

# Conversational Interfaces in Serious Games: Identifying Potentials and Future Research Directions based on a Systematic Literature Review

Barbara Göbl<sup>1</sup><sup>a</sup>, Simone Kriglstein<sup>1,2</sup><sup>b</sup> and Helmut Hlavacs<sup>1</sup><sup>c</sup>

<sup>1</sup>Faculty of Computer Science, University of Vienna, Austria

<sup>2</sup>Faculty of Informatics, Masaryk University, Czech Republic


**Keywords:** Natural Language Interfaces, Conversational Interfaces, Conversational Agents, Chatbots, Pedagogical Agents, Serious Games, Game-based Learning, Player-Computer Interaction.


**Abstract:** Conversational interfaces have become a popular approach to human-computer interaction in recent years. While currently often employed in a business context or as personal assistants, recent advances show that their application potential goes far beyond that. The following paper presents a systematic literature review on the integration of conversational interfaces using natural language in serious games. We provide an overview of application domains, designer's motivations and reasoning. Finally, we summarize potentials and pitfalls of this technology in serious games, identify research gaps and suggest directions for further research.


## 1 INTRODUCTION

In recent years, conversational interfaces (CIs), often also referred to as chatbots, have received increased attention as a means to create engaging human-computer interactions, guaranteeing ease-of-use and promising new, exciting experiences (Zadrozny et al., 2000; Følstad and Brandtzæg, 2017). Additionally, the popularity of messaging platforms not only provides developers with a wide range of implementation options but also underlines the wide-spread use of CIs (Shevat, 2017; Smutny and Schreiberova, 2020).

Among the many application domains, educational approaches are of increasing interest. We want to explore a specific field in this area, namely serious games (SGs). This paper focuses on the implementation of serious game CIs as natural language interfaces (NLIs), which allow the players to interact with the game based on text or speech. This approach may complement game-based learning in promising new ways, such as addressing the often daunting amounts of text in SGs by adding interactive dialogue. Additionally, pedagogical chatbots might profit from an engaging game environment to invite players to interact with them. In the course of this work, we aim to answer the following research questions:

<sup>a</sup>  <https://orcid.org/0000-0001-7186-076X>

<sup>b</sup>  <https://orcid.org/0000-0001-7817-5589>

<sup>c</sup>  <https://orcid.org/0000-0001-6837-674X>

**RQ 1:** What potentials and drawbacks does the combination of natural language-based conversational interfaces and serious games hold?

**RQ 2:** What aspects of natural language-based conversational interfaces in serious games need further examination?

The following paper will present a short background on CI classification and evaluation, leading up to a closer examination in the field of SGs. We conducted a systematic literature review and provide an analysis of the application of game-based learning, conversational agents and NLIs and summarize findings from presented evaluations. We identify promising new areas, potential pitfalls and point out how future projects may contribute to further bridge the use of NLIs and game-based learning.

## 2 BACKGROUND

This section will provide a short background on conversational systems, or chatbots, reflecting both the trends described above and the focus of the literature review presented below. To allow for a systematic approach to this broad field, we examine the roles such systems undertake, i.e. their intended purpose, the modalities, i.e. types of interfaces, how they are implemented and eventually, evaluated.

Conversational interfaces may be differentiated

according to their goals, namely task and non-task oriented (Chen et al., 2017). While the former serves a specific purpose, the latter mainly aims to entertain and chat with the user. Educational conversational agents may be considered as a subcategory of task-oriented approaches (Taouil et al., 2018). Conversation may be user- or agent-driven (Følstad et al., 2018) and is implemented in different modalities: voice- or text-based modes (Shevat, 2017; Hussain et al., 2019; Piccolo et al., 2018), or even rich interfaces that integrate buttons, text and graphics to support conversations (Shevat, 2017). Eventually, designers will also have to decide on the technical foundation of the system. Next to considerations regarding technical infrastructure and implementation platform, three types of dialogue handling have been identified: rule-based approaches, retrieval-based approaches and generative-based approaches (Hussain et al., 2019; Chen et al., 2017; Adiwardana et al., 2020). Technical implementation is strongly dependant on the types of interaction, tasks and other requirements.

Evaluation of conversational interfaces spans a large spectrum of aspects and techniques and has been extensively surveyed in previous work (see e.g. (Peras, 2018; Ren et al., 2019)). Considering the claim that conversational interfaces might be the next logical step in interface design (Shevat, 2017), one may use established usability metrics such as effectiveness, efficiency and satisfaction. A recent study refines this approach and maps metrics to usability characteristics effectiveness (e.g task completion, accuracy, expert and users assessment), efficiency (e.g. task completion time, mental effort, communication effort) and satisfaction (e.g. ease-of-use, context-dependant questions, 'want to use again') (Ren et al., 2019). A further point of discrimination is the time at which the CI is evaluated: many of the mentioned evaluation techniques focus on a finished product, while some used evaluations as an intermediary step in design (Maroengsit et al., 2019). Especially in iterative processes, intermediary evaluations might help to adapt the design and be based on low- or high-fidelity prototypes, or even help to plan the implementation. As an early step in the design of conversational interfaces, so called Wizard of Oz studies may be applied. In these studies, a human (the wizard) takes over the role of the bot and controls the dialogue with the user (Dahlbäck et al., 1993).

Regarding games, dialogue and chat-based interaction have a long history, ranging from text-based adventures to dialogue-rich modern day games. However, their integration into serious games and the integration of natural language processing certainly poses

a set of new challenges and tasks for this technology. In an educational context, conversational interfaces have been implemented for various applications. For example, as a motivational tool in language learning courses (Fryer et al., 2017), to lower barriers for teenagers in sex education (Crutzen et al., 2011) or to address psychological issues, such as stress release or to remedy symptoms of depression and anxiety (Fitzpatrick et al., 2017; Huang et al., 2015). However, these approaches do not provide a game environment that embeds these conversations and are not further scrutinized in this paper.

### 3 METHODOLOGY

To identify and analyse works that implement conversational interfaces in a game setting, we conducted a systematic literature review. Following query was used: ("*chat bot*" OR "*chatbot*" OR "*chatbot*" OR "*dialogue system*" OR "*conversational agent*" OR "*conversational agents*" OR "*natural language*") AND ("*serious games*" OR "*serious game*" OR "*game-based learning*" OR "*purposeful games*"). A search in databases SCOPUS, IEEE Explore, ACM Digital Library, SpringerLink resulted in a total number of 124 papers. After elimination of duplicates and further analysis, a body of  $N = 30$  works matched the criteria of NLIs embedded in serious games. Publication years ranged from 2008 to 2020. As can be seen in Table 1, some papers described different aspects of the same NLI. Overall, the body of work described 23 individual conversational agents. Different implementations or iteration steps were treated as separate entries in Table 1, whereas papers describing different aspects of the same implementation were summarized in one entry. After closer inspection, we found that 13 of these works presented evaluations that included aspects of the integrated NLI. Thus, a total of 17 papers focused on description of design, architecture and implementation or uses cases or evaluated unrelated aspects of the serious game. Due to their similar motivations and the holistic evaluations, solutions where only parts of the conversational interface where based on natural language interaction were considered as well.

### 4 RESULTS

Below, we analyse the resulting body of work ( $N = 30$ ) regarding application domains, types of evaluation, and specific purpose of the integrated conversational interfaces. To summarize how CIs may sup-

Table 1: On overview of the analysed body of literature.

Modality	Purpose	Motivations	Application Domain	Eval	Eval.Lvl	Authors
EA	soc & comm	sim	Health, Mental Health			(Augello et al., 2016c), (Augello et al., 2016a), (Augello et al., 2016b)
EA, text	know	imm, learn	History and Cultural Heritage			(Bellotti et al., 2011)
EA	know	imm, learn	History and Cultural Heritage			(Bellotti et al., 2008)
rich	controls	eng	IT Literacy	yes	1, 2	(Berger et al., 2019)
EA, text	soc & comm	sim	General Professional Skills	yes	1	(Callejas et al., 2014)
text	assess, analyse		Microbiology	yes	2	(Carpenter et al., 2020)
voice	assess, analyse		Speech Analysis			(Corrales-Astorgano et al., 2018)
EA, text	interact	eng, mot	Politics & Law			(Economou et al., 2014), (Economou et al., 2015)
EA	know	sim, learn	Research Skills	yes	Other	(Forsyth et al., 2015)
EA			General Professional Skills			(Jepp et al., 2010)
EA	soc & comm		General	yes	Other	(Lala et al., 2019)
EA, voice		learn	Microbiology	yes	2	(Lee et al., 2010)
EA, text	know	imm	History and Cultural Heritage	yes	1, 2	(Mori et al., 2013)
EA, voice	know	imm, learn	History and Cultural Heritage	yes		(Neto et al., 2015)
text	soc & comm	sim	General Professional Skills			(Othlinghaus and Hoppe, 2016)
EA	interact	imm, ia	History and Cultural Heritage			(Panzoli et al., 2010)
EA, voice	interact	sim, imm, learn	Language Learning			(Powers et al., 2008)
text		sim	Plastics Industry	yes		(Rojas-Barahona and Cerisara, 2014)
voice	assess, analyse		Disaster Awareness		1	(Sermet and Demir, 2019)
EA, voice	controls		IT Literacy			(Stefanidi et al., 2020)
		sim	Learning Disabilities			(Taouil et al., 2018)
EA		eng, mot	Health, Mental Health			(Tironi et al., 2019)
text	assess, analyse	learn	General	yes	1	(Toma et al., 2018)
voice	assess, analyse	ia	General	yes	1	(Toncu et al., 2020)
text	controls	acc	General			(Torrente et al., 2012)
EA, text	soc & comm, assess, analyse	sim	General Professional Skills	yes		(Vaassen and Daelemans, 2010)
text	know	learn	Microbiology	yes	1, 2	(Wiggins et al., 2019)

**Modality:** EA = embodied agent; **Purpose:** soc & comm = social and communication skills, know = present knowledge, controls = language-based commands, interact = learning through interaction, assess, analyse = automated assessment and analysis; **Motivation:** sim = simulate interaction, imm = immersion learn = improve learning outcome, eng = engagement, mot = motivation, ia = improve interaction, acc = accessibility, **Evaluation Level:** 1 - reaction, 2 - learning (Kirkpatrick and Kirkpatrick, 2006)

port the goals of the serious game, we further supplement this data with the motivations the authors state for the integration of a CIs. We considered motivations if explicitly mentioned or if authors underlined these aspects in relevant related works. Purposes of the bots were categorized based on our analysis of the examined works and summarized into 5 categories describing the NLI's task within the serious game: training social and communication skills, presenting knowledge via dialogue, learning through interaction, game controls, and assessment and analysis of player-provided data. Analysed works implemented NLIs using three different modalities, namely embodied agents (EA), text-based and voice-based interfaces. Some works did not explicitly state whether their design relies on voice- or text-based interaction.

#### 4.1 Application Domains, Purpose and Motivations

The analysed body of work suggests a broad range of applications domains, ranging from health and mental health to information technology (IT) literacy. Furthermore, a few motives come up repeatedly in the argumentation for the inclusion of CIs such as the simulation of human interaction ( $n = 8$ ), improving immersion ( $n = 6$ ), improving motivation ( $n = 2$ ) and engagement ( $n = 3$ ), improve learning outcomes ( $n = 8$ ) or improving interaction ( $n = 2$ ). Several works did not include specifics on their motivation to include CIs.

CIs' tasks may directly relate to the serious game's goal, i.e. in the form of presentation of knowledge through dialogue. For example, analysed works in the sector of history and cultural heritage feature embodied agents, and respective authoring tools, that convey historical or cultural facts throughout the conversation with the player (Bellotti et al., 2008; Bellotti et al., 2008; Mori et al., 2013). In other works, the interaction itself and related observation and exploration, rather than the dialogue's content, aim to support the learning goals. This is the case for social and communicative skill training but also serious game environments that allow the player to observe, explore and interact in self-determined approaches (Powers et al., 2008; Panzoli et al., 2010). A number of reviewed works ( $n = 6$ ) also includes speech- or text-based interfaces to gather natural language data as a basis for assessment and analysis, regarding e.g. the depth of reflection. Furthermore, NLIs often fulfill tasks related to game progress and navigation, e.g. by handling natural language player commands ( $n = 3$ ).

Provided details on the technical implementations varied widely in the examined works, thus complicat-

ing overall comparison. However, stated techniques included AIML (Artificial Intelligence Markup Language<sup>1</sup>) (Augello et al., 2016c; Othlinghaus and Hoppe, 2016), text-retrieval based approaches and metrics (e.g. ITF-IDF, (Bellotti et al., 2011; Mori et al., 2013)) or the use of external solutions and service providers such as Google Dialogflow<sup>2</sup> (Tironi et al., 2019), wit.ai<sup>3</sup> (Toncu et al., 2020) or the Microsoft Bot Framework<sup>4</sup> (Berger et al., 2019). For automated assessment and analysis, toolkits such as scikit-learn<sup>5</sup> (Carpenter et al., 2020) and openSMILE<sup>6</sup> (Corrales-Astorgano et al., 2018) were used.

#### 4.2 Evaluation and Findings

Most papers present an evaluation of the overall experience and do not look at the CI's and serious game's aspects separately. Thus, we based our analysis of the level of evaluation on the model of training evaluation (Kirkpatrick and Kirkpatrick, 2006).

Evaluations focusing on Kirkpatrick's first level, which relates to the players' "reaction" to the serious game, mostly concentrated on feedback regarding usability and satisfaction of the overall serious game. Evaluation of affective interaction, emotion classification and inducing emotion suggested that NLIs support the cause of affective interaction. One study found that combined verbal and non-verbal behaviour of embodied agents successfully conveyed social attitudes during virtual job interviews: In order to challenge and comfort the player, the embodied agent and its CI managed to support the portrayal of friendly, neutral and hostile attitudes (Callejas et al., 2014). By manipulation affect in pedagogical agents' discourse, simulating different moods, researchers managed to find a correlation between learner's arousal levels and learning outcomes (Forsyth et al., 2015). Additionally, pointing out the relevance of emotional aspects, different machine learning algorithms have been tested to identify optimal approaches to emotion classification based on natural language input (Vaassen and Daelemans, 2010). A model trained with manually annotated emotion classifications for sentences helped to identify a memory-based approach, TiMBL (Daelemans et al., 2005), as best machine learning solution to automate this task. Similarly, in an assessment of reflection-level of natural language input, written in-game, the potential of

<sup>1</sup><http://www.aiml.foundation/>

<sup>2</sup><https://cloud.google.com/dialogflow>

<sup>3</sup><https://wit.ai/>

<sup>4</sup><https://dev.botframework.com/>

<sup>5</sup><https://scikit-learn.org/stable/>

<sup>6</sup><https://www.audeering.com/opensmile/>



various algorithms to accurately determine reflection level in natural language data is demonstrated (Carpenter et al., 2020). Usability tests in regard to the overall game setting including CIs report positive attitudes and interest towards the chosen learning approaches (Mori et al., 2013; Toma et al., 2018) and report feedback commending entertainment factor and the innovative approach (Toncu et al., 2020). However, these works also demonstrate some drawbacks of the approach: for one, participants in a case study including automated assessment reported that they lack feedback to understand the automated process (Toma et al., 2018). For another, around 50% of participants in one study reported negative feelings about mismatched answers and a system's failure to correctly interpret commands (Toncu et al., 2020). In the game "Communicate", a system to provide highlighting of suggested dialogue options based on natural language input and hints in case of no-match scenarios was tested and found to have little effect on user behaviour and outcome (Lala et al., 2019). However, the authors find that providing open-text input possibilities increased the interaction with the system. "PrivaCity", which provides a mostly text-based interaction based game, reports some of the common issues that serious games face: user characteristics, such as game play frequency and game play skills, may strongly influence user satisfaction in that frequent players report lower satisfaction scores (Berger et al., 2019). Nevertheless, the authors report positive learning outcomes in self-report studies. A further proof-of-concept has been delivered by means of a Wizard of Oz study that finds a positive impact of NLI on learning (Lee et al., 2010). Another preliminary study cautiously supports the claims of improved learning and good satisfaction ratings (Neto et al., 2015) and points out that this approach shows better learning than presentation-based approaches. However, in another study comparing reading to a serious game implementing an intelligent dialogue system, the latter showed lower rankings in user reported learning outcomes and equal rankings in expert evaluations (Mori et al., 2013). Finally, a study investigating mixed initiative vs. no agent initiative settings showed that the former resulted in higher engagement but there was no overall difference in learning or frustration scores (Wiggins et al., 2019).

## 5 DISCUSSION

This paper aims to provide insights into the underpinnings of natural language interfaces in serious games. While several works stress the potential of this tech-

nology, a number of aspects needs further investigation. Below, we first aim to point out our findings in regard to RQ 1: "What potentials and drawbacks does the combination of natural-language-based conversational interfaces and serious games hold?". Thus, we discuss promising aspects of CIs in serious gaming environments but also to identify what issues need resolving. Subsequently, we address RQ 2, "What aspects of natural language-based conversational interfaces in serious games need further examination?", by identifying areas that have not been sufficiently examined and may be promising for future research.

### 5.1 Identified Potentials and Drawback of Serious Game Chatbots

Below, we aim to summarize findings and applications of NLIs in games and point out what potentials are identified in previous literature. Similar to Brandtzaeg's analysis of chatbot user motivations (Brandtzaeg and Følstad, 2017), reaching from productivity or entertainment to social motivations and curiosity, one might take a closer look on the motivations for serious game designers to include these systems. As pointed out previously, dialogue is a promising way to enhance text-based learning with more interactivity and reduce the stigma of supposedly boring reads (Bellotti et al., 2011). Additionally, the high amount ( $n = 16$ ) of implementations in the form of embodied agents further adds to the potential gain of natural language based interaction in serious games. This may be attributed to previous findings such as the 'embodied agent effect' (Atkinson, 2002) which reports increased motivation and learning through social agents in the form of faces or embodied agents that also add non-verbal cues. These agents not only improve interaction but support social elements as well (Augello et al., 2016c). A fact that is underlined by the various implementations tasking the agents, or other NLIs, in the above described studies with training social and communicative skills or simulating human-human interaction. Thus, this technique provides a safe environment for professional training and enables researchers and companies to simulate otherwise costly human resources in training and experiments (Forsyth et al., 2015; Augello et al., 2016b). A further interesting application for research is demonstrated by works using these techniques for automated assessment and analysis (Carpenter et al., 2020; Corrales-Astorgano et al., 2018; Vaassen and Daelemans, 2010). This is also promising for adaptive systems, such as intelligent tutoring systems.

However, our analysis also finds some drawbacks and pitfalls when integrating NLIs in serious games.

Especially in assessment tasks, the provided feedback is not always comprehensible for players (Toma et al., 2018), potentially hindering acceptance. Furthermore, no-match scenarios, in which the system cannot identify a matching answer, or mismatch scenarios represent an additional challenge. This is reflected in above discussed findings (Toncu et al., 2020), where 50% of participants report problems during interaction with the voice-based chatbot that failed to interpret commands correctly or did not follow them as expected. Previous literature also points out, that users' initially enthusiastic engagement with bots is often thwarted quickly if expectations are not met (Piccolo et al., 2018). Especially considering the many motivations discussed above, ranging from realistic virtual actors to improved immersion, these scenarios might interfere with the game's success - on both the level of usability and learning outcome. In line with recent suggestions (Jain et al., 2018), it might be helpful to clarify in advance what tasks and interactions the CI is able to handle and what its limits are.

Last, but not least, choices regarding implementation can only be made with the available resources in mind: available platforms, provided budget and of course, if applicable, whether data-driven implementations can be built on sufficiently large data sets. Previous work mentions, that this approach might currently not always be feasible due to the lack of available training data for specific use cases (Lala et al., 2019). Thus, applications embedded in a learning context might have more difficulty to train their bots than commercial or open-domain bots. This might be addressed by continuous training or narrowing down the type of questions and answers that the bot will provide (Gonda et al., 2018).

## 5.2 Research Gaps and Future Directions for Research

Especially in the field of evaluation many opportunities present themselves to future research. In serious game evaluation, the ultimate goal is to scrutinize a game's suitability and effectiveness regarding its goals and application setting (Emmerich and Bockholt, 2016). This underlines the approach of most of the above discussed works, that do not evaluate the NLI separately but take a holistic approach. However, comparative evaluation of solutions with or without the interface might provide helpful data to consider the costs and benefits of developing CIs for serious games. Generally speaking, as the area of NLP in serious games is still a rather young field, spurred on by the constant advancements of NLP technology, it still lacks systematic classification, categorization, evalu-

ation or general guidelines for future practitioners and researchers.

## 5.3 Limitations

Due to the varying level of detail provided in the analysed body of work, we tailored our overview to a level that would allow all works to be considered and categorised. This results in rather high-level categories in our analysis which may not fully represent data in more detailed works and does not fully accommodate previously introduced evaluation categories as discussed in section 2. Furthermore, the focus on conversational interfaces and may not fully consider all relevant aspects in a complex environment such as game-based learning and interaction. Motivations to combine natural language interfaces and game-aspect might profit from a further inspection of such interfaces in entertainment games, which was beyond the scope of this work.

## 6 CONCLUSION

The presented paper allows insights into application of natural language interaction and processing in the field of serious games. The examined body of literature points out many potentials of this combination, e.g. in terms of supporting immersion, engagement or learning outcome. However, this rather young field of research might profit from a more systemic approach based in the many preceding works regarding conversational interfaces.

## ACKNOWLEDGEMENTS

Barbara Göbl is supported by a DOC-Team scholarship of the Austrian Academy of Sciences.

## REFERENCES

- Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
- Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, 94(2):416–427.
- Augello, A., Gentile, M., and Dignum, F. (2016a). Social agents for learning in virtual environments. In *International Conference on Games and Learning Alliance*, pages 133–143. Springer.

- Augello, A., Gentile, M., Weideveld, L., and Dignum, F. (2016b). Dialogues as social practices for serious games. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pages 1732–1733. IOS Press.
- Augello, A., Gentile, M., Weideveld, L., and Dignum, F. (2016c). A model of a social chatbot. In *Intelligent Interactive Multimedia Systems and Services 2016*, pages 637–647. Springer.
- Bellotti, F., Berta, R., De Gloria, A., and Lavagnino, E. (2011). Towards a conversational agent architecture to favor knowledge discovery in serious games. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*, ACE '11. Association for Computing Machinery.
- Bellotti, F., Berta, R., De Gloria, A., Primavera, L., and Zappi, V. (2008). Travel in europe: An online environment to promote cultural heritage. *The IPSI BgD Transactions on Internet Research*, 14(1).
- Berger, E., Sæthre, T. H., and Divitini, M. (2019). Privacy. In *International Conference on Informatics in Schools: Situation, Evolution, and Perspectives*, pages 293–304. Springer.
- Brandtzaeg, P. B. and Følstad, A. (2017). Why people use chatbots. In Kompatsiaris, I., Cave, J., Satsiou, A., Carle, G., Passani, A., Kontopoulos, E., Diplaris, S., and McMillan, D., editors, *Internet Science*, pages 377–392. Cham. Springer International Publishing.
- Callejas, Z., Ravenet, B., Ochs, M., and Pelachaud, C. (2014). A model to generate adaptive multimodal job interviews with a virtual recruiter. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 3615–3619.
- Carpenter, D., Geden, M., Rowe, J., Azevedo, R., and Lester, J. (2020). Automated analysis of middle school students' written reflections during game-based learning. In *International Conference on Artificial Intelligence in Education*, pages 67–78. Springer.
- Chen, H., Liu, X., Yin, D., and Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35.
- Corrales-Astorgano, M., Martínez-Castilla, P., Mancebo, D. E., Aguilar, L., Ferreras, C. G., and Cardeñoso-Payo, V. (2018). Towards an automatic evaluation of the prosody of people with down syndrome. In *Iber-SPEECH*, pages 112–116.
- Crutzen, R., Peters, G.-J. Y., Portugal, S. D., Fisser, E. M., and Grolleman, J. J. (2011). An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study. *Journal of Adolescent Health*, 48(5):514–519.
- Daelemans, W., Van den Bosch, A., et al. (2005). *Memory-based language processing*. Cambridge University Press.
- Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of oz studies: Why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces, IUI '93*, page 193–200, New York. ACM.
- Economou, D., Doumanis, I., Bouki, V., Pedersen, F., Kathrani, P., Mentzelopoulos, M., and Georgalas, N. (2014). A dynamic role-playing platform for simulations in legal and political education. In *2014 International Conference on Interactive Mobile Communication Technologies and Learning (IMCL2014)*, pages 232–236. IEEE.
- Economou, D., Doumanis, I., Pedersen, F., Kathrani, P., Mentzelopoulos, M., and Bouki, V. (2015). Evaluation of a dynamic role-playing platform for simulations based on octalysis gamification framework. In *Intelligent Environments (Workshops)*, pages 388–395.
- Emmerich, K. and Bockholt, M. (2016). *Serious Games Evaluation: Processes, Models, and Concepts*, pages 265–283. Springer International Publishing, Cham.
- Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.
- Følstad, A. and Brandtzaeg, P. B. (2017). Chatbots and the new world of HCI. *Interactions*, 24(4):38–42.
- Følstad, A., Skjuve, M., and Brandtzaeg, P. B. (2018). Different chatbots for different purposes: towards a typology of chatbots to understand interaction design. In *International Conference on Internet Science*, pages 145–156. Springer.
- Forsyth, C. M., Graesser, A., Olney, A. M., Millis, K., Walker, B., and Cai, Z. (2015). Moody agents: affect and discourse during learning in a serious game. In *International Conference on Artificial Intelligence in Education*, pages 135–144. Springer.
- Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., and 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.
- Gonda, D. E., Luo, J., Wong, Y., and Lei, C. (2018). Evaluation of developing educational chatbots based on the seven principles for good teaching. In *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pages 446–453.
- Huang, J., Li, Q., Xue, Y., Cheng, T., Xu, S., Jia, J., and Feng, L. (2015). Teenchat: a chatterbot system for sensing and releasing adolescents' stress. In *International Conference on Health Information Science*, pages 133–145. Springer.
- Hussain, S., Sianaki, O. A., and Ababneh, N. (2019). A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*, pages 946–956. Springer.
- Jain, M., Kumar, P., Kota, R., and Patel, S. N. (2018). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 895–906.
- Jepp, P., Fradinho, M., and Pereira, J. M. (2010). An agent framework for a modular serious game. In *2010 Second International Conference on Games and Virtual Worlds for Serious Applications*, pages 19–26. IEEE.

- Kirkpatrick, D. and Kirkpatrick, J. (2006). *Evaluating training programs: The four levels*. Berrett-Koehler Publishers.
- Lala, R., Jeurig, J., and van Geest, M. (2019). Scaffolding open text input in a scripted communication skills learning environment. In *International Conference on Games and Learning Alliance*, pages 169–179. Springer.
- Lee, S. Y., Mott, B. W., and Lester, J. C. (2010). Optimizing story-based learning: An investigation of student narrative profiles. In *International conference on intelligent tutoring systems*, pages 155–165. Springer.
- Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., and Theeramunkong, T. (2019). A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International Conference on Information and Education Technology*, ICIET 2019, page 111–119, New York. ACM.
- Mori, D., Berta, R., De Gloria, A., Fiore, V., and Magnani, L. (2013). An easy to author dialogue management system for serious games. *Journal on Computing and Cultural Heritage*, 6(2):1–15.
- Neto, J., Ribeiro, C., Pereira, J., and Neto, M. J. (2015). Virtual agents and multi-modality of interaction in multimedia applications for cultural heritage. In *Proceedings of the 10th International Conference on Computer Graphics Theory and Applications.*, pages 446–453. SCITEPRESS-Science and Technology Publications.
- Othlinghaus, J. and Hoppe, H. U. (2016). Supporting group reflection in a virtual role-playing environment. In *International Conference on Intelligent Technologies for Interactive Entertainment*, pages 292–298. Springer.
- Panzoli, D., Qureshi, A., Dunwell, I., Petridis, P., de Freitas, S., and Rebolledo-Mendez, G. (2010). Levels of interaction (loi): a model for scaffolding learner engagement in an immersive environment. In *International Conference on Intelligent Tutoring Systems*, pages 393–395. Springer.
- Peras, D. (2018). Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, pages 89–97.
- Piccolo, L. S., Mensio, M., and Alani, H. (2018). Chasing the chatbots. In *International Conference on Internet Science*, pages 157–169. Springer.
- Powers, D. M., Leibbrandt, R., Pfitzner, D., Luerssen, M., Lewis, T., Abrahamyan, A., and Stevens, K. (2008). Language teaching in a mixed reality games environment. In *Proceedings of the 1st international conference on Pervasive Technologies Related to Assistive Environments*, pages 1–7, New York. ACM.
- Ren, R., Castro, J. W., Acuña, S. T., and de Lara, J. (2019). Evaluation techniques for chatbot usability: A systematic mapping study. *International Journal of Software Engineering and Knowledge Engineering*, 29(11n12):1673–1702.
- Rojas-Barahona, L. M. and Cerisara, C. (2014). Bayesian inverse reinforcement learning for modeling conversational agents in a virtual environment. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 503–514. Springer.
- Sermet, Y. and Demir, I. (2019). Flood action VR: a virtual reality framework for disaster awareness and emergency response training. In *ACM SIGGRAPH 2019 Posters*, pages 1–2.
- Shevat, A. (2017). *Designing bots: Creating conversational experiences*. O’Reilly Media, Inc.
- Smutny, P. and Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the facebook messenger. *Computers & Education*, page 103862.
- Stefanidi, E., Arampatzis, D., Leonidis, A., Korozi, M., Antona, M., and Papagiannakis, G. (2020). Magiplay: An augmented reality serious game allowing children to program intelligent environments. In *Transactions on Computational Science XXXVII*, pages 144–169. Springer.
- Taouil, M., Begdouri, A., and Majda, A. (2018). Adaptive dialogue system for disabled learners: Towards a learning disabilities model. In *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, pages 422–427. IEEE.
- Tironi, A., Mainetti, R., Pezzerà, M., and Borghese, N. A. (2019). An empathic virtual caregiver for assistance in exer-game-based rehabilitation therapies. In *2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–6. IEEE.
- Toma, I., Bacioiu, F., Dascalu, M., and Trausan-Matu, S. (2018). The edutainment platform: Interactive storytelling relying on semantic similarity. In *Challenges and Solutions in Smart Learning*, pages 87–96. Springer.
- Toncu, S., Toma, I., Dascalu, M., and Trausan-Matu, S. (2020). Escape from dungeon—modeling user intentions with natural language processing techniques. In *Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education*, pages 91–103. Springer.
- Torrente, J., Marchiori, E. J., Vallejo-Pinto, J. Á., Ortega-Moral, M., Moreno-Ger, P., and Fernández-Manjón, B. (2012). Eyes-free interfaces for educational games. In *2012 International Symposium on Computers in Education (SIEE)*, pages 1–6. IEEE.
- Vaassen, F. and Daelemans, W. (2010). Emotion classification in a serious game for training communication skills. *LOT Occasional Series*, 16:155–168.
- Wiggins, J. B., Kulkarni, M., Min, W., Boyer, K. E., Mott, B., Wiebe, E., and Lester, J. (2019). Take the initiative: Mixed initiative dialogue policies for pedagogical agents in game-based learning environments. In *International Conference on Artificial Intelligence in Education*, pages 314–318. Springer.
- Zadrozny, W., Budzikowska, M., Chai, J., Kambhatla, N., Levesque, S., and Nicolov, N. (2000). Natural Language Dialogue for Personalized Interaction. *Commun. ACM*, 43(8):116–120.