated in user acceptance testing (UAT) of the chatbot, completing nine tasks and providing feedback through a qualitative five-item questionnaire. Their input helped refine the software to meet user requirements. Spanish language teachers and professors further evaluated the chatbot using modified Nielsen heuristic forms, assessing language fluency, contextual understanding, and natural conversation flow (Benaida, 2023). This evaluation offered crucial insights for improving the chatbot and ensuring it met educational objectives.

Student participants took pretests and posttests to measure their progress in Spanish language proficiency in topics like present tense and adjectives after using the chatbot for one (1) week—from April 10–17, 2024. The Self-Determination Theory of Second Language Scale (SDT-L2) Intrinsic Orientation assessed their intrinsic motivation (Alamer, 2022). It consists of 20 items, evenly distributed to four (4) specific constructs: intrinsic orientation, identified orientation, introjected orientation, and external orientation. The researchers focused on the results of intrinsic orientation to answer the research question concerning learners' inherent interest and enjoyment in the language learning process. The System Usability Scale (SUS), on the other hand, evaluated the chatbot's usability. The SUS test is a standardized and reliable tool that measures the perceived ease of use, efficiency, and learnability of a system through 10 statements rated by participants on a scale of 1 to 5., ensuring comprehensive feedback on its impact and user experience (Brooke, 1995).

3.3 Procedure and Data Analysis

To evaluate the chatbot's effects on students' academic performance, the researchers analyzed pretest and posttest data. The primary statistical method used was the Paired Samples T-Test to compare mean scores before and after the chatbot intervention, determining its effectiveness. Effect size calculations and descriptive statistics were also used to assess practical significance and summarize performance improvements.

To gain deeper insights, a Repeated Measures ANOVA was conducted to examine changes over time, considering both academic performance and intrinsic motivation. This method provided a more comprehensive analysis, allowing for the detection of significant effects that might not be captured by the t-test alone, such as the interaction between test grades and intrinsic motivation.

The SDT-L2 is a validated scale designed to assess second language learners' motivation based on self-determination theory (Alamer, 2022). Mean scores before and after the chatbot interaction were compared using a paired t-test or Wilcoxon signed-rank test, depending on data normality. A significant increase in post-intervention scores would indicate enhanced motivation, while a decrease would suggest diminished motivation.

To assess the correlation between learning and intrinsic motivation, pretest and posttest scores, along with SDT-L2 Intrinsic Orientation results, were analyzed using Jamovi. Spearman's rank correlation coefficient indicated the strength and direction of the relationship, with a positive rho value suggesting improved learning with increased motivation. Additionally, regression analysis was performed to further explore the predictive relationship between intrinsic motivation and learning outcomes.

4. Results and Discussion

4.1 Development

The UAT of the chatbot showed that while most tasks across the learn bot, translation, and quiz features were successfully completed, errors occurred in Tasks 2 and 3 due to a bug in Chatfuel's logical connectors, which was quickly fixed. Despite these issues, feedback was largely positive, with 100% approval for the quiz function and relevance of study topics. The chatbot also received a 90% rating for motivating Spanish learning, though responsiveness (60%) and navigation (80%) were noted as areas for improvement. Participants were generally satisfied with the chatbot's potential in their Spanish classes, particularly highlighting response time as an area needing attention.

Two Spanish language professionals evaluated the chatbot using a 12-item modified Nielsen heuristic on a 5-point Likert scale (Benaida, 2023). The first evaluator awarded a perfect score, indicating a strong alignment with Spanish language teaching practices. The second evaluator scored slightly lower, averaging 93.33%, praising the conversational flow and clarity but suggesting improvements in error prevention and user control. While some features weren't implemented due to Chatfuel's limitations, essential corrections were made to enhance readability and support continuous learning.

4.2 Data Analysis

4.2.1 Chatbot Effects on Academic Performance

To answer the first research question regarding the chatbot's effects on students' academic performance, pretest and posttest scores were analyzed, revealing a positive trend. The mean score increased from 13.4 to 14.6, showing an average improvement of 1.2 points. A Paired Samples T-Test found a significant difference between pretest and posttest scores ($p = 0.001$), with a mean difference of -1.214 and a 95% confidence interval from -1.893 to -0.5356, strongly indicating that the chatbot led to statistically significant learning gains. Additionally, the standard deviation of posttest scores decreased from 2.48 to 2.30, suggesting more consistent student performance levels post-intervention. This consistency, alongside the median score increase from 14 to 15, confirms that the observed improvements were due to the chatbot's educational intervention rather than random variation. This addresses the gap in existing research by providing empirical evidence of a chatbot's impact on academic performance, thus advancing knowledge in educational technology. The Repeated Measures ANOVA supports this interpretation, revealing a significant main effect for test grades ($F(1, 27) = 575.23, p < .001$). This indicates a substantial and statistically significant improvement in test scores over time, affirming the chatbot's effectiveness in enhancing academic performance.

4.2.2 Chatbot Effects on Intrinsic Motivation Levels

To address the second research question regarding the chatbot's effects on intrinsic motivation, descriptive statistics showed a slight increase in the SDT-L2 intrinsic orientation scores, from 3.81 to 4.01, indicating a modest positive shift in motivation. However, the Paired Samples T-Test found this increase was not statistically significant ($p = 0.132$), with a mean difference of -0.200 and a 95% confidence interval ranging from -0.464 to 0.0645. Notably, the post-intervention standard deviation (0.506) was slightly lower than the pre-intervention score (0.587), possibly indicating that the chatbot helped create a more consistent level of motivation across the student group, even if the overall increase in motivation was not substantial. Contrary to the t-test results, the Repeated Measures ANOVA revealed a significant effect of the chatbot on intrinsic motivation ($F(1, 27) = 16.21, p < .001$). This indicates that while the mean increase in motivation appeared modest, the repeated measures design detected a significant change over time, likely due to greater statistical power. This finding highlights the chatbot's potential to meaningfully impact intrinsic motivation, even if this wasn't fully captured by the initial t-test.

4.2.3 Relationship Between Learning and Intrinsic Motivation

To explore the third research question on the students' academic performance and intrinsic motivation relationship, Spearman's *rho* values were analyzed. These values came from correlation matrices comparing Self-Determination Theory of Second Language Scale's intrinsic orientation scores post-intervention with posttest grades, and pre-intervention SDT-L2 scores with pretest grades (See Table 1). Post-intervention, Spearman's *rho* was -0.085 ($p = 0.668$), indicating a very weak, non-significant inverse correlation, showing no strong relationship between post-intervention intrinsic motivation and posttest performance (See Table 1). Pre-intervention, Spearman's *rho* was 0.179 ($p = 0.362$), suggesting a weak positive, non-significant correlation (See Table 1). Linear regression analyses further examined the relationship between learning and intrinsic motivation. The first model showed that 55.1% of the variance in posttest grades was explained by pretest grades and pre-intervention intrinsic motivation ($R = 0.743$, $R^2 = 0.551$, $F = 15.4$, $p < 0.001$), with pretest grades being a strong predictor ($p < 0.001$), but pre-intervention intrinsic motivation did not significantly predict posttest grades. The second model showed that only 7.5% of the variance in post-intervention intrinsic motivation could be explained ($R = 0.274$, $R^2 = 0.075$, $F = 1.01$, $p = 0.377$), indicating no significant prediction by pretest grades or intrinsic motivation, suggesting other factors are significant in language learning. The Repeated Measures ANOVA revealed a significant interaction between test grades and intrinsic motivation ($F(1, 27) = 8.00$, $p = 0.009$), indicating that changes in academic performance were influenced by motivation levels. While the overall correlation between motivation and performance might be weak, this interaction shows that intrinsic motivation significantly impacts how students benefit from the chatbot. In educational research, this highlights the importance of considering both academic and motivational factors.

Table 1. *Correlation Analyses Between Pretest and Posttest, and Pre- and Post-Intervention SDT-L2 Intrinsic Orientation*

		PRETESTG	POSTTESTG
PRESDTL2I	*Spearman's rho*	0.179	—
	df	26	—
	p-value	0.362	—
	N	28	—
POSTSDTL2I	*Spearman's rho*	—	-0.085
	df	—	26
	p-value	—	0.668
	N	—	28

5. Conclusion

The chatbot development and testing included data gathering, conversation outline, and development, with UAT involving five students. UAT demonstrated a 100% success rate in navigating lesson topics and managing translations but failed to select, switch topics, and handle various quiz types. Heuristic evaluations by Spanish language experts averaged 96.67%, indicating high usability but highlighting the need for interface improvements.

The analysis confirmed a significant increase in learning, with pretest to posttest mean score improvements validated by both a paired samples t-test ($p = 0.001$) and a Repeated Measures ANOVA ($F(1, 27) = 575.23$, $p < .001$). This underscores the chatbot's effectiveness in enhancing Spanish language skills, successfully addressing the first research question.

For the second research question, although the paired samples t-test suggested the increase in intrinsic motivation measured by the SDT-L2 questionnaire was not statistically significant ($p = 0.132$), the Repeated Measures ANOVA revealed a significant effect ($F(1, 27) = 16.21$, $p < .001$). This suggests that the chatbot had a meaningful impact on intrinsic motivation, even if the change was modest and not fully captured by the t-test.

To explore the third research question, the relationship between learning and intrinsic motivation was analyzed using Spearman's *rho* and linear regression. The findings showed no strong link between intrinsic motivation and learning outcomes (post-intervention: *rho* =

-0.085, p = 0.668; pre-intervention: rho = 0.179, p = 0.362). However, regression analysis indicated that pretest scores and pre-motivation scores significantly predicted posttest results (R^2 = 0.551). Additionally, a significant interaction effect between test grades and intrinsic motivation ($F(1, 27)$ = 8.00, p = 0.009) suggests that while motivation alone may not strongly predict performance, it plays a role in how students benefit from the chatbot intervention.

To sum, the chatbot effectively improved academic performance in Spanish language learning. However, further research is necessary to fully understand its impact on intrinsic motivation and explore other factors influencing learning outcomes. These insights will guide future improvements and research on educational chatbots.

References

Ellis, R. (2008). *The Study of Second Language Acquisition* (2nd ed.). Oxford University Press.

Sparks, R., & Ganschow, L. (1993). Searching for the cognitive locus of foreign language learning difficulties: Linking first and second language learning. *The Modern Language Journal, 77*(3), 289-302. https://doi.org/10.2307/329098

Sapungan, R. M. (2016). Learning Spanish language: Motivation and challenges affecting Arab college students. *Journal of Language and Literature Education*, 3(1).

Department of Education. (2017). DepEd enhances learners' foreign language skills through Special Program in Foreign Language. https://www.deped.gov.ph/2017/02/20/deped-enhances-learners-foreign-language-skills-through-special-program-in-foreign-language/. Accessed: 2023-07-16.

Bećirović, S., Brdarević-Čeljo, A., & Delić, H. (2021). The use of digital technology in foreign language learning. *SN Social Sciences, 1*(10), 246. https://doi.org/10.1007/s43545-021-00254-y

Palmer, G. (2022). 7 benefits of using technology in foreign language learning. *ValiantCEO*. https://valiantceo.com/7-benefits-of-using-technology-in-foreign-language-learning/

Dokukina, I., & Gumanova, J. (2020). The rise of chatbots – new personal assistants in foreign language learning. *Procedia Computer Science, 169*, 542-546. https://doi.org/10.1016/j.procs.2020.02.212

Rivera, J., & Ureta, J. (2018). Using chatbots to teach STEM related research concepts to high school students. In Y.-T. Wu et al. (Eds.), *Workshop Proceedings*, 338-339.

Chiu, N. E. A, Pamakid, K. R., & Salvo, T. A. (2023). Investigating the impact of a chatbot on motivation and learning among students studying the life and works of Rizal. *Xavier University–Ateneo de Cagayan.*

Haristiani, N., Danuwijaya, A. A., Rifa'i, M. M., & Sarila, H. (2019). GENGOBOT: A chatbot-based grammar application on mobile instant messaging as a language learning medium. *Journal of Engineering Science and Technology, 14*(6), 3158-3173.

Chuah, K.-M., & Kabilan, M. (2021). Teachers' views on the use of chatbots to support English language teaching in a mobile environment. *International Journal of Emerging Technologies in Learning (iJET), 16*, 223-237. https://doi.org/10.3991/ijet.v16i20.24917

Corvera, J. (2022). Increasing grade 9 students' TLE MPS using a blended approach with ManyChat. *International Journal of Trend in Scientific Research and Development, 6*(6), 288-292.

Nghi, T., Huu Phuc, T., & Nguyen Tat, T. (2019). Applying AI chatbot for teaching a foreign language: An empirical research. *International Journal of Scientific & Technology Research, 8*, 897-902.

Alamer, A. (2022). Basic psychological needs, motivational orientations, effort, and vocabulary knowledge: A comprehensive model. *Studies in Second Language Acquisition, 44*, 164-184. https://doi.org/10.1017/S027226312100005X

Torres-Robles, C. L. (2003). Spanish Verbs and Essential Grammar Review. Purdue University Calumet. Retrieved January 10, 2024, from https://smithrivas.com/wp-content/uploads/2019/09/spanish-verbs-and-essential-grammar-review.pdf

Spanish 1-4: Beginner, Elementary, Intermediate and Advanced. (n.d.). Retrieved from https://www.udemy.com/course/spanish-101-beginning-spanish-spanish-for-beginners/

Spanish Conjugations - Linguasorb. (2024). Retrieved January 10, 2024, from https://www.linguasorb.com/spanish/verbs/

Benaida, M. (2023). Developing and extending usability heuristics evaluation for user interface design via AHP. *Soft Computing, 27*(14), 9693–9707. https://doi.org/10.1007/s00500-022-07803-4

Brooke, J. (1995). SUS: A quick and dirty usability scale. *Usability Evaluation in Industry*, 189.