The current research trend of artificial intelligence in language learning: A systematic empirical literature review from an activity theory perspective

Hongzhi Yang
The University of Sydney, Australia

Suna Kyun
Sookmyung Women's University, South Korea

Although the field of artificial intelligence (AI) has rapidly developed, there has been little research to review, describe, and analyse the trends and development of empirical research on AI-supported language learning. This paper selected and analysed 25 empirical research papers on AI-supported language learning published in the last 15 years. These empirical studies were analysed using the activity theory from seven constituents: tool, subject, object, rules, community, division of labour, and outcome. A key contribution of this paper is the use of activity theory to illustrate the dynamic interactions and contradictions between the seven elements. AI-supported technology as a mediating tool demonstrated some effectiveness in language learning but needs further improvement in the use of language for communication and collaborative design. We argue that teachers' intervention and configuration of AI-supported language learning in the pedagogical design plays an important role in the effectiveness of learning. More research is needed to explore the use of AI-supported language learning in the classroom or the real-life learning context.

Implications for practice or policy:
- Research on AI-supported language learning should view teacher and students as active agents in interacting with technology and making transformations in real life learning situations.
- More research should focus on productive dialogue and communication in AI-supported language learning with collaborative design.
- A mixed module of AI-supported language learning and formal teacher instruction should be incorporated in pedagogical design.

Keywords: artificial intelligence, language teaching, language learning, activity theory, empirical literature review, AI-supported language learning

Introduction

The rapid development of artificial intelligence (AI) has impacted the use of information and communications technology in language learning as a sub-group of computer assisted language learning. While AI has proven to enhance language teaching and learning with appropriate guidance (e.g., Al-Kaisi et al., 2021; Chew & Chua, 2020; Dodigovic, 2007), the potential benefits and problems of using AI for language learning or teaching among first and second language learners have not been investigated systematically within a theoretical framework. Researchers have synthesised various types of AI-supported language learning tools (Kessler, 2018; Pokrivčáková, 2019) and predicted some scenarios in terms of how AI may impact language teaching in the future (Godwin-Jones, 2019). However, there has not been a systemic review of the trend and patterns of empirical research in this area. This empirical review study thus aimed to fill a review gap by analysing the trend and patterns emerging from the published empirical research on AI-supported language learning from 2007 to 2021.

Regarding the need to have a theoretical framework to guide this empirical review, activity theory has been viewed as a suitable theoretical model in computer-assisted language learning to illustrate and analyse the key factors, subject, tool, object, and context, at both individual and collective levels in teaching and
learning (Engeström, 1987; Lin et al., 2019; Liu et al., 2016). This is crucial for the success of the integration of information and communication technology in language education (Cash et al., 2015; Lin et al., 2019; Liu et al., 2016). More importantly, activity theory can show the interrelationship between different sociocultural elements and levels of activity by providing a holistic view on language learning with technology (Burston, 2015; Lin et al., 2019). As a sub-group of computer-assisted language learning, AI shares some features of it and its sub-branch, mobile assisted language learning, when the AI technology is used for mobile apps. Lin et al. (2019) used the seven components of activity theory to analyse the literature of mobile assisted language learning and its underlying design principles. The purpose of this review research was to apply activity theory as the working framework to analyse the selected empirical studies of AI-supported language learning published between 2007 and 2021 to reveal the research trends in this area.

The development of AI in education

AI in education

There are variations in the definition and understanding of AI. Some definitions have focused on AI as a set of skills or abilities of a digital computer, such as “computers which perform cognitive tasks, usually associated with human minds, particularly learning and problem-solving” (Baker & Smith, 2019; p. 10), or “the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings” (Encyclopaedia Britannica, 2016; p. 14). Others have focused on the computer’s ability and intelligent behaviours in interacting with human beings: computer systems that have been designed to:

[I]nteract with the world through capabilities (for example, visual perception and speech recognition) and intelligent behaviours (for example, assessing the available information and then taking the most sensible action to achieve a stated goal) that we would think of as essentially human (Luckin et al., 2016; p. 14).

When AI is used in AI-powered education, it provides the possibility for “more personalised, flexible, inclusive, and engaging” learning and a more sophisticated learning environment (Luckin et al., 2016; p. 11). One example is the collaboration between teacher and AI-powered education. AI-powered education technology can do some tasks, such as marking a large sample of students’ work that cannot be done by one teacher alone and providing learner needed support such as instant machine feedback (Pokrivčáková, 2019).

Researchers divide AI-powered education into three user categories: (1) learner-facing, used by students to learn a subject; (2) teacher-facing, referring to some automating tools or systems, used by teachers to reduce workload and increase output, such as marking, feedback and plagiarism checking; and (3) system-facing, used by administrative staff at the institutional level for managing some figures and patterns within and across institutions, such as attrition rate (Baker & Smith, 2019; Pokrivčáková, 2019). The application of AI in education includes learning profiling and prediction, assessment and evaluation, adaptivity and personalisation, and intelligent tutoring systems (Chen et al., 2022; Zawacki-Richter et al., 2019). The affordances of AI-powered education have been widely recognised in the academic field, evidenced by several systematic literature reviews of different aspects, such as the evaluation of machine learning (Zhai et al., 2020), AI in e-learning (Tang et al., 2021), and AI in deep learning (Guan et al., 2020). One common finding was the increasing number of research outputs on these aspects of AI-powered education. In addition, researchers called for more research on the application of AI in a real classroom context (Zhai et al., 2020) and a paradigm shift in AI-powered education from the focus on education technology to “pedagogical, cultural, social, economic, and ethical aspects” of education (Guan et al., 2020; p. 143).

There have also been several comprehensive reviews on AI-powered education. For instance, Chen et al. (2020) reviewed 45 articles in terms of annual distribution, journals, frequently used terms, institutions, and countries/regions to highlight the gaps in the application and theory of AI-powered education research. Similar to Zhai et al.’s (2020) review, Chen et al. (2020) also highlighted the need to apply AI in a real classroom setting and to incorporate AI applications with education theories. Another systematic review of the use of AI in higher education research revealed that most of the research has been conducted in the disciplines of computer science and STEM (Zawacki-Richter et al., 2019). Therefore, it is necessary to
extend AI-powered education research to other disciplines, such as language education. Further, a large-scale topic-based bibliometrics review which analysed 4,519 publications related to AI-powered education from 2000 to 2019 showed that language education with a focus on natural language processing has emerged as one research topic in AI-powered education. (Chen et al., 2022). Therefore, we identified the need for a specific systematic review of the use of AI in language education.

**AI in language education**

AI-supported technology in language learning can be viewed as a subset of computer-assisted language learning. The areas relevant to AI in computer assisted language learning are natural language processing, automatic marking and feedback systems, adaptive educational systems, and intelligent tutoring systems. (Pokrivčáková, 2019; Schulze, 2008). The shift from computer-assisted language learning to intelligent computer-assisted language learning featured big data processing and machine learning algorithms and has brought a substantial change in student-computer interaction (Kannan & Munday, 2018; Pokrivčáková, 2019). The benefits included reducing time, cost and learners’ frustration and anxiety, quick assessment with instant feedback, and predicting learners’ future performance (Pokrivčáková, 2019). For example, the intelligent language tutors collected learner data to build learner corpus and learner models, and to tailor learning content based on learners’ needs and progress. Teachers and researchers can also use this learner corpus to predict learners’ performance or learning challenges (Godwin-Jones, 2019).

Meanwhile there are three challenges in the area of AI-supported language learning. The first is a relative lack of empirical research in the aspects of pedagogical effects, learners’ interaction with AI-supported language learning, and teachers’ and students’ attitudes towards AI technology (Pokrivčáková, 2019). Together with the rapid development of AI technology in education, there has been a boom in research in the area of language learning. Therefore, researchers call for more systematic reviews and empirical analysis of AI in language education (Liang et al., 2021). The second challenge is the technology barrier, such as the dialogic competence of AI, which has imposed some difficulties in applying AI in language learning (Weigand, 2019). The third challenge is overcoming people’s perceptions and fear of AI; for example, whether language learning/teaching will be needed in the future due to the development of AI (Godwin-Jones, 2019).

To date, there have been several review studies on the use of AI in language education. Some focused on specific language skills supported by AI, such as the effect of an intelligent tutorial system on reading comprehension (e.g., Xu et al., 2019). Others looked at specific technology, such as chatbot (Smutny & Schreiberoova, 2020). Yet, there have been limited comprehensive research reviews related to the use of AI in language education (Liang et al., 2021). The systematic review by Liang et al. (2021) covered the research of the last 30 years and revealed that during the period 1999 to 2006, research published in this area comprised mainly of conceptual papers with limited empirical data support. This was confirmed by the Zawacki-Richter et al. (2019) review which identified 2007 as a significant year in the development of AI in education when iPhone’s Siri was introduced. Therefore, this systematic review focused on the empirical studies using AI in language learning from the year 2007 to 2021, with the aim being to illustrate the trends in the empirical research on AI in language learning. In addition, this review only included papers published in peer-reviewed journals. The criterion of being a peer reviewed paper has been viewed as a baseline of quality for published research in a specific field (Bond et al., 2020). In addition, this review summarised the limitations of each reviewed paper to bring insights about the future research needed.

**Theoretical framework: Activity theory**

Activity theory originated from Vygotsky’s (1978) triad model of subject, object, and tools for psychological development, which was expanded by Engeström (1987) to include contextual elements of rules, community and division of labour in addition to subject, object and tools (Figure 1). The six elements compose the unit of analysis, in the socio-historical context of both individual and collective levels (Koszalka & Wu, 2004).
Activity theory provides an analytical framework for analysing the need, activity, and outcome of the technology-supported learning environment (Jonassen & Rohrer-Murphy, 1999; Rambe, 2012). The subject is an individual or a group participating in the activity. The object is the motive or goal driving the subject to take the activity. The tools refer to either material or psychology artefacts mediating the relationship between the subject and object. The rules are the norms the subject follows and decides the cooperation between the participants. The division of labour is the organisation of the activity and distribution of responsibility among the participants. The community refers to the social group mediating the interaction between each element (Zheng et al., 2020).

Research using activity theory in analysing educational technology has increased (Hite & Thompson, 2019; Kaptelinin & Nardi, 2006; Zheng et al., 2020), especially in game-based learning (Carvalho et al., 2015) and the use of social media in learning (Rambe, 2012). Researchers argued that activity theory can “address the challenge of studying the interaction between technology and actors” (Karanasios et al., 2018; p. 439). Activity theory has been used as the conceptual framework in the literature review of mobile assisted language learning for reading (Lin et al., 2019). AI as a fast-developing technology has been used in language education. However, to date there has been no systematic review of the empirical research on AI-supported language learning.

In addition to the aspects illustrated in the technology-based learning model in the Liang et al. (2021) review, this review used activity theory as the theoretical framework. The aim was to show the interaction between the subject, AI technology and the objects, and analyse the process of human interaction with technology via the collective activity in which the subject participates (Kaptelinin & Nardi, 2006). The strengths of using activity theory as an analytical framework include: (1) it can illustrate the boundary between the artefact/tool and the subject in constructing consciousness; (2) it can illustrate the materialisation of consciousness from socially mediated activities; and (3) it can show the transcendence from individual to collective activities for analysing the object-oriented, tool-mediated activity system (Rambe, 2012).

The research questions that this review paper aimed to explore were:

1. What language skills and learning outcomes are focused in AI-supported language learning?
2. What trends in AI technology are applied in language learning?
3. What trends in research design are employed in AI-supported language learning?
4. How does activity theory illustrate AI-supported language learning?

**Methodology**

**Selection procedure**

The empirical review employed a systematic content analysis. The search for relevant articles was
conducted on the databases, Web of Science and ERIC. In an evaluation of 28 academic search systems in terms of coverage, recall, precision, efficiency, and reproducibility, Web of Science and ERIC have been rated as principal and supplementary systematic search systems respectively (Gusenbauer & Haddaway, 2020). Therefore, these two databases were selected for this systematic literature review. Table 1 lists the search string. The inclusion and exclusion criteria are listed in Table 2. The procedure for selecting and screening literature in this review was based on the PRISMA method (Page et al., 2021) and is shown in Figure 2.

Table 1
Initial search string

<table>
<thead>
<tr>
<th>Topic</th>
<th>Search terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial intelligence</td>
<td>“artificial intelligence” or “AI” or “intelligent support”</td>
</tr>
<tr>
<td>Language learning</td>
<td>“language learning” or “language education”</td>
</tr>
</tbody>
</table>

Table 2
Final inclusion and exclusion criteria

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published January 2007 – February 2021</td>
<td>Published before 2007</td>
</tr>
<tr>
<td>Published in English</td>
<td>Not in English</td>
</tr>
<tr>
<td>Empirical research including an experiment or case study of AI-supported language learning in various educational contexts</td>
<td>Not empirical research (e.g., review or conceptual papers) and papers which claimed to be empirical but had very little or no information about the methodology</td>
</tr>
<tr>
<td>Academic peer-reviewed journal article</td>
<td>Book reviews, editorial materials, chapters and review articles</td>
</tr>
<tr>
<td>AI supported learning or education in a language acquisition domain. Articles related to first language learning (L1) and second language learning (L2) and four specific areas of such L1 and L2 language learning, such as speaking, listening, reading, and writing</td>
<td>Studies on natural language processing but not related to the application of AI in language learning</td>
</tr>
<tr>
<td>Indexed in Web of Science and ERIC</td>
<td>Papers which had no access to the full text</td>
</tr>
</tbody>
</table>
Each researcher followed the selection and exclusion criteria in Figure 2 and kept a spreadsheet to record the number of papers found, excluded, and retained after each step in the screening process. Each researcher first searched for related articles in the Web of Science and Eric respectively using the key search words. Next, the lists of articles were screened based on the exclusion criteria 1 and 4. After discussion and confirmation by both researchers, 124 papers remained. The next stage consisted of three rounds of screening of the 124 papers, at both abstract and full-text levels. In the first screening round, articles not related to AI applied in language learning were rejected. This resulted in 74 articles being retained. In the second screening round, 5 papers were not retrievable, resulting in 69 articles. In the third screening round, following exclusion criteria steps 3, 4, and 5, and the exclusion of duplicated papers, a further 44 articles were excluded, leading to the selection of 25 articles for the final review.

To ensure coding consistency for inclusion or exclusion of articles, the two researchers were responsible for article identification and screening from each database respectively, double checking each other’s screening. Inter-rater reliability was conducted over three rounds (Table 3). The inter-rater agreement percentage was 94% in the third round and was considered excellent for the coding of inclusion and exclusion criteria (O’Connor & Joffe, 2020). A final consensus was achieved by discussions and looking for evidence based on the inclusion and exclusion criteria. The agreement percentage was limited to demonstrating the quality of inter-rater agreement in terms of accuracy and precision (Belur et al., 2021). When uncertainty or disagreement arose in the exclusion process, both researchers examined the full text of pertinent articles and discussed the decision to reject or retain.

**Figure 2:** The PRISMA diagram (slightly modified from Page et al., 2021)

<table>
<thead>
<tr>
<th>Identification of studies via databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records identified from:</td>
</tr>
<tr>
<td>Web of Science (n = 173)</td>
</tr>
<tr>
<td>Eric (n = 484)</td>
</tr>
<tr>
<td>Records removed before screening:</td>
</tr>
<tr>
<td>Records marked as ineligible by automation tools: Papers from non-journal such as conference proceeding, book review, web resources were excluded</td>
</tr>
<tr>
<td>Web of Science (n = 116)</td>
</tr>
<tr>
<td>Eric (n = 397)</td>
</tr>
<tr>
<td>Records (abstract and title) screened:</td>
</tr>
<tr>
<td>Web of Science (n = 57)</td>
</tr>
<tr>
<td>Eric (n = 67)</td>
</tr>
<tr>
<td>Records excluded (n = 80)</td>
</tr>
<tr>
<td>Web of Science (n = 38)</td>
</tr>
<tr>
<td>Eric (n = 12)</td>
</tr>
<tr>
<td>Reason: not related to the application of AI in language learning</td>
</tr>
<tr>
<td>Reports sought for retrieval:</td>
</tr>
<tr>
<td>Web of Science (n = 16)</td>
</tr>
<tr>
<td>Eric (n = 55)</td>
</tr>
<tr>
<td>Reports not retrieved (n = 5)</td>
</tr>
<tr>
<td>Web of Science (n = 2)</td>
</tr>
<tr>
<td>Eric (n = 3)</td>
</tr>
<tr>
<td>Reports of full papers assessed for eligibility:</td>
</tr>
<tr>
<td>Web of Science (n = 17)</td>
</tr>
<tr>
<td>Eric (n = 52)</td>
</tr>
<tr>
<td>Reports excluded (n = 44)</td>
</tr>
<tr>
<td>Reason 1: Not related to language learning (n = 10)</td>
</tr>
<tr>
<td>Reason 2: Non-empirical papers, such as reviews or conceptual articles were excluded (n = 11)</td>
</tr>
<tr>
<td>Reason 3: Some paper which claimed to be empirical but had very little or no information about the methodology were excluded (n = 8)</td>
</tr>
<tr>
<td>Reason 4: Studies on natural language processing but not related to the application of AI in language learning (n = 6)</td>
</tr>
<tr>
<td>Reason 5: Duplicate records removed (n = 6)</td>
</tr>
<tr>
<td>Studies included in review (n = 25)</td>
</tr>
</tbody>
</table>
Table 3
Inter-rater agreement percentage

<table>
<thead>
<tr>
<th></th>
<th>Screening (title and abstract)</th>
<th>Retrieving</th>
<th>Full paper assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of papers</td>
<td>124</td>
<td>74</td>
<td>69</td>
</tr>
<tr>
<td>Number of papers with agreement</td>
<td>114</td>
<td>68</td>
<td>65</td>
</tr>
<tr>
<td>Number of papers without agreement</td>
<td>10</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Agreement percentage</td>
<td>92%</td>
<td>92%</td>
<td>94%</td>
</tr>
</tbody>
</table>

Analysis

The empirical review comprised three rounds of analysis. First, we summarised and analysed the articles for review in terms of publication years (Figure 3) and aspects of human-AI interaction (Figure 10). The timeline and numbers of publications for each year showed that empirical research surged in 2019 and 2020. In the second round of analysis, all 25 studies were coded according to the protocol adapted from the extended activity theory (Ali et al., 2015; Engeström, 1987). This consisted of seven constituents – subject, object, tool, rules, community, division of labour, and outcome (Appendix A). The subject refers to the research participants in the AI-supported language learning activity system. The tool refers to the role (function) of AI in language learning. The object means the purpose/aim of using AI for language learning. The outcome refers to the application of AI to language learning or the results of the study. The rules are the guidelines or research design using AI, including aspects of course design, teaching, and learning. An AI-supported language learning community typically consists of administrators, an instructor, learners (in L1, L2, or foreign language settings), researchers, and developers who help to create the learning technology (Lin et al., 2019). The division of labour includes the task distribution among students, instructor, developers and researchers (Lin et al., 2019). The third round of analysis synthesised the analysis from the first two rounds with some specific foci, such as the trends of research in a timeline, the limitations of the research (Appendix B), the context of the research and the underpinning principles of research design and human-machine interaction.

Figure 3: Numbers of reviewed publications by year of publication

Results

The research participants in the papers reviewed, served as the most crucial entity in the activity system of AI-supported language learning. The characteristics of the participants were examined in terms of their education levels and target language learning contexts. As shown in Figure 4, tertiary level students accounted for the highest proportion (80.0%) of the participants (undergraduates, graduates and lifelong
learners), with secondary level students (consisting of either junior or senior high school students) accounting for 8.0%, kindergarten and primary levels being 8.0% and 4.0% of the total, respectively.

With regard to the language learning contexts (Figure 5), 23 studies (92.0%) concerned L2 studies including 17 studies for English as a foreign language (EFL) (68.0%), three studies for Chinese as a foreign language (12.0%), two studies for Russian as a foreign language (8.0%), and one study (4.0%) for Spanish as a foreign language in different countries. There were two studies (8.0%) focusing on L1, one related to Chinese reading ability and interest (e.g., Hsiao et al., 2015) and one other related to English writing (e.g., Weston-Sementelli et al., 2018). Although the majority of the reviewed research focused on EFL, there has been an emerging trend extending the research to the use of AI for learning languages other than English.

**Figure 4**: Participants’ education level

**Figure 5**: Studies’ language contexts

**Tools**

As a mediating tool, the reviewed papers used AI technology in various ways to support language learning and teaching: automatic writing evaluation systems (9 studies reviewed, 36%), AI-robots such as chatbots, intelligent or humanoid robots (4: 16%), AI agents such as pedagogical or conversational agents (3 studies reviewed, 12%), intelligent tutorial systems (2 studies reviewed, 8%), and AI-supported voice-based smartphone apps (2 studies reviewed, 8%) (Figure 6). Smartphone apps were supported by an AI program
in a number of studies. Wei and Zhang’s (2019) study used intelligent computer assisted pronunciation teaching apps. Al-Kaisi et al.’s (2021) study used an AI voice assistant apps. In addition, AI technology was used as machine learning analytical tools (3 studies reviewed, 12%) and as a mind-wave headset (1 study reviewed, 4%) to test learners’ mind waves in language learning, as well as in building a computational model to predict learners’ language competency level (1 study reviewed, 4%).

**Figure 6:** AI-supported language learning tools used in the studies

**Object**

Figure 7 summarises various aspects of language learning covered in the reviewed studies.

**Figure 7:** Language learning skills and numbers of outcomes of the reviewed papers
The objects of the studies were divided into three main categories with various foci: (1) intelligent tutoring systems; (2) AI-supported automatic assessment systems; (3) AI technology for assessing learner-related issues. The first category was about the development of AI-supported systems to sustain or tutor the learners’ language learning. Many studies focused on the impact of an ITS on students’ language error corrections (e.g., Dodigovic, 2007), reading and writing strategies (e.g., Weston-Sementelli et al., 2018), or assessing the difficulty scoring of grammar in an intelligent language tutoring system (e.g., Pandarova et al., 2019). In this category, the use of AI technology ranged from a web-based intelligent tutoring system, including an AI-supported pedagogical agent and an AI chatbot, to mobile-based intelligent language teaching apps, as well as an intelligent humanoid-robot tutor.

The aims of the studies were divided into four segments:

(a) The effectiveness of AI-supported pedagogical agents in students’ vocabulary learning (e.g., Theodoridou, 2011), and students’ productive dialogue (e.g., Ayedoun et al., 2019; Tegos et al., 2014).

(b) The assessment of learners’ real-time levels of attention and meditation as well as their brain-wave activities, when they interact with humans in both face-to-face and virtual environments, and with a web-based AI chatbot (e.g., Hsu, 2020), and the effectiveness of humanoid/chat-robot on students’ engagement (e.g., Chew & Chua, 2020).

(c) The effectiveness of intelligent computer assisted pronunciation teaching apps on Chinese language pronunciation (e.g., Wei & Zhang, 2018) and that of an AI-supported voice assistant application on students’ Russian learning (e.g., Al-Kaisi et al., 2021).

(d) The effectiveness of an intelligent robot on pre-kindergarteners’ reading ability, interest, learning behaviour and mitigating English language anxiety (e.g., Bao, 2019); the effectiveness of an English-speaking humanoid-robot tutor on children’s vocabulary learning and beliefs about humanoid-robots (e.g., van den Berghe et al., 2020).

Studies in the second category focused on AI-supported automatic assessment systems, including learners’ engagement (e.g., Lu, 2019; Zhang, 2017), affordances on continuous learning intention (e.g., Fu et al., 2020), and the effectiveness of automatic writing evaluation on students’ writing performance (e.g., Lu, 2019; Tang & Rich, 2017). Regarding the feedback provided by automatic writing evaluation systems, the revised studies focused on the precision of students’ uptake of automatic writing evaluation feedback (e.g., Bai & Hu, 2017), the differences between automatic writing evaluation and teacher feedback, and students’ engagement with these two types of feedback (e.g., Zhang & Hyland, 2018). Regarding research participants’ engagement with and attitudes towards automatic writing systems, the reviewed studies covered teacher and students’ views about the use of an AI-supported automatic assessment system (e.g., Urum 2020; Uzun, 2020), students’ use of automatic writing evaluation as a social appropriation (e.g., Jiang & Yu, 2020), and the technology acceptance model with automatic writing evaluation (e.g., Li et al., 2019). The review showed that the field of automatic writing evaluation systems has been a significant area of research and AI-supported automatic writing evaluation is an emerging part of this area. Studies in the third category used AI technology to assess learner-related issues, such as using an artificial neural network (ANN)-based computational model to identify the predictive factors for learners’ overall English competences (e.g., Yang et al., 2019), and using a support vector machine to test the effective pedagogical factor set distinguishing high- from low-achieving ESL primary school readers (e.g., Xiao & Hu, 2019).

Outcomes

Aligning with the aims of the studies, the reviewed studies revealed three categories of outcomes: (1) the effectiveness of intelligent tutoring systems on students’ language learning; (2) the AI-supported automatic writing feedback systems helped students in writing; and (3) the AI-supported computer model can reveal or predict some learner-related factors. First, some studies showed that using an intelligent tutoring system improved the quality of writing (e.g., Weston-Sementelli et al., 2018), the accuracy levels of the scoring system (e.g., Pandarova et al., 2019), or reduced error rate (e.g., Dodigovic, 2007). Specifically, the outcomes of these studies included:
(a) Conversational agent intervention helped students’ productive dialogue (e.g., Tegos et al., 2014), or exhibited the potential to enhance learners’ willingness to communicate (e.g., Ayedoun et al., 2019).

(b) EFL learners’ level of attention was highest when they were socialising with other humans. When their interlocutor was a chatbot, their level of meditation was highest (Hsu, 2020). The use of a humanoid robot increased the learner’s engagement level (e.g., Chew & Chua, 2020).

(c) Intelligent computer assisted pronunciation teaching apps improved the quality of students’ Chinese language pronunciation (e.g., Wei & Zhang, 2018) and demonstrated potential in supporting students’ foreign language learning at the beginning stage (e.g., Al-Kaisi et al., 2021).

(d) AI chatbot reduced speech-related anxieties (e.g., Bao, 2019). In terms of learners’ beliefs about humanoid robots, boys anthropomorphised the robot tutor less after the lesson than did girls (e.g., van den Berghe et al., 2020).

Second, some studies showed that the automatic writing evaluation assessment and feedback can effectively help students in EFL writing (e.g., Lu, 2019), especially in motivating students to rewrite and revise (e.g., Tang & Rich, 2017). In addition, it was found that AI-programmed automatic writing evaluation can supplement peer and instructor feedback in the EFL writing classroom (e.g., Bai & Hu, 2017), because students still value teachers’ feedback on content and organisation (e.g., Zhang & Hyland, 2018). Meanwhile, automatic writing evaluation needs to be perfected as it cannot provide proper evaluation of the text structure, content logic, and coherence (e.g., Lu, 2019). In addition, some studies argued that the effectiveness of automatic writing evaluation feedback depended upon how individual students engaged with the feedback behaviourally, emotionally and cognitively (e.g., Zhang, 2017). Other studies revealed that social presence, peer influence and immediate benefit had influences on both emotional and cognitive engagement (e.g., Fu et al., 2020). Learners’ behavioural intention to use automatic writing evaluation was directly determined by perceived usefulness, their attitude towards using it and computer self-efficacy (e.g., Li et al., 2019). Learners’ appropriation of the automatic writing evaluation feedback included three subprocesses: selecting, emotion-regulating, and goal setting (e.g., Jiang & Yu, 2020). In addition, artificial augmentation increased the prediction accuracy of students’ EFL writing (e.g., Uzun, 2020).

Third, some studies showed that the AI-supported computer model revealed the mutual relationships between phonological awareness, phonological short-term memory, and long-term memory abilities (e.g., Yang et al., 2019). These findings yielded some implications for a teacher’s L2 pedagogical design, especially with cognitive ability-related intervention strategies (Yang et al., 2019). Meanwhile in another study, the support vector machine produced a ranking list of the factors that can distinguish the high- from low-achieving EFL students (e.g., Xiao & Hu, 2019).

In addition to the generally positive results revealed in the reviewed studies, some yielded fewer positive results. For example, it was found that web-based pedagogical agent-assisted vocabulary systems could not improve learners’ vocabulary recall and retention, though most learners expressed satisfaction with the learning environment (e.g., Theodoridou, 2011). In the Turkish EFL context, teachers and students both had a negative and pessimistic view towards the reliability of the AI-based assessment system, because it only evaluated the level of memory rather than assessing foreign language skills and higher order thinking skills (e.g., Ulum, 2020).

Rules

The rules of an activity system refer to the research design that both researchers and participants follow. As shown in Figure 8, the research design in the reviewed studies included: (a) experimental studies (15 of 25 papers); (b) a quasi-experimental study (1 of 25 papers, Dodigovic, 2007); (c) some quantitative studies used predictive analysis methodologies based on machine learning (e.g., Pandarova et al., 2019; Uzun, 2020; Xiao & Hu, 2019); d) qualitative studies, including case studies (e.g., Chew & Chua, 2020; Zhang, 2017); and (d) mixed method studies using multiple data sources, including test results, questionnaires and interviews (e.g., Fu et al., 2020; Uzun, 2020). The qualitative studies generally had small samples.
example, Zhang (2017) had one student case and Chew and Chuan’s (2020) inquiry case study involved six students, indicating the limitations of these studies (Appendix B). In terms of data collection instruments, surveys were commonly used in most studies (e.g., Ayedoun et al., 2019; Bai & Hu, 2017; Bao, 2019; Li et al., 2019; Lu, 2019; Tang & Rich, 2017; Theodoridou, 2011; Uzun, 2020). Among them, three papers reported that they used validated or published surveys (Ayedoun et al., 2019; Li et al., 2019; Uzun 2020). The other papers used bespoke surveys for their studies. This could be due to the lack of validated survey instruments for the application of AI in language education, which is still an emerging area of research. This also indicated the need for future validation of these instruments. In addition, interviews were used in both qualitative case studies and experimental studies as complementary data sources (e.g., Bai & Hu, 2017; Fu et al., 2020; Jiang & Yu, 2020; Lu, 2019; Tang & Rich, 2017; Tegos et al., 2014; Ulum, 2020; Zhang & Hyland, 2018).

Figure 8: Research design in the reviewed studies

Community

An AI community typically consists of administrators, teachers, learners and developers of learning technology (Lin et al., 2019). Regarding the context of research in the reviewed studies (Figure 9), those for L1 studies included one on Chinese language conducted in Taiwan (e.g., Hsiao et al., 2015), and one on English writing conducted in the USA (e.g., Weston-Sementelli et al., 2018). Studies on EFL learning included eight conducted in China (e.g., Bai & Hu, 2017; Jiang & Yu, 2020; Li et al., 2019; Lu, 2019; Tang & Rich, 2017; Yang, 2019; Zhang, 2017; Zhang & Hyland, 2018), two in Turkey (e.g., Ulum, 2020; Uzun, 2020), and one in each of Taiwan (e.g., Hsu, 2020), Japan (e.g., Ayedoun et al., 2019), Thailand (e.g., Bao, 2019), UAE (e.g., Dodigovic, 2007), Germany (e.g., Pandarova et al., 2019), the Netherlands (e.g., van den Berghe et al., 2020), and Canada (e.g., Xiao & Hu, 2019). In addition, three studies on learning Chinese as a foreign language were conducted in China (e.g., Chew & Chua, 2020; Fu et al., 2020; Wei & Zhang, 2018), two studies on learning Russian as a foreign language were conducted as one each in Greece (e.g., Tegos et al., 2014) and Russia (e.g., Al-Kaisi et al., 2021) respectively. One study learning Spanish as a foreign language was conducted in the USA (Theodoridou, 2011).
Figure 9: Research contexts

Division of labour

In the context of language learning with technology, the division of labour is directed to the task distribution amongst learning system developers, researchers, instructors and students (Lin et al., 2019). In some of the reviewed studies, the developers were responsible for developing AI-supported intelligent tutors or pedagogical agents (e.g., Dodigovic, 2007; Tegos et al., 2014; Theodoridou, 2011). As for the research with already developed systems or apps, the researchers' roles in other studies included the design of the tasks or experiments and providing instructions to participants to follow the design (e.g., Al-Kaisi et al., 2021; Ayedoun et al., 2019; Bao, 2019; Hsiao et al, 2015; Hsu, 2020; Li, 2019; Pandarova et al., 2019; Tang & Rich, 2017; Uzun, 2020; van den Berghe et al., 2020; Wei & Zhang, 2018; Weston-Sementelli et al., 2018; Zhang, 2017), building models based on data analysis (e.g., Bai & Hu, 2017; Fu et al., 2020;) and conducting surveys or interviews (e.g., Chew & Chua, 2020; Jiang & Yu, 2020; Li et al., 2019; Ulum, 2020; Xiao & Hu, 2019; Zhang & Hyland, 2018). In some studies, teachers as participants were also responsible for providing instructions for the design of learning content. However, in the reviewed studies, learners as the research participants had a rather passive role, following the design and/or instruction provided by the researchers and teachers.

Discussion

Aligning with the two research questions posed for this systematic review, this section first discusses the trend of language learning, the development of technology and design emerging from the empirical review. After that, the implications informed by using activity theory are discussed.

What language skills and learning outcomes are focused in AI-supported language learning?

With regard to the aspects of language learning, the reviewed studies emphasised the effectiveness of AI in supporting students’ learning of vocabulary, pronunciation, four language skills and dialogue. There was attention on some learner-related issues, such as their attention level, engagement, interest, and attitude or assessing their competence and achievement level. The outcomes included improved writing quality, accuracy, productive dialogue, reduced speech-related anxieties and increased engagement level. This review found that there has been more research focusing on language learning anxiety (e.g., Ayedoun et al., 2019; Bao, 2019; Li et al., 2019; Uzun, 2020), as compared to the two studies in the Liang et al. (2021) review which focused on learning anxiety in the higher education sector.
The reviewed studies on intelligent tutor/agent/robots focused more on language practice and revealed the effectiveness in improving learners’ pronunciation, error correction and their willingness to communicate, engagement in language learning and the lessening of anxiety. Although there was no particular effectiveness observed in some areas, such as vocabulary recall, the use of AI-supported language learning could save the time and labour of teachers and be an effective tool for engaging students in language learning. Meanwhile, some studies shifted from testing the effectiveness of AI-supported language learning to learners’ experience and learning engagement (e.g., Bao, 2019; Chew & Chua, 2020; van den Berghe, 2020). In addition, some studies showed that teachers’ intervention and configuration of the AI-supported language learning in the pedagogical design played an important role in the effectiveness of learning (e.g., Tegos et al., 2014).

Regarding the use of AI in automatic writing evaluation, the trend in studies reviewed showed a movement from testing the effectiveness of AI-supported automatic writing evaluation (e.g., Lu, 2019) to a focus on the learner and their engagement with automatic writing evaluation feedback (e.g., Bai & Hu, 2017; Li et al., 2019; Zhang, 2017; Zhang & Hyland, 2018). Although the findings of our review studies revealed that automatic writing evaluation feedback focused more on mechanical errors, and that most learners’ engagement with automatic writing evaluation feedback remained at the level of superficial error correction, researchers believed that it would benefit students’ learning by using automatic writing evaluation feedback in combination with teacher and/or peer feedback (Lu, 2019; Zhang, 2017). The future development of automatic writing evaluation should aim for evaluation of the text structure, content logic, ideas and coherence (Lu, 2019).

What trends in AI technology are applied in language learning?

In the studies reviewed, AI technology was applied as a set of tools with some human attributes to support language learning, reflected in the name of the design, such as intelligent tutor, humanoid robots, or analytical tools for analysing learner-related factors. As shown in the analysis of tools in Figure 6, the intelligent tutor has developed from being intangible and embedded in the online system (e.g., online-based tutoring system) without a concrete image, to mobile apps with text or audio for interaction (e.g., text-based or voice-based chatbot), and to more tangible and independent agents with a humanoid outlook, such as a humanoid robot. Online-based intelligent tutors or agents in language learning have been researched for a long time as part of the research for to develop adaptive learning systems. With the development of artificial intelligence, the intelligent tutoring system has been enhanced by incorporating personalised learning, which provides a tailored learning experience to individual learners based on their knowledge and preferences.

With the increasing popularity of mobile technology, educational apps, such as chatbot-based mobile apps for language learning have been developed, mobilising language learning anywhere and anytime with a mobile device. In other words, learners can easily learn through the mobile apps whenever and wherever they feel the need to learn in their daily life. Some studies reviewed recorded that chatbot-based mobile apps increased the interest and engagement of learners by the affordance of stimulating conversations (e.g., Zang & Aslan, 2021). At the same time, others reported that voice-based mobile apps improved learners’ communicative capability as well as grammar, reading, and writing skills (e.g., Al-Kaisi et al., 2021). Generally, the AI technology trends in language learning in the reviewed studies were aligned with those in general education, especially in the areas of building learner profile, assessment and evaluation, pedagogical agent/chatbot, and intelligent tutoring systems (Cheng et al., 2022; Zawacki-Richter et al., 2019).

Furthermore, in some studies reviewed AI-based humanoid robots, or social robots, were used as instructional tools in the language learning area, in particular in children’s language learning in recent years (Jamet et al., 2018; Kanero et al., 2018; van den Berghe et al., 2020). In these studies, a humanoid robot facilitated children’s thinking and perceptions in the way that some children considered the robot as a social person. NAO, which is one of the humanoid robots used as an instructional tool, is very similar in appearance to the children, and performs movements that closely mimic human movements. NAO is able to manipulate small objects, to make deictic gestures useful for learning, and to mimic gestures or signs.
NAO can orally articulate language, and the rate and tone of his voice are both easily able to be parametrised, which is particularly useful in learning reading and pronunciation (Jamet et al., 2018). The characteristics of these humanoid robots (e.g., NAO) are sufficient for children to have a positive attitude and motivation for their language learning in the classroom, and therefore the educational possibilities and applications of humanoid robots have been actively researched.

**What trends in research design are employed in AI-supported language learning?**

As reviewed in this paper, the dominant research design was an experimental or quasi-experimental design, with some using case studies or interviews. This was in line with the finding from the Liang et al. (2021) review. Most of research was carried out to prove the effectiveness of AI in language learning via a variety of experimental settings. Based on the results of the experimental design, the application of AI-supported language learning technology in a real classroom setting or learning outside of classrooms will be needed. This finding aligned with the need for applying AI in physical classroom settings suggested by other systematic reviews (Chen et al., 2020; Zhai et al., 2020). In addition, while the research has mostly focused on the learner’s interaction with AI on an individual basis, future research will be needed on collaborative language learning design supported by AI. Furthermore, as there were only two studies incorporating a teachers’ configuration of AI-supported language learning, teachers’ intervention using AI-supported language learning in the pedagogical design need to be considered in future research.

**How does activity theory illustrate AI-supported language learning?**

The use of activity theory for this research illustrated two aspects of the AI-support language learning. First the activity theory framework showed the dynamic interaction between the seven elements of the activity system. Second the concept of contradiction provided some insights in analysing the interaction between the research participants, technology and object, as well as outcomes. We acknowledge that the coding of the seven components of the activity theory may have overlaps (Lin et al., 2019).

First, the activity theory framework revealed the reciprocal interaction between the human-technology object in context with rules, community and division of labour, rather than a direct interaction between people and technology (Kapteinlinin & Nardi, 2006). The tools, including both technological (AI technology) and semiotic tools (language resources), mediated the research participants and the object of language learning (Rambe, 2012). For example, the interaction between the learner and the AI-supported enhanced learners’ language learning via an intelligent agent. The design was based on computational processing that mimics the thinking of humans, including automation to function (Kessler, 2018), such as emotion recognition/expression, and body language automation (LED eye with various colours) (Chew & Chua, 2020; Hsiao et al., 2015). These functions mediated learners’ cognition and learning process, such as a higher level of motivation and better performance in reading literacy when learning with a robot as distinct to learning with a tablet, because the sound-light effects and interaction with the robot attracted the learner’s attention (Hsiao et al., 2015).

Analysis of the reviewed studies for the interaction patterns between the research participants and the tool (Figure 10) showed that most research focused on the human-AI interaction on an individual basis with only two studies incorporating some collaborative design. This echoed another review on the use of AI in higher education that very few studies focused on the use of AI to facilitate learner collaboration (Zawacki-Richter et al., 2019). In addition, among the reviewed empirical studies, only two included the teacher’s interaction with AI. For example, Tegos et al. (2014) found that the teacher-configured AI-supported conversation agent provided beneficial language learning experience and enhanced their subject-related discussion, and some students enjoyed the collaborative discussion with some directed intervention. Similarly, Hsiao (2015) also revealed that the AI-supported robot enhanced the children’s collaborative reading activity. This confirms that knowledge generation and learning is a process embedded in different types of interaction between the participants, the tools, and the community of a collective activity (Kapteinlinin & Nardi, 2006).

The use of activity theory illustrated the reviewed research as collective activities through analysing an object-oriented, tool-mediated collective activity system (Rambe, 2012), which was also a community of
multiple points of view (Engeström, 2001). The review of the empirical research illustrated not only the role of learners but also the roles of the designers and researchers in the community. As shown in the columns of rules and division of labour (Appendix A), in most studies learners and teachers were still adopting a passive role in following the experimental design. The research of learner-related issues, such as predicting a learner’s performance level, followed the principle of extracting patterns from big data (e.g., Uzun, 2020; Xiao & Hu, 2019; Yang et al., 2019). This indicates the need for future research viewing teacher and students as active agents in interacting with technology and making transformations in real life learning contexts (Goodfellow et al., 2016; Weigand, 2019).

![Figure 10: Human-AI interaction](image)

The second principle of activity theory provided for in this review was the concept of contradiction. In this review, we applied two types of contradiction in the activity system: primary contradiction which exists within each component of the central activity, and secondary contradiction existing between the components, such as between the subject and the tool (Engeström, 1987). For example, the review indicated that most of the interactions between learners and AI were based on the non-communicative use of language such as pronunciation, vocabulary learning and error corrections, though some studies focused on reading and writing, while the outcomes were measured by objective tests. There were relatively few studies focusing on productive dialogue and communication, whereas with the positive results that AI-supported language learning promoted learners’ collaboration (Hsiao et al., 2015; Tegos et al., 2014). This confirms that most AI-supported language learning technology/virtual assistants do not represent an effective communication scenario (Godwin-Jones, 2019). One reason could be that language use in communication as the core of human intelligence cannot be grasped by formal rules or programmes (Weigand, 2019). Some robots used as an intelligent agent had difficulties in recognising inaccurate pronunciation and multiple voices at once (Chew & Chua, 2020). Students’ and teachers’ prior exposure to AI may affect their perceptions. Contradictions can result in tensions but may also trigger transformation in the activity systems (Kamanga & Alexander, 2020; p. 3). This indicates the need for resolution from both components, such as the improvement of the intelligent agents and the automatic writing evaluation system or complementing them with other assessment strategies. From an activity theory perspective, to resolve the contradiction, this may involve changing the established norms and existing power structures (Kamanga & Alexander, 2020). AI-supported language learning used in conjunction with teacher pedagogy would enhance effectiveness and experience in language learning.
The development of the internet has evolved rapidly from Web 1.0, through Web 2.0 and now into the era of Web 3.0. This has affected the way learning has been constructed. For instance, e-learning 1.0 was featured with a more teacher-centred information delivery model. In Web 2.0, with the affordances of exchanging and creating information enabled by social networks, the e-learning 2.0 also moved to collaborative and project-based learning models underpinned by constructivism and connectionism (Rubens et al., 2014). Web 3.0 is more structured and intelligent than Web 2.0 in terms of storing and processing information following semantic rules (Morris, 2011). The digital technology used in Web 3.0 includes AI systems that could run smart programs to assist users, virtual community with human avatars, and intelligent agents (Bidarra & Cardoso, 2007). Accordingly, e-learning in Web 3.0 should be more collaborative and intelligent with the support of intelligent agents (Rubens et al., 2014). However, this review showed that some features of AI technology have not been fully utilised to support collaborative language learning and group interaction, which was also revealed by other systematic reviews (Liang et al., 2021; Zawacki-Richter et al., 2019).

Another gap revealed in this systematic review was that very little research discussed the ethical concerns of AI application, agreeing with the Zawacki-Richter et al. (2019) review. Only Chew and Chua’s (2020) study included the ethical-pedagogical reflection on the use of an AI-programmed robot. Although half of the reviewed papers were concerned with the generalisability issue due to the small sample size (Appendix B), a few studies discussed the challenges and problems in AI application. For example, the Ulum (2020) study pointed out the inadequate function of the AI-supported assessment system, especially the lack of reliability in assessing learners’ language skills. Hsiao et al. (2015) were concerned with the high cost of AI-based robots which affected their wider application in an education setting. These gaps indicate the areas for future research.

**Conclusion, limitations, and future research direction**

The reviewed studies were analysed in terms of the seven components of activity theory, revealing the trends of research in the field of AI-supported language learning. Based on the analysis of the interconnection between these seven components, this paper noted some pedagogical implications and proposed future research areas. One pedagogical implication is that a mixed module of AI-supported language learning and formal teacher instruction should be incorporated in the pedagogical design. One way to do this is to design bridging activities between formal classroom instruction with after-class online learning, in terms of providing support and resources for learners to work autonomously online (Little & Thorne, 2017). The second implication is to use AI-supported language learning to support learner collaboration in language learning, as research shows that learners still prefer working with their peers in conjunction with the AI-supported language learning.

We acknowledge that one limitation of this review paper was the relatively small sample for review. Due to the accessibility issue, the small size of the reviewed study cannot ensure the generalisability of the findings to other language learning contexts. The second limitation is that we did not carry out a separate quality appraisal of the studies. The third limitation lies in the method of calculating inter-rater reliability, which could be further improved by using the $k$ statistic to measure accuracy and precision (Belur et al., 2021).

The trend towards hybrid learning models also indicates a future direction for the research on AI-supported language learning, moving from an experimental setting to the research on learner’s interaction with AI in real life situations. Another future research direction would be the mediation of AI-supported language learning in collaborative learning design. This review suggests several future research directions, including how language and meaning are negotiated in the interaction with AI, teacher and students’ cultural role in the interaction of AI, as well as the impact on the power structure during their interactions with AI.

**References**


Corresponding author: Suna Kyun, skyun@sookmyung.ac.kr

Copyright: Articles published in the Australasian Journal of Educational Technology (AJET) are available under Creative Commons Attribution Non-Commercial No Derivatives Licence ([CC BY-NC-ND 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0)). Authors retain copyright in their work and grant AJET right of first publication under CC BY-NC-ND 4.0.

## Appendix A
### Analysing research studies using activity theory

<table>
<thead>
<tr>
<th>Authors</th>
<th>Subjects</th>
<th>Tool</th>
<th>Object</th>
<th>Outcome</th>
<th>Rules</th>
<th>Community</th>
<th>Division of labour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dodigovic, 2007</td>
<td>266 university students</td>
<td>AI-supported Intelligent tutor</td>
<td>To investigate whether the AI-supported intelligent tutor has effect on the second language (L2) learning outcomes.</td>
<td>Artificial intelligence was an efficient instrument of error remediation, reducing the error rate by an average of 83%.</td>
<td>Quasi-Experimental Design. (pre-test, treatment, post-test)</td>
<td>Researcher Intelligent tutoring system developers L2 Students from Taiwan, Australia and UAE</td>
<td>Developers developed an AI-supported intelligent tutor. Researcher designed quasi-experiment and provided pre-test, treatment, and post-test to students.</td>
</tr>
<tr>
<td>Theodoridou, 2011</td>
<td>47 university students</td>
<td>AI-supported pedagogical agents for vocabulary learning system</td>
<td>To investigate whether pedagogical agents have an effect on learners’ vocabulary recall and retention. To investigate learners’ reactions and attitudes on the pedagogical agents in Spanish vocabulary learning.</td>
<td>AI-supported pedagogical agents could not improve learners’ vocabulary recall and retention, but most learners expressed satisfaction with the learning environment.</td>
<td>Experimental design (24 learners in control group, 23 learners in experimental group)</td>
<td>Researcher The developer of the AI-supported Pedagogical agents Students enrolled Spanish as a second language classes in a university in the US</td>
<td>Teacher designed research model and collected data. Teacher designed the domain issue/topic and configured agent’s behaviours and rules.</td>
</tr>
<tr>
<td>Tegos et al., 2014</td>
<td>30 university students</td>
<td>Conversational agent (Mantor Chat) based on an intelligent tutoring system</td>
<td>To test the effectiveness of a conversational agent on students’ productive dialogue.</td>
<td>Conversational agent (Mantor Chat) intervention can help students’ productive dialogue.</td>
<td>Pilot study with a post-questionnaire, focus group interviews and discourse analysis of students’ interaction</td>
<td>Researcher Teacher Students (in, English as a second language setting) at a university of Ukraine</td>
<td>Researcher designed research model and collected data. Teacher designed the domain issue/topic and configured agent’s behaviours and rules.</td>
</tr>
<tr>
<td>Hsiao et al., 2015</td>
<td>57 pre-kindergarten</td>
<td>An intelligent robot (iRobiQ)</td>
<td>To test the effectiveness of an intelligent robot (iRobiQ) on pre-children’s reading</td>
<td>An intelligent robot (iRobiQ) can enhance children’s reading</td>
<td>Experimental design (27 in control group, 27 in experimental group)</td>
<td>Researchers Children from pre-kindergarten</td>
<td>Researcher assigned children into experimental and control groups,</td>
</tr>
<tr>
<td>Study</td>
<td>Participants</td>
<td>Methodology</td>
<td>Findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>-------------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bai &amp; Hu, 2017</td>
<td>30 Chinese university students</td>
<td>An automated writing assessment system, Pigai</td>
<td>Investigated the precision of feedback provided by Pigai system and students’ uptake of such feedback.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang, 2017</td>
<td>One Chinese university student</td>
<td>An automated writing assessment system, Pigai</td>
<td>Investigated students’ engagement (behavioural, emotional, cognitive) with automated writing assessment feedback.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Weston-Sementelli et al. (2018)**

175 undergraduate students

Two intelligent tutoring systems, iSTART and Writing Pal

To investigate the impact of the two intelligent tutorial systems on students’ reading and writing. Combined (iSTART & Writing Pal) strategy training condition produced higher quality writing tasks than other conditions (iSTART only, Writing Pal only, and control groups).

Researcher
Undergraduate students (in L1 setting,) at a university in the US

Researchers
Students into four groups, provided pre-tests, and analysed the students’ outcomes.

**Wei & Zhang, 2018**

36 international students who enrolled in a Chinese university

An intelligent computer assisted pronunciation teaching application system, Erya

To test the effectiveness of intelligent computer assisted pronunciation teaching app - Erya on Chinese as second language learner’s pronunciation learning. The intelligent computer assisted pronunciation teaching app can help the learner to complete a large number of pronunciation exercises and improve the pronunciation quality.

Researchers
Teachers Chinese as second language from a university in Beijing

The teacher assigned students Chinese pronunciation exercise/tasks via the app. Students completed the assigned task via the apps. Researchers kept records of students’ exercise times and success rate.

**Zhang & Hyland, 2018**

Two Chinese university students and their English teachers

Automatic writing evaluation system, Pigai, and teacher feedback

To investigate the difference between teacher feedback and Pigai feedback. To investigate how L2 students engage with teacher feedback and Pigai feedback on their Teachers’ comments on content and organisation were highly valued by students. Experienced teachers can offer more comprehensive corrective feedback on student writing.

Qualitative study: students’ writing drafts; Pigai feedback and submission information; teacher feedback on the draft; interviews with the participant

Researchers, teachers and students

Researcher designed the study and interviewed students. Teacher assigned students a writing task and provided feedback. Students completed the essay, submitted to Pigai and revised based on Pigai feedback.

Researcher assigned students into four groups, provided pre-tests, and analysed the students’ outcomes.
Ayedoun et al., 2019  | 40 Japanese university students  | Conversational agent equipped with conversational strategies and affective backchannels  | To investigate the impact of a conversational agent with conversational strategies and affective backchannels on learners’ willingness to communicate in a second language. | The combining of conversational strategies and affective backchannels enhanced L2 learners and also the affective backchannels only version of the system had the potential to enhance their willingness to communicate to some extent. | Experimental design (affective backchannels only, conversational strategies only, and combined groups)  | The researchers assigned students into three groups; provided pre-tests (confidence, anxiety, & desire to communicate tests), treatment sessions, post-tests, and analysed the students’ outcomes. |

Bao, 2019  | 40 adults employed in a large financial institution in Thailand  | English speaking AI chatbot  | To investigate whether an AI chatbot can mitigate English language anxiety. | AI chatbot reduced speech-related anxieties and learning inhibitions of English as a second language students. | Experimental design (19 in control, and 21 AI treatment groups)  | Researcher assigned participants into control and experimental groups, measured learners’ anxiety, attitudes toward daily chatbot usage, conducted interviews and IELTES tests of participants’ English-speaking ability. |

Lu, 2019  | 114 Chinese university students and 30 teachers  | Automatic writing evaluation system, Pigai, and teacher-feedback  | To explore the effectiveness of Pigai on helping students with their English writing. - Both teachers and students had a positive attitude to Pigai but found that automatic writing evaluation system cannot provide proper evaluation on the text structure, content logic, and Pigai effectively helped the students with their English writing.  | Experimental design (pre- and post-tests; experiment); questionnaires and interviews  | Researcher, teacher and students  | The teacher assigned the students’ writing task and then read the students compositions online and added the teacher’s feedback and comments. Students took part in the experiment including pre- and post-tests and submitted their writing to Pigai. The researcher conducted a
<table>
<thead>
<tr>
<th>Name(s)</th>
<th>Year</th>
<th>Type of Learners</th>
<th>Methodological Approach</th>
<th>Findings</th>
<th>Questionnaire and Interviews with</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li et al., 2019</td>
<td></td>
<td>Chinese university students</td>
<td>Automatic writing evaluation system, Pigai</td>
<td>Results revealed that learners’ behavioural intention to use Pigai was directly determined by</td>
<td>30 teachers and 200 students</td>
<td>Students completed the questionnaire and interviews with 30 teachers and 200 students.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>perceived usefulness, attitude towards using, and perceived ease of use when using Pigai.</td>
<td></td>
<td>The researchers analysed the data to validate the proposed model.</td>
</tr>
<tr>
<td>Pandarova et al., 2019</td>
<td>787 high school students</td>
<td>Adaptive language learning technologies, current print and digital English learning materials</td>
<td>To advance the (semi-) automatic difficulty scoring of grammar exercise items to be used in dynamic difficulty adaptation in an intelligent language tutoring system.</td>
<td>The intelligent language learning technologies encouraged prediction accuracy levels.</td>
<td>Four native speaker experts; 787 9th and 10th graders in two preparatory high schools in German</td>
<td>Native speakers evaluated item quality and possible solutions. The researcher administered the test to 787 students.</td>
</tr>
<tr>
<td>Yang et al., 2019</td>
<td>15 Chinese middle school students</td>
<td>Neural network based computational model</td>
<td>Using neural network model to find the predicting factors to learners’ overall English competencies.</td>
<td>The neural network model can show the mutual relationships among the phonological awareness, phonological short-term memory, and long-term memory abilities.</td>
<td>Middle school students in an English as a foreign language class; their teachers; the</td>
<td>The students completed three learning tasks. Teachers gave observational evaluation for their performance. The researcher carried out the experiment, and analysed the data.</td>
</tr>
<tr>
<td>Authors</td>
<td>Methodology</td>
<td>Analysis Techniques</td>
<td>Findings</td>
<td>Additional Notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xiao &amp; Hu, 2019</td>
<td>Primary school learners and teachers</td>
<td>Using support vector machine algorithm to analyse pedagogical factors associated with reading materials, classroom organisation, reading strategies, in-class and post-reading activities.</td>
<td>The effective pedagogical factor distinguished high-from low-achieving primary school readers.</td>
<td>The students took the test. Teachers participated in the questionnaires. Researchers did the statistical analysis by using support vector machine.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jiang &amp; Yu, 2020</td>
<td>An English teacher and nine university students</td>
<td>Automatic writing evaluation system, Pigai</td>
<td>The findings revealed three forms of appropriation (i.e., regular, partial, and rare) among the students, and the students further differed in their internalisation of the resources.</td>
<td>A teacher and students in an English as a foreign language setting in a Chinese university</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chew &amp; Chua, 2020</td>
<td>Six university students</td>
<td>Autonomous programmable robot, NAO</td>
<td>The novel learning experience was more fun and interesting, and thus engagement from the axis of novelty, interactivity, motivation and interest was enhanced.</td>
<td>Each student would learn Chinese with NAO. NAO recorded students’ responses. The researchers observed students’ engagement with the Chinese learning sessions and interviewed students.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Year</td>
<td>Participants</td>
<td>Methodology/Technology</td>
<td>Findings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>------</td>
<td>--------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fu et al., 2020</td>
<td></td>
<td>260 Chinese university students</td>
<td>AI-enabled automatic scoring application, LAIX (To test 6 hypotheses and to examine the role and affordances of automatic scoring application on cognitive/emotional engagement and following continuous learning intention.)</td>
<td>The results revealed that social presence, peer influence and immediate benefits of the automatic scoring application influenced on emotional and cognitive engagement.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsu, 2020</td>
<td></td>
<td>30 university students</td>
<td>Mindwave headset, NeuroSky; a virtual platform, and an AI chatbot (To assess learners’ real-time levels of attention, meditation, and their brainwave activities in each of the three contexts: with another human in person, with another person through a virtual platform, and with an artificial intelligence chatbot.)</td>
<td>The EFL learners’ level of attention was highest when they were socialising with other humans in person. When their interlocutor was a chatbot, their level of meditation was highest. When they were interacting with another person in a virtual environment, both their attention and meditation were lowest.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ulum, 2020</td>
<td></td>
<td>30 students and 10 teachers</td>
<td>An AI-supported automated (To investigate EFL students and teachers’ views about the use of)</td>
<td>The students and teachers showed a negative view towards</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Researchers built and tested a model of the role and affordances of AI-enabled automatic scoring application and conducted questionnaires and interviews with the students. Researchers conducted an English proficiency with the participants and assigned participants to undertake two designed tasks in three socialisation settings (i.e., with AI Chatbot, in-person socialisation and with another person through a virtual environment).
<table>
<thead>
<tr>
<th>Researcher</th>
<th>Year</th>
<th>Participants</th>
<th>Methodology</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>van den Berghe et al., 2020</td>
<td></td>
<td>104 5-year-old children in Netherlands</td>
<td>Versant English Test and individual interviews with teachers</td>
<td>The application of the AI-based Versant test and perceived it as unreliable and invalid.</td>
</tr>
<tr>
<td>Uzun, 2020</td>
<td></td>
<td>102 university students in Turkey</td>
<td>Machine learning method, Parameter K’ algorithm</td>
<td>There was positive correlation between children’s’ anthropomorphism before the lessons and post-test scores. Boys anthropomorphised the robot tutor less after the lesson than girls.</td>
</tr>
<tr>
<td>Al-Kais et al., 2021</td>
<td></td>
<td>24 university students</td>
<td>AI-supported Voice assistant smartphone App, Alice</td>
<td>The voice assistant had a wide didactic potential to enhance the independent studies of foreign language, specifically in speaking and listening.</td>
</tr>
</tbody>
</table>

Researchers assigned children into two iconic-gesture and no-iconic-gesture conditions group, measured children’s pre/post-test, and analysed the children’s outcomes.

Researchers assigned students into control and experimental groups and conducted pre-test, treatment and post-tests (reading, writing, listening, speaking of Russian) with the students.
## Appendix B
### Limitations of the reviewed studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Limitations</th>
</tr>
</thead>
</table>
| Dodigovic, 2007              | Short length of research time  
Small number of participants  |
| Theodoridou, 2011            | Not mentioned                                                                                                                                                                                                |
| Tegos et al., 2014           | Short length of research time  
Small number of participants  |
| Hsiao et al., 2015           | The AI-based robot was too expensive, and therefore the use of robot was limited in only four places  
Teachers’ strategies to use AI-based robot in their instruction is necessarily needed, otherwise AI-based robot is not effective in instruction  |
| Tang & Rich, 2017            | Not mentioned                                                                                                                                                                                                |
| Bai & Hu, 2017               | Only computer-generated data were analysed, overlooking students’ self-initiated revision to correct students’ writing errors  |
| Zhang, 2017                  | Only one case study, which was difficult to be generalised widely  |
| Weston-Sementelli et al., 2018| Limited number of learning contexts  |
| Wei & Zhang, 2018            | The findings may not always be generalisable as the research site is a homogenous setting  
The participants used mobile phones slightly for non-reading purposes  
It was not easy to sustain participants’ motivation and determination for continued learning outside class  |
| Zhang & Hyland, 2018         | The shared annotations interfered with the reading process  
For high-achievement learners, annotations from low-achievement students were useless  
For low-achievement learners, they did not benefit from annotations made by people with the same proficiency level as them, indicating that difficult words were not annotated |
| Ayedoun et al., 2019         | A sample size of 40 students in one Asian country could not be generalised widely  
Only one context was used for experiments, which was not enough to measure students’ affective strategies for L2 learning  |
| Bao, 2019                    | A small sample size led to limited generalisability  
Duration of the research only 4 weeks, which was not long enough for ideal learning attitude change experiments  |
| Lu, 2019                     | A small sample size led to limited generalisability  
The method for calculating the amount of time spent is limiting, a more effective method for a more detailed analysis is needed  |
| Li et al., 2019              | Since learners used their own mobile device, each with a different configuration, the learning became inconsistent among all the learners, leading to hindrance of the experiment  
The presentation of certain digital content was not suitable for some of the mobile devices, leading to problems with mobile form factors (e.g., e-reader or old tablets) |
Pandarova et al., 2019  
A small sample size led to limited generalisability  
Students were of different backgrounds (e.g., age and English proficiency)  
No observation of learners’ peer interactions

Yang et al., 2019  
Not mentioned

Xiao & Hu, 2019  
Not mentioned

Jiang & Yu, 2020  
Only eight EFL teachers from Taipei were invited as participants of the study  
Only one elementary school in Taipei participated in the evaluation of mobile EFL reading system  
Students from different regions of Taiwan differ in their instructional needs and should be represented appropriately in future studies

Chew & Chua, 2020  
Since many principles in extensive reading were applied to the mobile reading mode, the positive results should not be directly linked to mean the replacement of paper-based reading, despite the assumption of the study that the mobile reading mode led to higher interests and motivation

Fu et al., 2020  
A small sample size led to limited generalisability

Hsu, 2020  
The treatment duration was too short  
The experimental group did not engage in self-paced learning completely due to having to learn at both the computer assisted language learning lab and standard classroom  
The validity was questioned because the unit tests were part of the course curriculum, making the interval data for statistical analysis questionable

Ulum, 2020  
The findings are drawn from differences-in-differences results with an estimate on the effects of the three treatments  
Causality could not be inferred outside of the Kisumu County to other areas in Kenya

van den Berghe et al., 2020  
Not mentioned

Uzun, 2020  
A small number of participants  
The algorithm to predict L2 writing performance is too simple (predicts only pass or fail)

Al-Kais et al., 2021  
A small sample size led to limited generalisability  
Only students at the beginner level (Russian as a foreign language) were tested