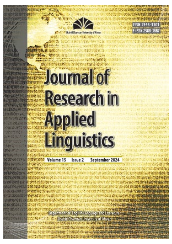
Journal of Research in Applied Linguistics
ISSN: 2345-3303 – E-ISSN: 2588-3887 – https://rals.scu.ac.ir
© 2024 – Published by Shahid Chamran University of Ahvaz

Shahid Chamran University of Ahvaz



Please cite this paper as follows:

García Laborda, J., Madarova, S., & Magal Royo, T. (2024). Issues in the design and implementation of chatbots for oral language assessment. *Journal of Research in Applied Linguistics*, *15*(2), 43-54. https://doi.org/10.22055/rals.2024.45822.3211

Research Paper

Issues in the Design and Implementation of Chatbots for Oral Language Assessment

Jesus García Laborda[1], Slavka Madarova [2], & Teresa Magal Royo[3]

[1]Corresponding author, Filologia Moderna-Intituto Franklin, College of Education, Universidad de Alcala, Alcala Henares, Spain; *jesus.garcialaborda@uah.es*

[2]Linguistica Aplicada, E.T.S. de Ing. de Caminos Canales, Universidad Politacnica de Madrid, Madrid, Spain; *slavka.madarova@gmail.com*

[3]Department, E.T.S.I. Aeroespacial y Diseño Industrial, Universidad Politecnica de Valencia, Valencia, Spain; *tmagal@degi.upv.es*

Received: 11/01/2024 Accepted: 24/09/2024

Abstract

Oral assessment in computer-assisted language learning is one of the best-known challenges at a technical and implementation level in an official language certification. The case of Spain is especially critical since the government has delayed for years the completion of the listening comprehension and oral expression test in the University Access Test (EVAU). This article presents first the evolution of oral tests at a general level, then a SWOT analysis of the potential of such implementation, and, third, how to implement them and the paper concludes that there is evidence that chatbots adapted to language learning can also be used for evaluation. Chatbot-assisted language learning with artificial intelligence adapted to voice recognition and its processing to obtain semi-automatic assessment supervised by the teacher can become a tool to be implemented within the language learning processes and / or included in the language certification methodology. Another interesting aspect could be the development and design of the interfaces of future adapted chatbots that must consider multimodality in the interaction between man-machine so that communication is effective and can be validated based on the knowledge available to the student.

Keywords: Chatbots; Language Assessment; Language Testing Examination; Oral Testing; University Entrance Examination; Spain.

1. Introduction

Oral testing is usually the most complex skill to deliver in high-stakes language testing (Soodmand Afshar, 2020), especially when the budget is very limited. The case of Spain is particularly critical since the University Entrance Examination (EVAU) is delivered in all 14 regions with a different construct despite the fact that it counts evenly towards admissions in public universities in the whole country. Additionally, the EVAU English examination has been carried out in a very similar way for more than 25 years without an oral and speaking section despite the, at least, three law decrees that stated its obligatory implementation over many years. The EVAU language testing is acknowledged to have quite a negative washback. There are several reasons for this: focus on traditional skills, grammar and vocabulary; an old type of items; scarce research in the Spanish University Entrance Examination (EVAU) and many other aspects. Researchers like García Laborda, Fernández Álvarez, Amengual Pizarro, Magal-Royo and others have pointed out the need to implement new types of assessments, especially technology-based (Fernandez Alvarez, García Laborda & Magal-Royo, 2022), to modernize the original exam including the new and expected oral/speaking sections.

For many, the most important omissions are still the verbal skills (listening and speaking). It is evident that a realistic assessment cannot be fully comprehended today without a speaking/oral section. However, despite the large number of proposals, the oral testing does not seem to be planned for implementation even in the new law in 2024, whose model has not been delivered yet, mostly due to the different conditions to be found in the different classrooms used for assessment in the public universities in Spain.

Since 2022, the Spanish Ministry of Education has indicated its intention to implement a new university entrance examination which appears to be a potentially beneficial and challenging possibility for improvement. Due to the COVID-19 pandemic and new legal relations based in the current new educational system inspired by the last law (LOMLOE), the government proved to be open to new proposals that include the integration of technology for this specific test. Since February 2022, the Spanish Rectors' Conference (CRUE) has shown some interest in revising the whole exam. However, CRUE lacks the legal validity to introduce changes in the legal system though it has been analyzing and reviewing EVAU carried out in most of the autonomous communities since October 2022.

Indicators show an improvement in the English language proficiency of students in Spain: however, this remains to be confirmed as, further information is anticipated from the forthcoming PISA Report, which will be conducted by Cambridge Assessments in 2025. Given that EVAU is a criterial test (at B1 in the CEFR), it is possible that the test may not show a direct correspondence between a candidate's greater knowledge of English and the performance in EVAU. In other words, a student whose competence might be C2 may receive a comparable score to one with a B1 level if both students perform well on the test. Thus, while in other countries it is a standard practice to know the actual competence of a student, in each Spanish region the exam is supposedly limited to measuring compliance with minimum knowledge threshold that is closely aligned with the LOMLOE. Again, is it impossible to conduct a genuine assessment without the inclusion of both listening and speaking tasks.

At this point, what seems to be at stake is the reliability and practicality of EVAU given the need to implement oral tasks. The benefits of technology-based delivery in language tests has also received positive feedback from the students (Díez Arcón & Martín-Monje, 2023; Huang & Li, 2024)). Additionally, previous research has proved the possibility to use chatbots for language assessment, and since chatbot tend to be dynamic in their response, they could implement an interactive computer-based delivery such as that suggested by Matt Poehner and his Dynamic Assessment (Poehner & Lantolf, 2013; Poehner & Leontjev, 2020; Poehner & Lantolf, 2023) In the light of aforementioned premises, this paper aims to address the principles of using chatbots in language testing, with a particular emphasis on their use in an automated way, including a study of applicability in the assessment tests of oral competencies in foreign languages and a possible implementation of an adapted chatbot for oral and speaking tasks into the new proposal of the Spanish language examination for EVAU.

A chatbot's interface design for learning and assessing oral tests requires synchronous use of at least two types of communication on an easy and affordable ubiquitous device (Magal-Royo & Garcia Laborda, 2022). The article points out the basics related to the design of a tool created with multimodal interfaces so that it offers a reasonably standard alternative to oral testing assessment through adapted chatbots. In the design of adapted chatbots the concept of multimodality is employed, which allows for the most efficient communication and interaction channels.

In short, this paper intends to address two main objectives: first, to provide an overview of the challenges presented by the implementation of chatbots as speaking skill assessment tools; second, to create a SWOT analysis stating the pros and cons of such an implementation.

2. Assessment and Technology

Language teachers spend a lot of time monitoring the progress of their students, both in the classroom during teaching and learning activities and outside of class, teaching the use of L2, reviewing their pronunciation, working on the range and precision of their vocabulary, showing the use of syntactic rules and the adequacy of their use of the language. In fact, it is possible that the aspect that is the least emphasized in the classroom, is what should be precisely the ultimate objective of teaching: communication (Pahissa & Tragant, 2009). Most teachers are able to assess the progress of their high school students informally (through observation and formative assessment), although it is not uncommon to have a misconception about a given student (Ross & Okabe, 2006). This naturally leads to the question of why do we want to evaluate our high school students? The answer is obvious: the pursuit of equity. All students in the second year of Baccalaureate are going to face a test to access university studies more or less close to their interests (Yin et al., 2021). The differences in the Baccalaureate and EVAU qualifications shape the future of a person: on many occasions, for their entire life. Therefore, teachers must strive for impartiality and fairness in the educational process, taking students' potential needs and wishes into account. The examinations assure the same opportunities for all students due to homogeneous and equal conditions (Moledo et al., 2014).

Testing and assessment processes also allow us to obtain a verification of students' progress; they help confirm the evaluations of each teacher and favor decision-making about the needs of students and even the potential success they will have in their university studies (Olani, 2009; Cerdeira et al., 2018). In the absence of real knowledge, exams do not provide objective data but indicators of the competence (often not of their performance as is generally said) especially when there is a significant effect of anxiety (Harley et al., 2021) and even more in high-stakes exams (In'nami, 2006; Finney et al., 2018). Assessment tests, and particularly exams, provide a certain degree of standardization (Stenlund et al., 2018), through which we judge the performance and progress of students and subsequently compare them with each other and with the criteria and objectives of the Baccalaureate curriculum. Nevertheless, despite the implementation of assessment procedures, it remains challenging to ascertain the genuine competence of a student. Although the judgment of the teachers and the quality of the examinations aim to provide a reliable evaluation, the reality is that there is an increasingly evident imbalance between the curriculum and the methods of its evaluation in Baccalaureate in Spain.

The use of technology and standardization of language tests for academic and professional process has been reflected in the technical reports and related research worldwide. A number of reference language tests have been developed, such as the TOEFL (USA), Cambridge Suite, IELTS, PASS (United Kingdom), or university entrance tests with L2 sections such as OSYM (Turkey), Abitur (Germany), Unified Exam (Russia), Swedish Scholastic Aptitude Test, (Switzerland), SAT (USA), Nyūgaku Shiken (Japan), just to mention a few. These tests have systematized the process of evaluating oral skills through structured computer environments that allow for massive user access simultaneously (Magal-Royo & García Laborda, 2018). In other words, the majority of educational and assessment bodies are seeking nationwide computerized tests with a high number of simultaneous users. In addition, these tests should introduce new types of items and be equally valid and much less expensive (García Laborda & Martín Monje, 2013). Furthermore, Goertler and Gacs (2018) state that a slight but growing advantage is seen in the use of computers compared to traditional exam tasks such as writing due to the normalization of keyboard use.

Innovation in language assessment must be seen in the light of, at least, three categories: (a) items or tasks; (b) design, composition and delivery; and (c) innovations and personal factors (García Laborda & Fernández Álvarez, 2021). In educational contexts, language tests should consider measuring linguistic competence as well as 21st-century skills, although their evolution has yet to be studied in light of the use of technology in 2020 due to the COVID-19 Pandemic (García Laborda & Amengual Pizarro, 2022). Regarding the EVAU, which is the exam with the greatest impact in the Spanish educational system, it is difficult to justify the disinterest of Ministry of Education in improving the current systems to assess the command of foreign language in the university entrance examination as a whole.

Exams, especially those associated with standardized certifications, particularly those provided by Cambridge Suite exams and to a lesser extent Langcert, APTIS or Linguaskill, should provide educational agents, including administrators of programs, parents, and potential employers, with the assurance that students meet performance standards that ensure the required level of proficiency in a language for academic or employment purposes. To this end, the EVAU tasks are ideally based on real materials. Each student must demonstrate the L2 linguistic ability for L2 use in an academic context, which is the particular case of the TOEFL exam – either computer-based or paper-based. It is therefore essential that the results of the exams are communicated effectively to the Ministries of Education - and the rest of the educational agents – to provide an insight if, at a global or specific level, progress is being made in the teaching of a Foreign Language and whether standards have been reached that are necessary for the performance of certain work or academic tasks.

3. Literature Review

3.1. Chatbots in Language learning and Teaching

In the educational field, chatbots have recently been used in education to motivate students in scientific learning or to help instructors to manage activities such as instructions (Okonkwo & Ibijola, 2021; Kumar, 2021). Regarding foreign languages, chatbots have drawn the attention of language teaching researchers for their ability to communicate with users in the target language (Fryer et al., 2021).

Chatbots for language assessment have four characteristics of enormous interest for this research: 1) they are available and have enormous flexibility, so they can be used at any time in the weeks/months prior to the test; 2) students can practice their language skills with chatbots whenever they want for test preparation - something that a human partner could not easily do (Haristiani, 2019; Winkler & Soellner, 2018); 3) in certain situations, chatbots can provide students with much broader input than their schoolmates when acting in an L2: 4) likewise, chatbots facilitate continuous practice, freeing the L2 teacher from repetitive work and increasing the possibilities of interaction in L2 practice

Chatbot-assisted language learning refers to a method of using human-machine interaction to facilitate daily language practice for students through natural language conversation (e.g., conversation practice, Fryer et al., 2017), answering related questions with language learning (e.g., reading storybooks), allowing to conduct assessments (although we believe this aspect is still underdeveloped) and providing feedback (Martín Mazón, 2021). For example, through a vocabulary test (Jia et al., 2012), recently, artificial intelligence and machine learning techniques have shown that it is possible to improve the ability of chatbots to adapt to unstructured input from end users, as it is often the case in the teaching of foreign languages. It is evident that this approach offers a potential solution to the need for an online evaluation system, with certain limitations. When considered together with other measures already proposed by other researchers (Peñate Cabrera, 2014), it has the potential to solve the serious deficiency of the oral evaluation in the university entrance exam.

Active dialogue and an immersive environment in language assessment are drivers that favor communicative competence and improve production in an oral English exam. A more interactive and authentic language environment enabled by chatbot-supported activities can improve L2 assessment results, highlighting the potential of chatbots to reduce the shyness that some students may feel during language practice when compared to speaking with a human partner. Chatbots, however, can help reduce the emotional distance in the automated oral assessment system by providing a dialogue for the student to interact with the required tasks. Many studies affirm the immense potential of chatbots in reducing student anxiety (Ayedoun et al., 2015) and promoting motivation for being something new, which improves short-term performance (Ayedoun et al., 2019). The importance of this normalization lies in the fact that it eliminates the bias introduced by factors other than linguistics in the teaching of languages through technology. Perhaps the biggest concern in the use of chatbots in the oral evaluation (not so much in the written one, which is much more developed) is their ability to handle complex structures / tasks, particularly when the student has a distinctive pronunciation that differs from that of the native speaker. It is therefore important to recognize the limitation in the chatbot's ability to process several complex sentences concurrently as this differs from human-human interaction in a real language learning context. Nonetheless it is to be assumed that AI will provide larger corpora each time favoring machine learning and improving communication with end users (in this case, test examinees). To implement the use of chatbots more effectively, it is crucial to know how chatbots have been used for current language learning and what improvements could be incorporated into the assessment of chatbot-compatible languages.

In general, recent progress in chatbot research has focused on the particular features which permit the implementation of chatbots in language assessment. For example, Haristiani analyzes the different types of chatbots used in language learning and concludes that Cleverbot and Mondly are the most widely used chatbot applications (Haristiani, 2019; Fryer et al., 2020) since they are flexible and have a great capacity to learn from interaction with humans. Haristiani (2019) also proposes a series of improvements, as indicated in the project by Kim who carried out a similar study and found that the lack of meaningful interaction was the most significant issue for the direct use of chatbots for speech recognition and conversation building for language learning which would greatly complicate their use in exams today (Kim et al., 2021).

4. Methodology

In order to observe the potential of chatbots for language assessment, this paper employs a two stages process. The first stage utilizes a SWOT analysis to indicate the pros and cons as well as opportunities and threats associated with their use. Some insights from this analysis have already been addressed in the previous sections. The second stage relates the implementation from a technical perspective.

A SWOT analysis is a strategic planning tool used to evaluate the strengths, weaknesses, opportunities, and threats. In this context, it is used to outline the most significant features of this implementation, in relation to our use of chatbots for language assessment. The analysis provides insights into the internal and external factors that could affect the success of this use and, subsequently, the validity of the results, including features such as reliability, fairness and practicality.

5. A SWOT analysis of Chatbots for the University Entrance Examination

In the field of applied linguistics and language teaching, SWOT analyses offer valuable insights for evaluating and implementing technological tools and innovations in language education (Table 1). As digital tools such as the one this study addresses become increasingly integrated into the field of applied linguistics, this paper will evaluate the effectiveness and suitability of chatbots for the EVAU according to the researchers.

5.1 Strengths

Regarding the strengths of the SWOT analysis, the researchers identified four primary aspects. Advances in chatbot technology, particularly through voice-to-text applications, have significantly enhanced the development of oral chatbots. These applications have become increasingly effective, especially in voice recognition, which facilitates accurate transcription and response generation. Partially utilizing this approach, e-rubrics and AI-driven scoring systems can provide timely feedback, enabling conversations that closely mimic real-life interactions. Second, the study highlights the use of user-friendly, standardized interfaces with avatars that closely resemble real speakers. Additionally, improvements in e-rating have made automated evaluations more consistent with human assessments thanks to the enhancement of existing applications. Finally, the increased accessibility of these technologies allows for usage across various platforms—such as iOS, Android, and web browsers—under supervised conditions.

5.2 Weaknesses

The weaknesses associated with implementing chatbots are primarily related to their current availability. Chatbots require significant financial investment to maintain and enhance their performance capabilities. Without ongoing efforts, chatbots risk becoming obsolete relatively quickly. However, this limitation can be mitigated through collaboration with external organizations, such as Cambridge Assessment or the British Council. The second issue pertains to the limited language support provided by chatbots; only languages with commercially viable industries (e.g., English—as examined in this study—Spanish, and Standard Chinese [Mandarin]) are typically supported. A third potential weakness is the overreliance on technology, which includes the dependence on databases and central processing units, delivery engines, and secure file protection systems. The final challenge involves accessibility in test examinee training, as equivalent chatbots should be widely available. In this regard, initiatives like "Speak & Improve" (Cambridge Assessment) can aid candidates in adequately preparing for tests. Additionally, it is worth mentioning that this availability of chatbots has led to the normalization of interactions with avatars in semi-real environments (all the conversations in a regular test are not real, i.e. not fully authentic, but semi-real).

5.3 Opportunities

The most significant opportunity lies in enhancing the speaking proficiency of Spanish students, particularly given the current educational context. The speaking skill is often underemphasized due to the strong washback effect of the University Entrance Examination in Spain. Including speaking as a component in this examination could provide substantial benefits to students and lead to transformative changes in the teaching of English. An increased emphasis on speaking would not only advance research in language acquisition but also promote the development of advanced language tools that facilitate both learning and communication. Furthermore, advancements in technology, particularly through artificial intelligence and machine learning, could further improve the capabilities of these applications.

5.4 Threats

The most significant threat is the limited availability of training platforms. For chatbots to be effectively implemented for educational and assessment purposes, they must be accessible to all students; otherwise, the integrity of the assessment process may be compromised. A second issue is the possibility that the educational boards may fail to acknowledge the importance of this assessment method, which could result in insufficient funding for continuous progress and system updates. Furthermore, rapid technological advancements could quickly render any application or platform obsolete, and unforeseen legal changes may also pose additional challenges.

Table 1. *Summary of the SWOT Analysis for the Use of Chatbots in Language Testing*

STRENGTHS (+)	WEAKNESSES (-)
Advances in chatbot technology	High financial cost
Enhanced Voice-to-Text Technology and voice recognition	Limited institutional support
AI-Driven Scoring and Feedback	Dependence on technology
Positive Washback in language learning	Accessibility for training
Development of User-Friendly Interfaces	Normalization of use and interaction
Improved Automated Evaluations	
Increased Accessibility across multiple platforms	
OPPORTUNITIES (+)	THREATS (-)
Enhancing speaking proficiency of Spanish students.	Limited availability of training chatbots
Addressing the underemphasis on speaking skills due to the University Entrance Examination's washback effect	Accessibility issues compromising the integrity of the assessment process
Including speaking in the University Entrance Examination	Lack of recognition by educational boards of the importance of chatbot assessments
Advancing research in language acquisition and assessment	Insufficient financial investment for continuous progress and system updates
Promoting the development of advanced language tools	Rapid technological advancements potentially rendering applications or platforms obsolete
Leveraging technology advancements in AI and machine learning for language learning	Unforeseen legal changes posing additional challenges

6. Chatbot for Language Testing Design

The previous section looked at previous research on the subject. It is also necessary to define the function and context of a chatbot for language testing. According to Weir, the validity of a way of delivering a test or a section of a test would be related to Context Validity (Weir, 2005; Darabi Bazvand & Ahmadi, 2020), which includes other alternatives like face-to-face testing, use of mobiles and laptops, use of social networks, instant messaging tools, and so on. One of the rules to validate the means of delivery would be that there was a state of normalization (flow) where test candidates do not experience significant differences across different delivery means, since otherwise we would be assessing additional cognitive factors (i.e., computer skills).

Therefore, an adapted chatbot, with a series of preconfigured rules based on levels of grammar and/or syntactic skills, could be considered an appropriate tool to semi-automatically assess a student's ability in oral mediation and the development of a conversation under a previous context. The use of chatbot as a language learning tool would require a prior explanatory context that would be used in an oral listening and speaking test to give way to the conversation within a comfortable and recognizable interactive environment for the student. The communicative value of chatbots would allow the student a training oriented towards their own learning that would be reconfigured based on the knowledge learned by the student over time, since chatbots provide feedback and adapt to all conversations established with the student.

As observed in Figure 1, the following four main elements must be considered when developing an adapted chatbot: 1) language learning and assessment environment; 2) type of interaction; 3) type of task, and 4) language context. One further possibility is to use a chatbot as a tool included in a language exam, starting from the previous knowledge that the student already has recognized (level A1, level B2, etc.) (CEFR, 2018) and taking into account their knowledge of digital skills, currently aligned with the European Framework of digital skills DIGCOMP (Vuorikari et al., 2016; Vuorikari et al., 2022) and based on the security of a digital certification for citizens (European Commission, 2021).

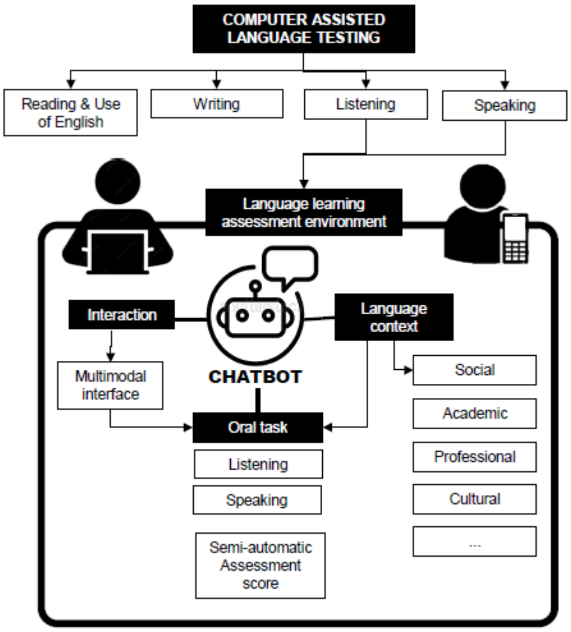


Figure 1. *Elements to be Consider in an Adapted Chatbot for Language Learning and Assessment in an Oral Task*

7. Chatbots and the Need for Effective Multimodal Communication

Chatbots were conceived as a means of developing a human-machine communication interface that could effectively simulate and replicate a dialog that was both intelligent and engaging. The primary type of communication focuses on a text created by a keyboard entry and an answer by playing a text on a screen (Nguyen et al., 2021).

With the technological advancement of digital devices, the modes of interaction, and the increasingly sophisticated artificial intelligence systems, we can find two types of chatbots that are well differentiated depending on their way of working. The first type is the simple chatbot that works based on a series of previously created commands and keywords and where the conversation is limited to using the keywords to rebuild the conversational phrase with a specific protocol. It uses a pattern-based language such as those created with Artificial Intelligence Markup Language (AIML) or use algorithms in combination with patterns to create a hierarchical structure that allows both questions and responses to be debugged and categorized. The second category comprises intelligent chatbots based on artificial neural networks that provide feedback from user conversations and become increasingly proficient based on the knowledge acquired on a specific subject or area of knowledge. Neural networks are a form of control and evaluation of the interaction in which the system is automatically improved to find the best answer to a specific question or query.

The communication system of a chatbot, known as conversational interaction, is based on management of an information processing system generated by the human user as input and an information processing system generated by the digital device as output (Følstad & Brandtzæg, 2017). The conversational interfaces used in a conversational chatbot are of two types: 1) a conversational user interface (CUI) that uses natural language, either in the form of text or a combination of text and speech; or 2) a voice user interface (VUI) that allows you to interact with the chatbot using only your voice. In fact, the CUI type interface can include the VUI type interface. Curiously, most of the chatbot proposals at a commercial level today are designed with the objective of retaining a single input channel, for example, the keyboard or the voice recording and recognition, and a single output channel, such as text displayed on the screen. The screen or the response of a synthetic voice is not considered in relation to the fact that in human conversation, there is a synchronization of multiple interaction modes that enrich the communication of what is transmitted. A chatbot must recognize a question and generate a response that is as humane as possible (Nordberg & Guribye, 2023) (see Figure 2).

This fact requires knowledge of several natural language principles focused on three key technologies, Natural Language Processing (NLP), Natural Language Understanding (NCL), and Natural Language Generation (NGL). The NLP breaks the user's query into sentences or words, allowing text to be standardized, spelling corrected, words tagged, and even to recognize sentiment factors. The NCL establishes the recognition and use of recognizable lexicons, synonyms, themes, or specific words to find the best answer. For this, algorithms are used which construct dialogue flows that indicate the most accurate response that the chatbot will issue. The NGL creates the most accurate response based on a natural language according to the established communication context. The latest generation of chatbots, based on conversational Artificial Intelligence, allow the development of sophisticated communication interfaces, based on the available database, personal preferences, and contextual understanding that help to formalize a natural and realistic dialogue (Börsting & Hesenius, 2021).

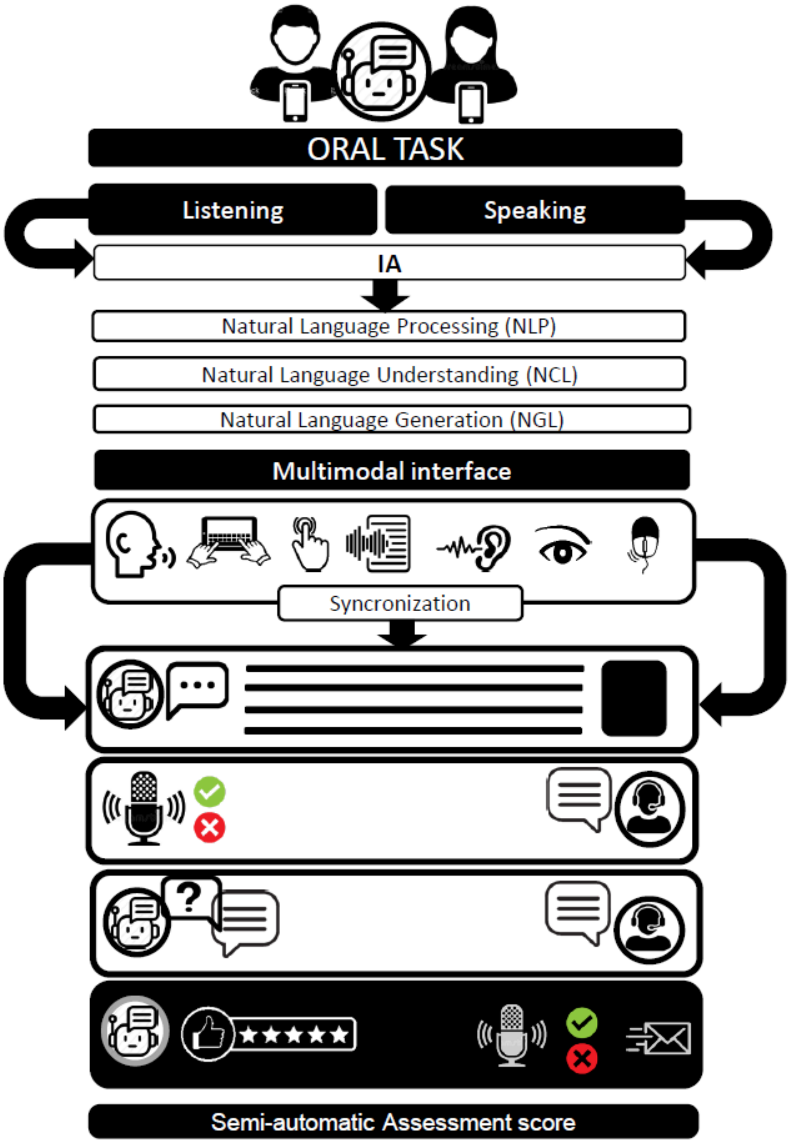

ORAL TASK
Listening
Speaking
IA
Natural Language Processing (NLP)
Natural Language Understanding (NCL)
Natural Language Generation (NGL)
Multimodal interface
Syncronization
Semi-automatic Assessment score

Figure 2. *Chatbot Framework for Assessment in Language Learning in an Oral Task*

8. Conclusion

Chatbots can be used in a variety of ways in language testing. One common use is for oral exams, where the chatbot simulates a conversation with the user and evaluates their listening and speaking skills. Written exams can also be administered through chatbots with the assessment tool that considers the user´s grammar, vocabulary and writing style. Furthermore, in addition to the administering test, chatbots can provided personalized feedback to students. They can identify areas where the students need improvement and provide targeted exercises and resources to help them improve their language learning skill. The SWOT analysis identified significant positive aspects in chatbot technology, such as improved voice-to-text capabilities and AI-driven scoring systems, which enable realistic conversations and automated feedback. User-friendly interfaces and increased accessibility across multiple platforms were also highlighted. On the other hand, weaknesses involve the high financial cost required to maintain and update chatbots, limited language support, overreliance on technology, and the need for widespread training platforms. Opportunities lie in improving speaking proficiency among Spanish students by incorporating speaking components into assessments, which could revolutionize language teaching and foster technological advancements in AI and machine learning. External threats comprise limited availability of training resources, insufficient financial support from educational boards, rapid technological changes that could make current platforms obsolete, legal uncertainties, and competition from established language assessment providers. The analysis underscores both the potential and the challenges of integrating chatbots into language education.

From a technical perspective, some chatbots facilitate gaming experiences, while others use gamification techniques to make language learning more engaging and enjoyable. There are also some limitations to their use in language testing. One of the biggest limitations is their inability to understand certain accents or nuances of language. This can lead to inaccurate evaluation and frustration for students. Another limitation is the potential for cheating. Since chatbots are automated, it can be difficult to ensure that users are not receiving outside help during the test. Finally, chatbots may be not suitable for all types of testing, such as tests that require physical interaction or non-verbal communication.

Despite these limitations, it is believed that chatbots are an important tool for language educators and have the potential to significantly enhance not only language learning outcomes but also serve as an assessment language for self-learning.

Conflict of Interest

The authors declare no conflict of interests.

References

Ayedoun E., Hayashi Y., & Seta, K. (2015). A Conversational agent to encourage willingness to communicate in the context of English as a foreign language. *Procedia Computer Science, 60*(1), 1433-1442. https://www.doi.org/10.1016/j.procs.2015.08.219

Ayedoun, E., Hayashi, Y., & Seta, K. (2019). L2 learners' preferences of dialogue agents: A key to achieve adaptive motivational support? In Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., & Luckin, R. (Eds.), *Artificial Intelligence in Education. AIED 2019. Lecture Notes in Computer Science, 11626* (pp. 19-23). Springer, Cham. https://www.doi.org/10.1007/978-3-030-23207-8_4

Börsting, I., & Hesenius, M. (2021). Towards a systematic approach for chatbot development in digital work environments. In Klumpp, M., & Ruiner, C. (Eds.), *Digital Supply Chains and the Human Factor. Lecture Notes in Logistics* (pp. 79-94). Springer, Cham. https://www.doi.org/10.1007/978-3-030-58430-6_5

Darabi Bazvand, A., & Ahmadi, A. (2020). Interpreting the validity of a high-stakes test in light of the argument-based framework: Implications for test improvement. *Journal of Research in Applied Linguistics, 11*(1), 66-88. https://www.doi.org/10.22055/rals.2020.15417

Cerdeira J. M., Catela Nunes L., Balcão Reis A., & Seabra, C. (2018). Predictors of student success in higher education: Secondary school internal scores versus national exams. *Higher Education Quarterly*, *72*(4), 303-313. https://doi.org/10.1111/hequ.12158

Díez-Arcón, P., & Martin-Monje, E. (2023). Language teacher development in computer-mediated collaborative work and digital peer assessment: An innovative proposal. *Journal of Research in Applied Linguistics*, *14*(2), 40-54. doi: 10.22055/rals.2023.44054.3080

European Comission (2021). *Proposal for a regulation of the European Parliament and of the Council amending Regulation (EU) No 910/2014 as regards establishing a framework for a European Digital Identity*. EUR-Lex, European Digital Identity Regulation. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0281

Fernandez Alvarez, M., García Laborda, J., & Magal-Royo, T. (2022). Subrepresentación del constructo en exámenes estandarizados de lengua extranjera en España: propuesta de examen asistido por ordenador. *Porta Linguarum Revista Interuniversitaria De Didáctica De Las Lenguas Extranjeras, Monográfico 2022*, 27–45. https://doi.org/10.30827/portalin.vi.21393

Finney, S. J., Myers, A. J., & Mathers, C. E. (2018) Test instructions do not moderate the indirect effect of perceived test importance on test performance in low-stakes testing contexts. *International Journal of Testing*, *18*(4), 297-322. https://doi.org/10.1080/15305058.2017.1396466

Fryer, L. K., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language, Learning and Technology, 24*(2), 8-22. http://hdl.handle.net/10125/44719

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners. *Computers in Human Behavior 75*, 461–468. https://doi.org/10.1016/j.chb.2017.05.045

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *Interactions, 24*(4), 38–42. https://www.doi.org/10.1145/3085558

García Laborda, J., & Amengual Pizarro, M. (2022). Revisando el mundo de la evaluación lingüística. Análisis DAFO de 2020. In Romera Ciria, M., & Camino Bueno Alastuey, M. (Coord.) *Didáctica de la lengua, multimodalidad y nuevos entornos de aprendizaje* (pp. 265-278). Graó Editors, Barcelona, Spain.

García Laborda, J., & Fernández Álvarez, M. (2021). Multilevel language tests: Walking into the land of the unexplored. *Language Learning & Technology, 25*(2), 1–25. http://hdl.handle.net/10125/73428

García Laborda, J., & Martín-Monje, E. (2013). Item and test construct definition for the new Spanish baccalaureate final evaluation: A proposal. *International Journal of English Studies, 13*(2), 69-88. https://doi.org/10.6018/ijes.13.2.185921

Goertler, S., & Gacs, A. (2018). Assessment in online German: Assessment methods and results. Die Unterrichtspraxis. *Teaching German, 51*(2), 156–174. https://www.jstor.org/stable/90026423

Guapacha Chamorro, M. E. (2022). Cognitive validity evidence of computer-and paper-based writing tests and differences in the impact on EFL test-takers in classroom assessment. *Assessing Writing, 51.* https://www.doi.org/10.1016/j.asw.2021.100594

Haristiani, N. (2019). Artificial intelligence (AI) chatbot as language learning medium: An inquiry. *Journal of Physics: Conference Series, 1387, International Conference on Education, Science and Technology, 13–16 March 2019, Padang, Indonesia*. https://www.doi.org/10.1088/1742-6596/1387/1/012020

Harley, J. M., Mantou Lou, N., Liu, Y., Cutumisu, M., Daniels, L. M, Leighton, J. P. & Nadon, L. (2021). University students' negative emotions in a computer-based examination: the roles of trait test-emotion, prior test-taking methods and gender. *Assessment & Evaluation in Higher Education, 46*(6), 956-972. https://doi.org/10.1080/02602938.2020.1836123

Huang W, Hew K.F. & Fryer L.K. (2021). Chatbots for language learning. Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning, 38*(1), 237-257. https://doi.org/10.1111/jcal.12610

In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System, 34*(3), 317-340. https://doi.org/10.1016/j.system.2006.04.005

Jia, J., Chen, Y., Ding, Z., & Ruan, M. (2012). Effects of a vocabulary acquisition and assessment system on students' performance in a blended learning class for English subject. *Computers & education, 58*(1), 63-76. https://doi.org/10.1016/j.compedu.2011.08.002

Kim, H. S., Kim, N. Y., & Cha, Y. (2021). Is it beneficial to use AI chatbots to improve learners' speaking performance? *The Journal of Asia TEFL, 18*(1), 161-178. https://www.doi.org/10.18823/asiatefl.2021.18.1.10.161

Kumar, J. A. (2021). Educational chatbots for project-based learning: investigating learning outcomes for a team-based design course. *International Journal of Education Technology High Education*, *18*(65). https://www.doi.org/10.1186/s41239-021-00302-w

Lorenzo Moledo, M., Argos González, J., Hernández García, J., & Vera Vila, J. (2014). El acceso y la entrada del estudiante a la universidad: situación y propuestas de mejora facilitadoras del tránsito = Access and student entrance to the University: status and improvement proposals facilitating transit. *Educación XXI: Revista de la Facultad de Educación, 17*(1), 15-38. https://doi.org/10.5944/educxx1.17.1.9951

Magal Royo, T., & García Laborda, J. (2018). Standardization of design interfaces applied to language test on-line through ubiquitous devices. *International Journal of Interactive Mobile Technologies (iJIM), 12*(4), 21-31. https://doi.org/10.3991/ijim.v12i4.9197

Magal Royo, T., & García Laborda, J. (2022). *Communicative competence of mediation assessment language learning through the use of chatbots. EDULEARN22 Proceedings*. 14th International Conference on Education and New Learning Technologies, 4-6 July 2022 Palma, Spain, 3463-3469. https://www.doi.org/10.21125/edulearn.2022.0846

Martín Mazón, R. (2021). A chatbot on syntactic issues: A proposal for innovative help in the first year of baccalaureate classroom. *Alcalibe: Revista Centro Asociado a la UNED Ciudad de la Cerámica, 21*, 83-110. http://www.alcalibe.es/images/Alcalibe_21/un_chabot_sobre_cuestiones_sintacticas.pdf

Nguyen, Q., Sidorova, A., & Torres, R. (2021). User interactions with chatbot interfaces vs. Menu-based interfaces. *Computers in Human Behavior, 128*. https://www.doi.org/10.1016/j.chb.2021.107093

Nordberg, O. E., & Guribye, F. (2023). Interacting with the news through voice user interfaces. In Følstad, A., et al. (Eds.), *Chatbot Research and Design. CONVERSATIONS 2022. Lecture Notes in Computer Science, 13815*. Springer, Cham. https://doi.org/10.1007/978-3-031-25581-6_2

Okonkwo, C. W., & Ibijola, A. A (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence, 2*. https://doi.org/10.1016/j.caeai.2021.100033

Olani, A. (2009). Predicting first year university students' academic success. *Electronic Journal of Research in Educational Psychology, 7*(3), 1053-1072. https://doi.org/10.25115/ejrep.v7i19.1351

Pahissa, I., & Tragant, E. (2009). Grammar and the non-native secondary school teacher in Catalonia. *Language Awareness, 18*(1), 47-60. https://doi.org/10.1080/09658410802307931

Peñate Cabrera, M. (2014). Choosing a speaking test in English as a foreign language for the university entrance exam. *Didáctica (lengua y literatura), 26*, 377-400. http://hdl.handle.net/11162/179007

Poehner, M. E., & Lantolf, J. P. (2023). Advancing L2 dynamic assessment: Innovations in Chinese contexts. *Language Assessment Quarterly, 20*(1), 1-19. https://doi.org/10.1080/15434303.2022.2158465

Poehner, M. E., & Leontjev, D. (2020). To correct or to cooperate: Mediational processes and L2 development. *Language Teaching Research, 24*(3), 295-316. https://doi.org/10.1177/1362168818783212

Ross, S. J., & Okabe, J. (2006). The subjective and objective interface of bias detection on language tests. *International Journal of Testing, 6*(3), 229-253. https://doi.org/10.1207/s15327574ijt0603_2

Stenlund, T., Lyrén, P. E., & Eklöf, H. (2018). The successful test taker: exploring test-taking behavior profiles through cluster analysis. *European journal of psychology of education, 33*(2), 403-417. https://doi.org/10.1007/s10212-017-0332-2

Soodmand Afshar, H. (2020). Test-takers' perceptions of paired speaking tests and the role of interlocutor variables in pairing. *Journal of Research in Applied Linguistics*, *11*(1), 89-123. doi: 10.22055/rals.2020.15418

Vuorikari, R., Punie, Y., Carretero Gomez, S., & Van Den Brande, G. (2016). *DigComp 2.0: The digital competence framework for citizens. Update phase 1: The conceptual reference model. EUR 27948 EN*. Luxembourg, Publications Office of the European Union. https://www.doi.org/10.2791/11517

Vuorikari, R., Kluzer, S. & Punie, Y., (2022). *DigComp 2.2: The digital competence framework for citizens with new examples of knowledge, skills and attitudes, EUR 31006 EN*. Luxembourg, Publications Office of the European Union. https:///www.doi.org/10.2760/490274

Weir, C. J. (2005). *Language testing and validation. An evidence-based approach*. Research and practice in applied linguistics (RPAL), Palgrave Macmillan, London. https://www.doi.org/10.1057/9780230514577

Winkler, R.., & Soellner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of Management Annual Meeting Proceedings* (Vol. 1, 15903). https://doi.org/10.5465/AMBPP.2018.15903abstract

Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing Children's reading with a conversational agent vs. a human partner. *Computers & Education, 161*. https://doi.org/10.1016/j.compedu.2020.104059

Yin, J., Goh, T. T., Yang, B., & Xiaobin, Y. (2021). Conversation technology with micro-learning: The impact of chatbot-based learning on students' learning motivation and performance. *Journal of Educational Computing Research, 59*(1), 154–177. https://doi.org/10.1177/0735633120952067