IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 30, 2020, accepted August 21, 2020, date of publication September 3, 2020, date of current version September 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021383

A Systematic Literature Review of Intelligent Tutoring Systems With Dialogue in Natural Language

JOSÉ PALADINES [1,2] **AND JAIME RAMÍREZ** [1]

[1]Computer Science School, Universidad Politécnica de Madrid, 28660 Madrid, Spain
[2]Thecnical Science Faculty, Universidad Estatal del Sur de Manabí, 130650 Manabí, Ecuador

Corresponding author: José Paladines (jose.paladines@unesum.edu.ec)

This work was supported by the Universidad Estatal del Sur de Manabí.

ABSTRACT Intelligent tutoring systems (ITSs) are computer programs that provide instruction adapted to the needs of individual students. Dialog systems are computer programs that communicate with human users by using natural language. This paper presents a systematic literature review to address ITSs that incorporate dialog systems and have been implemented in the last twenty years. The review found 33 ITSs and focused on answering the following five research questions. a) What ITSs with natural language dialogue have been developed? b) What is the main purpose of the tutoring dialogue in each system? c) What are the pedagogical features of the teaching process performed by the ITSs with natural language dialogue? d) What natural language understanding approach does each system employ to understand students' utterances? e) What evidence exists related to the evaluation of ITSs with natural language dialogue? The results of this review reveal that most ITSs are directed toward science, technology, engineering, and mathematics (STEM) domains at the university level, and the majority of the selected ITSs implement the expectations and misconceptions tailored approach. Furthermore, most ITSs use dialog to help students learn how to solve a problem by applying rules, laws, etc. (the apply level in Bloom's taxonomy). With regard to the instructional approach, the selected ITSs help students write correct explanations or answers for deep questions; assist students in problem solving; or support a reflective dialogue motivated by either previously provided content or the result of a simulation. Additionally, we found empirical evaluations for 90.91% of the selected ITSs that measure the learning gains and/or assess the impacts of different tutoring strategies.

INDEX TERMS Intelligent tutoring system, natural language dialogue, natural language processing, systematic literature review.

I. INTRODUCTION

Intelligent tutoring systems (ITSs) are computer programs that provide instruction adapted to the needs of individual students; i.e., they perform functions inherent to the tutorial process (presenting information that must be learned, asking questions or assigning tasks, providing feedback, etc.) to cause a cognitive and motivational change in the student. To accomplish this goal, ITSs leverage artificial intelligence techniques to define content models (the subject to be taught) as well as the tutoring strategies to be employed with each student; i.e., they specify ''what'' and ''how'' to teach [1].

According to what was reported by du Boulay [2], there is substantial empirical evidence of the effectiveness of ITSs [3]–[6]. Other studies have found some evidence of the connection between students' metacognitive decisions while working with an ITS and their learning gains [7].

A dialog system is a computer program that communicates with a human user by using natural language. Dialog systems are currently gaining interest in different fields of application, such as e-commerce, personal assistants, and call centers.

Natural language tutoring systems implement several well-validated instructional strategies, such as active participation in learning [8], adequate feedback [9], collaborative interaction [10], or the presence of impasse or cognitive imbalance [11]. These strategies are reported in the literature

The associate editor coordinating the review of this manuscript and approving it for publication was Okyay Kaynak.

related to the ITSs developed over recent years that have dialogue in natural language. Additionally, the literature shows an evolution in the tutoring provided by some of these systems, providing a better understanding of their functionality and learning benefits. One example of this evolution is described in [12]. This work shows an ITS called Autotutor and the subsequent versions of this ITS, which during the past two decades have been incorporating improvements to the tutoring process. In another study [13], it can be seen that starting from SQL-Tutor, many tutors based on constraints have been developed to teach well-defined and ill-defined tasks. Moreover, this work also revealed how the natural language dialogue strategies implemented by these tutors have improved over time. In [14], a brief survey is presented that describes some of the ITSs that support conversational dialogues together with the natural language processing (NLP) techniques that facilitate having free entry of words and sentences.

However, to the best of our knowledge, to date, no systematic literature review of ITSs with dialogue in natural language has been presented. Hence, the main contribution of this article is to provide an overview of this kind of system that covers, for each system, the teaching purpose of the dialogue, the main pedagogical features of its tutoring dialogue, the natural language understanding approach adopted, and the empirical support of its learning effectiveness.

To identify evidence that pertains to the ITSs with dialogue in natural language, we have carried out a systematic literature review (SLR) based on the Kitchenham and Charters guidelines [15]. They define an SLR as a means of identifying, evaluating and interpreting all of the available research relevant to a certain topic area. This process includes the identification and classification of contributions in a specific field of interest to provide a framework/background to appropriately position new research activities. Thus, the main goal of our SLR is to identify and structure the existing knowledge about ITSs that integrate natural language dialogue in the tutoring process.

The remainder of the paper is structured as follows. Section 2 presents the theoretical foundations for ITSs and dialog-based systems. Section 3 describes the method that we followed to collect and compare the different ITSs that exist in the literature. Section 4 reports the results of the literature review along with an analysis of the answers found for research questions. Section 5 provides a discussion of previous results. Section 6 discusses some threats to the validity of this study. Section 7 describes some future trends in this field. Finally, Section 8 outlines the conclusions of this research and suggests some future lines of work.

II. THEORETICAL FOUNDATIONS

A. INTELLIGENT TUTORING SYSTEMS

The term ITS has its origin in the publication of Carbonell [16], in which a system called Scholar is detailed. This system consists of a set of programs that use a semantic representation of the student's knowledge to create a dialogue with him/her. Later, Sleeman and Brown [1] enumerated the main features that an ITS should possess: it should be adaptive to the student, be able to exchange the control with the student, and possess domain-specific knowledge. Then, Wenger [17] took inspiration from these ideas to present a reference architecture for an ITS composed of four modules: communication, tutorial, student and expert.

In Wenger's architecture, the **communication module** is the interface between the student and the ITS, through which either the student's tutoring requests are received to evaluate his/her solution or the ITS provides an immediate response based on the student's behavior. The **tutorial module** is the core of the ITS and implements the tutoring strategy. The **student module** contains information on the student that is related to his/her knowledge, previous actions (logs), learning style, etc. Such a student module typically features some inference mechanisms to diagnose what the student knows or not from his/her answers, actions and/or utterances. All of this information is essential to decide the best tutoring strategy for each student. The **expert module** keeps the knowledge domain, i.e., the concepts to be taught and/or the tasks that the student must learn to perform.

Generally, ITSs provide feedback and hints in each step of the problem-solving process. In some cases, feedback and hints are shown immediately every time the student performs a step [18], [19], whereas in other cases, the ITS waits until the student has submitted a whole solution [20].

Systematic reviews of ITSs [21]–[23] have shown that ITS research has successfully provided techniques and systems to help students acquire specific cognitive and metacognitive knowledge in different areas (e.g., medicine, mathematics, science, database design). ITS research has also proposed different strategies for tutoring, such as the identification and correction of errors [24] and the construction of self-explanations [25], [26].

As student modeling is a central element of the ITS design, researchers have adopted a variety of approaches depending on how the student model is represented. As a result, we have considered the ITS types mentioned in [4]: model-tracing tutors (MTTs), constraint-based modeling (CBM), Bayesian network modeling (BNM), and expectation and misconceptions tailoring (EMT). Next, we will briefly explain each approach.

In MTTs, domain knowledge is captured through rules, and the student model is required to "trace" the student input and find a sequence of rule executions whose final result matches the student's contributions [27]. To provide the input, some operations or student's actions are available in the user interface that enable student to advance towards the solution of the problem.

In CBM, domain knowledge is represented as a set of constraints, and the student model has to identify the constraints that the student violates during the resolution of the problem [28]. This type of ITS is particularly well-suited to ill-defined domains in which there could be many possible solutions to the same problem.

In BNM, the student model contains causal connections between pieces of target domain knowledge and observable student actions. Additionally, every piece of target domain knowledge has an associated probability that represents the system's best assumption that the student knows that piece of knowledge [29].

In EMT, the student model represents a set of missing pieces of information in an expected response and a set of errors and misconceptions expressed by the student thus far [30].

Although the scope of this SLR is the ITSs with natural language dialogue, it is worth noting that there are many ITSs developed after Scholar and before 1998 where the communication module did not focus on natural language feedback. Instead, the communication module of these ITSs was implemented entirely using a GUI for student input and feedback that was sometimes accompanied with some text.

B. DIALOG-BASED SYSTEMS

A dialog system is a computer program that communicates with a human user by using natural language. A key component of a dialog system is the dialog manager, which is responsible for keeping track of the state and flow of the conversation as well as making decisions on the most appropriate next system action (e.g., the answer).

A fundamental aspect of dialog systems is NLP. NLP is considered to be a subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content [31]. NLP is also employed to convert information stored in natural language to a machine-understandable format.

NLP systems vary in the employed techniques, are built for different purposes, and can differ in focus to link natural language inputs and outputs with their domain models.

A subfield of NLP is natural language understanding (NLU), which addresses the understanding of the user's utterances. There are three main approaches for NLU: symbolic, statistics, and hybrid approaches [32].

The symbolic approach is based on linguistic or lexicographical knowledge. This knowledge is specified by a language expert in form of a lexicon and a grammar that later can be employed to analyze the user inputs. On the other hand, the statistical approach is based on quantitative methods and uses a *corpus* to train a text classifier without knowledge of grammars or lexicons provided as direct input [33]. So, this approach relies on discovering the cooccurrence relationships of the words in text excerpts of a *corpus*. Finally, the hybrid approach results from a combination of the other two approaches, with the goal of complementing each other. For example, syntactic features can be used to complement the training of a text classifier.

Tutoring dialogue is considered to be a dynamic form of instruction that can be highly adaptable to the individual needs of students and provides opportunities for students to make their thinking transparent to a tutor [34]. Thus, tutorial dialog systems aim to provide feedback in natural language using a wide range of tutoring tactics similar to those employed by a human teacher; they facilitate high interactivity and provide opportunities to reflect on existing (right or wrong) knowledge and integrate new knowledge [35].

To achieve a fluent conversation in a tutoring dialogue, the dialog manager must be flexible, prompt and goal oriented. The system must be flexible to accommodate to any input from the student. In addition, it must be prompt to take advantage of learning opportunities that arise and, above all, to ensure that the existing dialogue plans are fulfilled without failure while providing appropriate feedback. Therefore, a dialog manager must emulate the pragmatics of human-human tutoring dialogues [34].

In the literature, dialog systems are also called conversational agents (CAs) [36]. There are two types of CAs (not mutually exclusive): linguistic conversational agents that handle the conversation in written or spoken form [37] and embodied conversational agents that give a more realistic appearance to the dialogue by using the attributes of an animated humanoid body, i.e., facial expressions, movement of the mouth, and eye gaze [38].

III. METHOD

We conducted an SLR following the guide proposed by Kitchenham and Charters [15]. According to this guide, the process includes several activities, which can be grouped into the following three phases: planning of the SLR (research questions, search strategy, inclusion and exclusion criteria), conducting the SLR (selection process of primary studies, data extraction and synthesis) and reporting the SLR (results and analysis). These activities are described below.

A. RESEARCH QUESTIONS

This systematic review responds to the following research questions:

RQ1: What are the essential features of the existing ITSs that have natural language dialogue?

RQ2: What is the main purpose of the tutoring dialogue in each system?

RQ3: What are the instructional approaches and the support resources used by the ITSs with natural language dialogue?

RQ4: What is the NLU approach implemented by each system?

RQ5: How have the ITSs with natural language dialogue been evaluated empirically?

B. SEARCH STRATEGY

The search was carried out within the period between January 1998 and January 2018 in the following digital libraries: ACM Digital Library, Elsevier Scopus, Elsevier Science Direct, IEEE Xplore, Semantic Scholar and SpringerLink.

The search query was defined by using different keywords related to intelligent tutoring systems and dialog systems, in the following way: (‘‘Intelligent Tutoring*’’ OR ‘‘Tutor*’’ OR ‘‘Pedagogical Agent’’ OR ‘‘Intelligent Learning

Environment'') AND (''Dialogue*'' OR ''Tutorial Dialogue System'' OR ''Natural Language *'' OR ''Conversational'') AND (''Teaching'' OR ''Educational Training'' OR ''Procedural Training'').

To ensure that we were not leaving out relevant documents (false negatives), we repeated the search with new synonymous keywords until the search did not return any new document that met the inclusion criteria.

C. INCLUSION AND EXCLUSION CRITERIA

We included studies that had been peer-reviewed; describe some contributions of ITSs whose tutoring is based on dialogue with natural language; explain the main purpose of the dialogue; specify the adopted NLU approach; and describe how the tutoring dialogue was conducted. For some ITSs, we also included papers that described further details of their NLU approach or empirical studies related to them.

We excluded studies based on the following: those that are surveys, not peer-reviewed, duplicated, or not written in English; if they describe conversational agents that have not been integrated into ITSs; or if either their tutoring process or their NLU approach is not explained in sufficient detail.

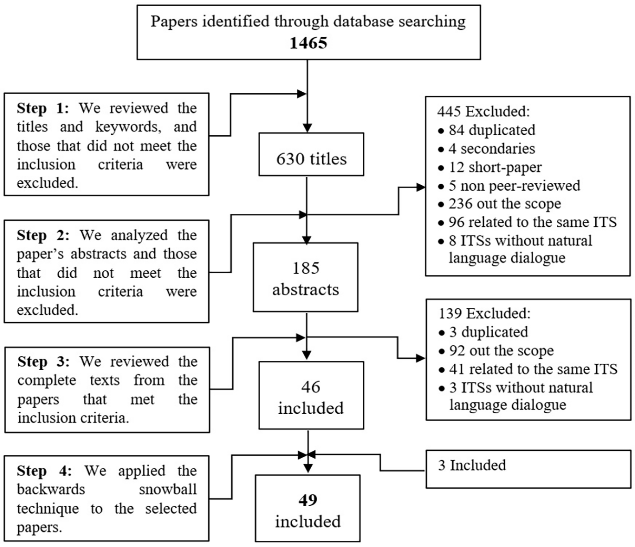
Papers identified through database searching
1465
Step 1: We reviewed the titles and keywords, and those that did not meet the inclusion criteria were excluded.
630 titles
445 Excluded:
• 84 duplicated
• 4 secondaries
• 12 short-paper
• 5 non peer-reviewed
• 236 out the scope
• 96 related to the same ITS
• 8 ITSs without natural language dialogue
Step 2: We analyzed the paper's abstracts and those that did not meet the inclusion criteria were excluded.
185 abstracts
139 Excluded:
• 3 duplicated
• 92 out the scope
• 41 related to the same ITS
• 3 ITSs without natural language dialogue
Step 3: We reviewed the complete texts from the papers that met the inclusion criteria.
46 included
3 Included
Step 4: We applied the backwards snowball technique to the selected papers.
49 included

FIGURE 1. Paper selection flowchart.

D. SELECTION PROCESS

As outlined in Figure 1, the selection process consisted of four steps. In step 1, the papers identified through database searching were screened while accounting for the title and keywords. Then, in step 2, we reviewed the abstracts. As a result of this review, we selected 185 out of 630 papers. Next, we examined in detail the full text of those 185 papers in step 3, and we were left with 46 papers. Finally, in step 4, we applied the backward snowball technique to the selected papers, and we obtained three more papers, reaching a total of 49 primary studies.

E. DATA EXTRACTION

To answer the five research questions, apart from the id, the system name, and the paper references, we defined the following nine features:

- ITS type, knowledge area and educational level. (addressing RQ1).
- Purpose of the dialogue (addressing RQ2).
- Instructional approach, support resource and test format (addressing RQ3).
- NLU approach (addressing RQ4).
- Empirical evaluation (addressing RQ5).

IV. RESULTS AND ANALYSIS

Of the 49 primary studies analyzed, a total of 33 ITSs were identified. Table 1 presents the general information on each study: system's name, reference, relationship to research question, paper type, venue, and year of publication; and Table 2 shows a summary of the main characteristics of each ITS. Furthermore, we also tabulated the data with regard to each research question in Tables 3-7.

A. RQ1: ESSENTIAL FEATURES OF AN ITS

In this research question, we sought to identify the essential features of the ITSs included in this study. To accomplish this task, we chose the three essential features shown in Table 3. For each feature, we defined some categories based on the content of the selected studies and other surveys. Next, we detailed how the ITSs are classified by using these categories for each essential feature.

1) ITS TYPE

For this essential feature, we will use the classification mentioned in section II.A.

a: EXPECTATION AND MISCONCEPTIONS TAILORED

Table 3 shows that the ITSs of EMT type are the most frequent. This approach is quite common in human tutoring [39], [40]. Basically, these systems model student knowledge through matching the students' answers with a list of good answers that represent expectations (learning objectives) and/or a list of anticipated misconceptions in the domain [30]. However, not all of these ITSs work with both expectations and misconceptions: some of them focus on only one of these sets to provide feedback, as will be explained below.

In this group of EMT systems, we can find Autotutor [41] and its versions, namely, Aries [42], Autotutor 3D [43], Autotutor Affect-Sensitive [44], AutoTutor Lite [45], DeepTutor [46], Gaze Tutor [47], Guru [48], and Why2-Autotutor [49]. All of the ITSs of the Autotutor family share a tutoring strategy based on both expectations and misconceptions.

AutoTutor Lite is a web-based version designed for integration into third-party systems that works with a simplified student model called learner's characteristic curves based on the relevance and newness of the student contributions [50].

TABLE 1. Data extraction form.

ITS name	Ref.	Relationship with RQ	Article Type	Venue	Year
Abdullah	[54]	1,2,3,4,5	Conference	International Conference on Agents and Artificial Intelligence	2015
Aries	[42]	1,2,3,4,5	Chapter	Serious Games and Edutainment Applications	2011
	[92]	4	Conference	International Conference on Computing and Intelligent Systems	2011
Atlas-Andes	[70]	1,2,3,4,5	Conference	Artificial Intelligence in Education	2001
AutoTutor	[41]	1,2,3,4,5	Journal	Cognitive Systems Research	1999
	[112]	5	Journal	Interactive Learning Environments	2000
	[107]	5	Conference	Artificial Intelligence in Education	2003
AutoTutor 3D	[43]	1,2,3,4,5	Conference	28th Annual Meeting of the Cognitive Science Society	2006
AutoTutor Affect-Sensitive	[44]	1,2,3,4,5	Journal	ACM Transaction Interactive Intelligent System	2012
AutoTutor Lite	[45]	1,2,3,4,5	Journal	International Journal of Learning Technology	2015
Beetle II	[56]	1,2,3,4,5	Journal	International Journal of Artificial Intelligence in Education	2014
	[86]	4,5	Conference	International Conference on Artificial Intelligence in Education	2013
	[102]	5	Demo	Association for Computational Linguistics	2010
	[103]	5	Conference	Association for Computational Linguistics	2010
	[104]	5	Conference	Sustaining TEL: From Innovation to Learning and Practice	2010
BRCA Gist	[52]	1,2,3,4,5	Journal	Learning and Individual Differences	2016
CIRCSIM-Tutor	[57]	1,2,3,4,5	Journal	Artificial Intelligence in Medicine	2006
	[83]	4	Conference	International Conference on Artificial Intelligence in Education	2001
	[108]	5	Conference	Int. Conf. on Information Technology: Coding and Computing	2004
DeepTutor	[46]	1,2,3,4	Chapter	AI Magazine	2013
	[100]	5	Conference	International Conference on Intelligent Tutoring Systems	2014
Dialog	[58]	1,2,3,4	Conference	Cognitive Systems	2007
EER-Tutor	[76]	1,2,3,4	Workshop	2006 Workshop on ITS for Ill-defined Domains	2006
	[101]	5	Conference	International Conference on Computers in Education	2010
Gaze Tutor	[47]	1,2,3,4,5	Journal	International Journal Human - Computer Studies	2012
Geometry Explanation Tutor	[59]	1,2,3,4,5	Conference	International Conference on Intelligent Tutoring Systems	2004
Guru	[48]	1,2,3,4	Chapter	Advances in Intelligent Tutoring Systems	2010
	[94]	4	Chapter	Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches	2012
	[109]	5	Conference	International Conference on Intelligent Tutoring Systems	2012
ITSpoke	[60]	1,2,3,4,5	Journal	Speech Communication	2006
Jacob	[71]	1,2,3,4	Conference	Advances in Multimodal Interfaces - ICMI 2000	2000
Kermit-SE	[75]	1,2,3,4,5	Journal	International Journal of Knowledge-Based and Intelligent Engineering Systems	2006
Lana	[55]	1,2,3,4	Conference	IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications	2017
Ms. Lindquist	[73]	1,2,3,4,5	Journal	International Journal of Artificial Intelligence in Education	2008
My Science Tutor	[61]	1,2,3,4,5	Journal	ACM Transactions on Speech and Language Processing	2011
Normit-SE	[74]	1,2,3,4,5	Conference	International Conference on Artificial Intelligence in Education	2005
Oscar	[62]	1,2,3,4,5	Journal	Computers & Education	2014
Paco	[72]	1,2,3,4,5	Conference	International Conference on Intelligent Tutoring Systems	2002
ProPL	[63]	1,2,3,4,5	Journal	Computer Science Education	2005
ReportTutor	[64]	1,2,3,4,5	Journal	Advances in Health Sciences Education	2008
Rimac	[65]	1,2,3,4,5	Conference	International Conference on Artificial Intelligence in Education	2015
RMT	[66]	1,2,3,4,5	Journal	Behavior Research Methods	2008
SCoT	[77]	1,2,3,4,5	Workshop	Workshop on Dialogue-based Intelligent	2004
VCAEST	[53]	1,2,3,4,5	Conference	Human Factors and Ergonomics Society	2016
Why2-Atlas	[67]	1,2,3,4	Conference	International Conference on Intelligent Tutoring Systems	2002
	[96]	4,5	Conference	International Conference Florida Artificial Intell. Research Society	2006
Why2-AutoTutor	[49]	1,2,3,4,5	Conference	25th Annual Conference of the Cognitive Science Society	2003
	[110]	5	Conference	International Conference on Intelligent Tutoring Systems	2004
	[111]	5	Journal	Cognitive Science	2007

In addition, AutoTutor Lite supports shareable knowledge objects (SKO) to create the materials to be learned by the students [51].

BRCA Gist [52] and VCAEST [53] were developed using AutoTutor Lite for the generation of tutorial dialog.

Although the application of the EMT strategy is not explicitly indicated in the studies related to Abdullah [54], Lana [55], Beetle II [56], CIRCSIM-Tutor [57], Dialog [58], Geometry Explanation Tutor [59], ITSpoke [60], My Science Tutor [61], Oscar [62], ProPL [63], ReportTutor [64], Rimac [65], RMT [66] and Why2-Atlas [67], after analyzing the way in which they perform the tutoring, we concluded that they belong to this type of ITS. Next, let us justify this conclusion for each of these systems.

Abdullah and Lana compare students' short answers with the expected right answers.

TABLE 2. Summary of ITSs characteristics.

RQ	Characteristics		ITS																														
			Abdullah	Aries	Atlas-Andes	AutoTutor	AutoTutor 3D	AutoTutor AS	AutoTutor Lite	Beetle II	BRCA Gist	CIRCSIM-Tutor	DeepTutor	Dialog	EER-Tutor	Gaze Tutor	Geom. Exp. Tutor	Guru	ITSpoke	Jacob	Kermit-SE	Lana	Ms. Lindquist	My Science Tutor	Normit-SE	Oscar	Paco	ProPL	ReportTutor	Rimac	RMT	SCoT	VCAEST	Why2-Atlas	Wh2-AutoTutor
1	ITS type	Expectation and misconceptions tailored	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x			x		x		x		x	x		x		x	x	x
		Model-tracing tutor			x															x			x				x								
		Constraint-based model													x						x				x										
		Bayesian network model																														x			
	Knowledge area	Natural science			x		x			x			x			x		x	x			x		x						x				x	x
		Computer science				x		x							x						x				x	x		x							
		Others	x	x					x											x							x				x	x			
		Medicine									x	x																	x				x		
		Mathematics												x			x						x												
	Educational level	College			x	x	x	x		x	x	x		x	x	x			x		x				x	x		x			x			x	x
		High school		x									x				x	x					x				x			x		x			
		Others																				x							x				x		
		Elementary school	x																					x											
		Not reported							x											x															
2	Understand level	Explanation of concepts		x		x		x	x		x					x		x						x							x				
		Short answer questioning	x																			x													
		Explanation of executed actions																														x			
	Apply level	Justification of solutions					x			x			x						x											x			x	x	x
		Application of rules, laws, and theorems			x							x		x			x						x												
		Execute an exercise																		x					x		x								
		Resolution based on schemas																								x		x							
	Analyze level																												x						
	Create level														x						x														
3	Test formats	Long answer		x		x	x	x	x	x	x		x	x	x	x	x	x	x		x			x	x				x		x	x	x	x	x
		Short answer	x		x							x								x		x	x			x	x	x		x					
4	Symbolic		x		x					x		x		x	x					x	x	x	x	x	x	x	x	x	x	x		x			
	Statistical					x	x	x	x		x																				x		x		x
	Hybrid			x									x			x	x	x	x															x	
5	Impact of different tutoring strategies	Comparison of different strategies supported by dialogue		x				x					x		x	x										x			x	x					
		Tutoring feedback supported by dialog vs limited tutoring feedback			x					x							x				x		x		x							x			
		Tutoring feedback supported by dialog vs limited tutoring feedback vs random feedback							x																										
	ITS versus traditional learning method	Short-term learning	x	x		x	x			x	x	x						x					x					x			x		x	x	
		Long-term learning																																	x
	Student impressions		x							x		x									x					x		x	x				x		
	Component evaluation	NLU			x	x																		x											
		Voice recognizer																						x								x			
		Emotion recognizer																	x																
	Not reported													x						x		x													
	Positive argumentation																										x								

Beetle II uses a knowledge base to verify the factual and the explanation correctness of the students' responses. Thus, knowledge base can be considered a set of explicit and inferred expectations.

CIRCSIM-Tutor keeps a student model that represents the student's errors and their possible misconceptions.

Then, we have some ITSs (Rimac, ITSpoke, Why2-Atlas and ProPL) that have been implemented by employing tutorial dialog toolkits, such as TuTalk [68] and Atlas [69]. These toolkits support the definition of flexible tutoring strategies, which can rely on expectations and/or typical errors.

Dialog addresses misconceptions that are detected in students' proofs. To detect these misconceptions, the system relies on a domain reasoner that evaluates the soundness, granularity (acceptable size of the proof) and relevance (useful with respect to the goal) of the proof steps with respect to the expectations.

Geometry Explanation Tutor focuses on helping students to improve explanations that appear to aim at the right idea but are not sufficiently precise with respect to the expected explanation.

My Science Tutor, during spoken dialogues, asks questions that are designed to elicit student responses that will map to the elements of the targeted semantic frames (statements to be learned). Information extracted from student responses is integrated into the session context that represents which points have been addressed by the student, which have not, which were expressed correctly, and which represented misconceptions.

Oscar relies on erroneous or incomplete solutions to provide feedback.

ReportTutor assesses the correction of students' reports by checking if they contain some erroneous or missing information.

RMT engages students in a natural language dialogue, evaluating student responses against sets of expected answers.

b: MODEL-TRACING TUTOR

These ITSs (Atlas-Andes [70], Jacob [71], Paco [72] and Ms. Lindquist [73]) monitor the learning progression of the students by keeping track of their actions as performed through the user interface. Thus, these ITSs employ a dialogue to help the student advance towards resolution of a problem.

c: CONSTRAINT-BASED MODEL

The ITSs of CBM type (Normit-SE [74], Kermit-SE [75] and EER-Tutor [76]) model the domain knowledge by means of a set of constraints that have to be met during the problem-solving process. So, the violation of a constraint will reveal a knowledge gap or misconception that will have to be corrected in the student's knowledge. As we will explain later on, the support for the self-explanation will play a key role in the revision of the student's knowledge.

d: BAYESIAN NETWORK MODEL

The BNM type is used to represent a student model of multiple variables with which you can determine the probability that the student has acquired some knowledge or has a misconception. The Bayesian network is an appropriate modelling option because the task of inferring a student's cognitive state from their responses to questions involves a great deal of uncertainty. According to the selected studies, SCoT [77] is the only member of this subgroup.

2) KNOWLEDGE AREA

Each of the referenced ITSs addresses one of the following knowledge areas: natural science, computer science, medicine, mathematics, and others. From these, the ITSs of natural science in the domains of physics, biology and electronics stand out to a greater extent, followed by the ITSs of computer science.

In addition to showing their areas of knowledge, it is worth noting that ITSs provide tutoring in three different languages: Dialog (German), Abdullah and Lana (Arabic) and the other ITSs (English).

3) EDUCATIONAL LEVEL

This essential feature refers to the educational level at which the ITS has been utilized. We will consider the following educational levels: elementary school, high school, college, and others. This essential feature highlights a higher percentage of ITSs aimed at the college level. For AutoTutor Lite and Jacob, this information was not reported in any paper.

B. RQ2: PURPOSE OF DIALOGUE

This research question was used with the intention of knowing which type of cognitive processes are supported by each ITS. To classify the ITSs, as shown in Table 4, we considered the levels of cognitive processes defined in the revised edition of Bloom's taxonomy [78]. In turn, in the first two levels (understand and apply levels), we distinguished different subgroups of ITSs that account for how students demonstrate their knowledge of the target matter.

1) UNDERSTAND LEVEL

This type of system typically adapts the dialogue to facilitate the building of connections between new knowledge and prior knowledge in the students' mind. In this sense, the dialogues serve several purposes: to help students provide correct answers to questions and explanations; to remedy misconceptions; and to summarize the addressed topics to make clear the learning of the main concepts.

a: EXPLANATION OF CONCEPTS

Within the group of systems at the understand level, most of the systems are aimed at making explanations of concepts. In this subgroup are Aries, AutoTutor, AutoTutor Affective – Sensitive, AutoTutor Lite, BRCA Gist, Gaze Tutor, Guru, My Science Tutor, and RMT. All of these systems require

TABLE 3. Essential features of an ITS.

Essential Features		ITS	Freq.	%
ITS type	Expectation and misconceptions tailored	Abdullah, Aries, AutoTutor, AutoTutor 3D, AutoTutor Affect – Sensitive, AutoTutor Lite, Beetle II, BRCA Gist, CIRCSIM-Tutor, DeepTutor, Dialog, Gaze Tutor, Geometry Explanation Tutor, Guru, ITSpoke, Lana, My Science Tutor, Oscar, ProPL, ReportTutor, Rimac, RMT,VCAEST, Why2-Autotutor, Why2-Atlas.	25	75.76
	Model-tracing tutor	Atlas-Andes, Jacob, Ms. Lindquist, Paco.	4	12.12
	Constraint-based model	EER-Tutor, Kermit-SE, Normit-SE.	3	9.09
	Bayesian network model	SCoT	1	3.03
Knowledge area	Natural science	Atlas-Andes, AutoTutor 3D, Beetle II, DeepTutor, Gaze Tutor, Guru, ITSpoke, Lana, My Science Tutor, Rimac, Why2-Autotutor, Why2-Atlas.	12	36.36
	Computer science	Autotutor, AutoTutor Affect – Sensitive, EER-Tutor, Kermit-SE, Normit-SE, Oscar, ProPL.	7	21.21
	Others	Abdullah, Aries, AutoTutor Lite, Jacob, Paco, RMT, SCoT.	7	21.21
	Medicine	BRCA Gist, CIRCSIM-Tutor, ReportTutor, VCAEST.	4	12.12
	Mathematics	Dialog, Geometry Explanation Tutor, Ms. Lindquist.	3	9.09
Educational level	College	Atlas-Andes, AutoTutor, AutoTutor 3D, AutoTutor Affect – Sensitive, Beetle II, BRCA Gist, CIRCSIM-Tutor, Dialog, EER-Tutor, Gaze Tutor, ITSpoke, Kermit-SE, Normit-SE, Oscar, ProPL, RMT, Why2-Autotutor, Why2-Atlas.	18	54.55
	High school	Aries, DeepTutor, Geometry Explanation Tutor, Guru, Ms. Lindquist, Paco, Rimac, SCoT.	8	24.24
	Others	Lana, ReportTutor, VCAEST.	3	9.06
	Elementary school	Abdullah, My Science Tutor.	2	6.06
	Not reported	AutoTutor Lite, Jacob.	2	6.06

TABLE 4. Purpose of the dialogue.

Purpose of Dialogue		ITS	Freq.	%
Understand level	Explanation of concepts	Aries, AutoTutor, AutoTutor Affective – Sensitive, AutoTutor Lite, BRCA Gist, Gaze Tutor, Guru, My Science Tutor, RMT.	9	27.27
	Short answer questioning	Abdullah, Lana.	2	6.06
	Explanation of executed actions	SCoT.	1	3.03
Apply level	Justification of solutions	AutoTutor 3D, Beetle II, DeepTutor, ITSpoke, Rimac, VCAEST, Why2-Atlas, Why2-AutoTutor.	8	24.24
	Application of rules, laws and theorems	Atlas-Andes, CIRCSIM-Tutor, Dialog, Geometry Explanation Tutor, Ms. Lindquist.	5	15.15
	Execute an exercise	Jacob, Normit-SE, Paco.	3	9.09
	Resolution based on schemas	Oscar, ProPL.	2	6.06
Analyze level		ReportTutor.	1	3.03
Create level		EER-Tutor. Kermit-SE.	2	6.06

students to provide explanations in response to their questioning rather than short answers that could lead to superficial knowledge. In the case of Guru, it presents students with brief lectures from which they produce summaries, complete conceptual maps and finish cloze tasks.

b: SHORT ANSWER QUESTIONING

Abdullah and Lana comprise another subgroup of ITSs, which share the feature of formulating short answer questions to the students.

c: EXPLANATION OF EXECUTED ACTIONS

Another ITS of this group is SCoT. However, instead of commenting on the student's answers, SCoT examines the actions that the student performed in a DC-Train simulator and then involves him/her in a reflective dialogue [79].

2) APPLY LEVEL

The ITSs of this group generate dialogues in natural language to accompany the students in problem solving, guiding the student through the solving process. Nevertheless, some of these ITSs wait for students to complete the resolution of a problem and then engage them in dialogues to review their comprehension of the concepts. Here, it is worth noting that this approach is underpinned by the study presented in [80]. According to its conclusions, students can ask more questions and better discuss their reasoning processes in the dialogues that are generated after the resolution of a problem than in the dialogues generated during the resolution of the problem.

a: JUSTIFICATION OF SOLUTIONS

Most of the ITSs in this subgroup (AutoTutor 3D, DeepTutor, ITSpoke, Why2-Atlas and Why2-AutoTutor) pose a physics problem to students and then help the students to solve the

TABLE 5. Pedagogical features of ITSs.

Pedagogical Feature		ITS	Freq.	%
Instructional approach	Generation of explanations to justify solutions	Aries, AutoTutor, Autotutor 3D, AutoTutor Affective – Sensitive, AutoTutor Lite, Beetle II, BRCA Gist, DeepTutor, Gaze Tutor, Guru, ITSpoke, My Science Tutor, Rimac, RMT, SCoT, VCAEST, Why2-Atlas, Why2-AutoTutor.	18	54.55
	Support for problem solving	Atlas-Andes, CIRCSIM-Tutor, Dialog, EER-Tutor, Geometry Explanation Tutor, Kermit-SE, Ms. Lindquist, Oscar, ProPL, ReportTutor.	10	30.30
	Clarify and direct procedures	Paco, Jacob, Normit-SE.	3	9.09
	Ask questions - answer	Abdullah, Lana.	2	6.06
Support resources	Animated agent	Aries, AutoTutor, AutoTutor 3D, AutoTutor Affect-Sensitive, AutoTutor Lite, BRGA Gist, DeepTutor, Gaze Tutor, Guru, Jacob, Lana, My Science Tutor, RMT, VCAEST, Why2-AutoTutor.	15	44.12
	Images	Abdullah, Atlas-Andes, AutoTutor, AutoTutor Affective-Sensitive, AutoTutor Lite, BRCA Gist, DeepTutor, Gaze Tutor, Guru, Geometry Explanation Tutor, My Science Tutor, Oscar, RMT.	13	38.24
	Audio and video	Abdullah, Aries, Lana, My Science Tutor, BRCA Gist, Guru, ITSpoke, Oscar, Rimac, SCoT, VCAEST.	11	32.35
	Simulation	AutoTutor 3D, Beetle II, Jacob, My Science Tutor, Paco, SCoT, VCAEST.	7	20.59
	Option menu	EER-Tutor, Kermit-SE, Ms. Lindquist, Normit-SE, Paco, Rimac.	6	17.65
	Conceptual maps	CIRCSIM-Tutor, Guru.	2	5.88
	Not reported	ProPL, Why2-Atlas.	2	5.88
	Table	CIRCSIM-Tutor.	1	2.94
	Virtual slides	ReportTutor	1	2.94
Input Text formats	Long answer	Aries, AutoTutor, AutoTutor 3D, AutoTutor Affective-Sensitive, AutoTutor Lite, Beetle II, BRCA Gist, DeepTutor, Dialog, EER-Tutor, Gaze Tutor, Geometry Explanation Tutor, Guru, ITSpoke, Kermit-SE, My Science Tutor, Normit-SE, ReportTutor, RMT, SCoT, VCAEST, Why2-Autotutor, Why2-Atlas.	23	69.70
	Short answer	Abdullah, Atlas-Andes, CIRCSIM-Tutor, Jacob, Lana, Ms. Lindquist, Oscar, Paco, ProPL, Rimac.	10	30.30

problem step by step by means of a mixed initiative dialogue. In some steps, students are asked to justify their right or wrong decisions/actions.

In contrast, Rimac and VCAEST wait for students to have completed the resolution of the problem, and then, they engage students in a reflective dialogue.

b: APPLICATION OF RULES, LAWS, AND THEOREMS

Unlike the ITSs of the previous group, this type of ITS supports dialogs based on a mixed language that combines natural language and a formal language. Thus, during the resolution of the problem, these ITSs help students to build a mathematical expression or fill in a table by applying some rules, laws, or theorems.

For example, in Atlas-Andes, students must solve problems related to acceleration with the support of the Conceptual Helper, which offers unsolicited help.

CIRCSIM-Tutor generates a dialog based on a student's prediction on seven cardiovascular parameters previously submitted in a table.

In Dialog, the students must prove a theorem in set theory.

Geometry Explanation Tutor asks students to explain in their own words each of the steps that they have followed to solve a problem related to the angle theorems.

Ms. Lindquist aids the student in building an algebra formula from a problem statement by providing feedback step by step.

c: EXECUTE AN EXERCISE

In this subgroup, we have the ITSs (Jacob, Normit-SE and Paco) that help students to perform tasks by applying predefined procedures step by step. In this context, a procedure consists of a sequence of steps that must be followed in a fixed order.

d: RESOLUTION BASED ON SCHEMAS

In this subgroup, we can find ITSs (Oscar and ProPL) that work with schemas of the solution and guide students to identify and/or instantiate the required schemas or templates. In the case of ProPL, firstly, students have to identify programming goals (decomposition task) from a problem statement with the help of the tutor. Afterward, they have to identify the right schema to attain each goal by means a tutor supported dialogue and provide details about how to instantiate the schemas for the problem at hand (composition task).

3) ANALYZE LEVEL

In this level, we included only ReportTutor, because it is the only ITS that involves the analysis of some pieces of information to elaborate a diagnosis report, specifically, on melanoma cases. In this sense, it helps students learn how to document and identify all relevant features found in some virtual slides.

TABLE 6. NLU approach.

NLU approach	ITS	Freq.	%
Symbolic	Abdullah, Atlas-Andes, Beetle II, CIRCSIM-Tutor, Dialog, EER-Tutor, Jacob, Kermit-SE, Lana, Ms. Lindquist, My Science Tutor, Normit-SE, Oscar, Paco, ProPL, ReportTutor, Rimac, SCoT.	18	54.55
Statistical	AutoTutor, AutoTutor 3D, AutoTutor Affect - Sensitive, AutoTutor Lite, BRCA Gist, RMT, VCAEST, Why2-AutoTutor	8	24.24
Hybrid	Aries, DeepTutor, Gaze Tutor, Geometry Explanation Tutor, Guru, ITSpoke, Why2-Atlas.	7	21.21

4) CREATE LEVEL

This level comprises only two ITSs (EER-Tutor and Kermit-SE) that support the creation of an Entity Relationship data model. While using these ITSs, students are free to make the design decisions they consider necessary to build a correct data model, provided that they meet certain domain constraints.

C. RQ3: PEDAGOGICAL FEATURES

This research question was posed to know the pedagogical features of the ITSs. To accomplish this task, we established three relevant features related to the tutoring process: the instructional approach, the support resources on which it relies, and the input format of the texts through which the ITSs evaluate the knowledge acquired by the students. The categories for each feature that we present in Table 5 were defined: for the instructional approach from the considered studies; for the support resource we were inspired by the section on complementary media described in [12]; and for the input text format, we based it on the classification established in [81].

1) INSTRUCTIONAL APPROACH

The instructional approach is the strategy followed by the ITS to ensure that the students learn the target matter and reach their learning objectives.

Even though we know that the ITS instructional approach is generally inspired by the theory of Socratic and constructivist learning, we examined the approach that each ITS carries out during the teaching process.

As we could not find in the related literature any classification that fits the specific requirements of this study, we developed our own classification defined by means of the four categories outlined in Table 5.

Additionally, we want to note that the descriptions of the ITSs shown below are based on what is explained in the selected studies. Therefore, we may be leaving out some relevant features of these ITSs that were not mentioned in the selected studies.

a: GENERATION OF EXPLANATIONS TO JUSTIFY SOLUTIONS

In this instructional approach, the ITS enables the student to actively elaborate explanations and justifications of a previous student input (e.g., a previous student's answer, a student prediction related to a simulation) in a turn-based dialog. In each turn, the ITS compares the student's response to the expectations (right answers) and/or misconceptions that are prepared for each question. To facilitate the treatment of misconceptions, the ITS has a set of anticipated incorrect answers (bugs) and their corresponding remediations.

We found that most of the ITSs implement this approach, for example, AutoTutor and its descendants. AutoTutor presents the student with a deep question and encourages him/her to provide a sufficiently detailed answer in a ‘‘hint – prompt – assertion’’ cycle [82].

The AutoTutor dialog mechanism has been implemented in all of its descendants with certain variations. For example, AutoTutor 3D, in addition to asking deep questions to the students, also requests predictions about simulations and responds to them.

In Aries, the dialog mechanism is similar, but it differs in that the dialog has a conversation of three-way ‘‘trialogs’’ (the participant and two virtual agents, Dr. Quinn and Glass). There are three types of trialogs. A standard trialog occurs when Glass observes Dr. Quinn teach the player. A vicarious trialog occurs when the player watches Dr. Quinn teach Glass. Finally, a teaching trialog occurs when the player teaches Glass as Dr. Quinn observes. The type of trialog that occurs for a particular question depends on the level of knowledge exhibited by the player. All of the trialogs follow the same ‘‘question – hint – prompt - summary’’ sequence.

Guru's authors proposed an extended dialog cycle including the following phases ‘‘direct instruction – prompt – feedback – verification – question – feedback’’ to enhance the effectiveness of the evaluations of the student's comprehension.

AutoTutor Affective-Sensitive can recognize the affective and cognitive states of the student, and based on them, adapts their dialog movements.

BRCA Gist and VCAEST apply deep reasoning questions supplemented with simpler self-reflection questions and indirect tutoring.

Apart from Autotutor and its descendants, there are other remarkable ITSs that implement this instructional approach, as we will show below.

In the case of Beetle II, it provides dynamic feedback from a specific context, fostering a reflexive dialog through a ‘‘predict – check – evaluate’’ cycle. The specific context is based on the simulations in which the students are asked to predict the behavior of the circuit and explain the prediction. Then, they must verify their predictions in the circuit simulator. Next, the system asks if the simulated results coincide with their predictions and requests students to explain what they have observed. After analyzing each student's answer, Beetle II can provide different types of feedback:

TABLE 7. Evidence in the evaluation of the ITSs.

Evidence		ITS	Freq.	%
Impact of different tutoring strategies	Comparison of different strategies supported by dialogue	Aries, AutoTutor Affective-Sensitive, DeepTutor, EER-Tutor, Gaze Tutor, Oscar, ReportTutor, Rimac.	8	24.24
	Tutoring feedback supported by dialog vs limited tutoring feedback	Atlas-Andes, Beetle II, Geometry Explanation Tutor, Kermit-SE, Ms. Lindquist, Normit-SE, SCoT.	7	21.21
	Tutoring feedback supported by dialog vs limited tutoring feedback vs random feedback	AutoTutor Lite.	1	3.03
ITS versus Traditional learning method	Short-term learning	Abdullah, Aries, AutoTutor, AutoTutor 3D, Beetle II, BRCA Gist, CIRCSIM-Tutor, Guru, Ms. Lindquist, ProPL, RMT, VCAEST, Why2-Atlas.	13	39.39
	Long-term learning	Why2-AutoTutor.	1	3.03
Student impressions		Abdullah, Beetle II, CIRCSIM-Tutor, Kermit-SE, Oscar, ProPL, ReportTutor, VCAEST.	8	24.24
Component evaluation	NLU	Atlas-Andes, AutoTutor, My Science Tutor.	3	9.09
	Voice recognizer	My Science Tutor, SCoT.	2	6.06
	Emotion recognizer	ITSpoke.	1	3.03
Not reported		Dialog, Jacob, Lana.	3	9.09
Positive argumentation		Paco.	1	3.03

acknowledging the correct part of the answer; suggesting a slide to read with background material; prompting for missing parts of the answer; hinting (with low or high specificity); and giving away the correct answer.

My Science Tutor asks questions based on what students see in interactive videos to help students provide answers that demonstrate their understanding of some topics. Follow-up questions and media presentations are designed to scaffold learning by providing hints about the important elements of the investigation that the student did not include in his/her explanation or misunderstood.

In Rimac, the dialogs are only initiated by the tutors, and the student initially issues short answers and then explanations. Rimac is designed to be sensitive to the level of abstraction of the student input at various points during the dialog. Thus, during the dialog, Rimac prompts the student to abstract or specialize his/her explanations, when appropriate.

SCoT formulates questions based on the actions performed by the student in the simulator DC-Train and discusses wrong or partially correct answers (related to wrong actions in the simulator) with the student by applying activity recipes to guide the dialog.

In ITSpoke and Why2-Atlas, the student must construct an essay that is then analyzed and used as a basis for a tutorial dialog. In this dialog, they expose criticisms of the essay and help student to rewrite it to address any defect that remains. In ITSpoke, if the student does not correct all of the defects, the tutor tries again but with a different dialog. Unlike ITSpoke and Why2-Atlas, Why2-AutoTutor focuses mainly on obtaining the correct contributions (expectations) of the student rather than correcting all of the defects. If all of the expectations are not met, it will resort to prompts and hints to guide the student toward them.

b: SUPPORT FOR PROBLEM SOLVING

The ITSs of this group can offer support for problem solving either on demand; when they detect a student's mistake or a student's misconception; or when the solution is not sufficiently complete. Next, let us mention the most relevant approaches implemented in these ITSs.

Atlas-Andes, Geometry Explanation Tutor and ProPL initiate their dialog when misconceptions are detected, or the solutions are not complete or precise enough. In the case of ProPL, it focuses on eliciting a right response from the student by: providing him/her a hint (e.g. point out something in the problem statement); or requesting the student a generalization or a synthesis that allows him/her to improve his/her previous answer (e.g. asking to imagine a scenario that provokes a program failure).

Geometry Explanation Tutor responses often take the form of questions, which is meant to make the student see that the rule he/she stated is overly general. After several iterations, the student must refine an explanation that initially was imprecise and overly general into an accurate statement of the geometry rule.

CIRCSIM-Tutor ask students qualitative predictions (increase, decrease, or does not change) on seven cardiovascular parameters. Then, if the ITS detects some erroneous prediction, it will begin a reflective dialogue with questions aimed at the revision of the misconception in the student's mind (e.g., erroneous cause-effect relations). This system generates a hint only when the student makes a mistake on the first try at a question. Otherwise, it gives the student the right answer.

In EER-Tutor and Kermit-SE, dialogs begin once the student has violated a restriction. There is a dialog for each type of error, and if there are several errors, the most appropriate error is selected for the discussion. In the case of Kermit-SE, the dialog consists of the following stages: (i) informs of the error and asks the student the reason for the decision; (ii) prompts him/her to understand why the decision was wrong; (iii) prompts him/her to give the right decision; and (iv) prompts him/her to review the domain concept related to the error through a question.

In ReportTutor, the participant makes a diagnosis from a bundle of virtual slides and proceeds to write a report. If he/she is lost in the writing, he/she can ask for help and receive suggestions for the item under study. When the report is finished, the system verifies whether it has neither errors nor missing attributes and provides visual feedback.

c: CLARIFY AND DIRECT PROCEDURES

The clarification and direction of the procedures is another instructional approach that we identified in the selected studies. This group could be a subgroup of the previous group, but we preferred to consider it a separated group to highlight that in this group, students need support to carry out a predefined procedure or task. Instead, the ITSs of the previous group are aimed to support the resolution of problems whose solution cannot be built step by step following a predefined procedure. We postulate that this difference has a clear impact in how the ITSs of these two groups guide students throughout the solving process. Therefore, it makes sense to present these two groups separately.

Within this group, we included Paco, Jacob and Normit-SE. Both Jacob and Paco provide instruction and assistance for tasks by giving hints on the next action to be done and immediate feedback on the executed actions. Paco also offers confirmation when the student does something correctly and corrections and encouragement when he/she makes an error. Additionally, in the case of Paco, the practice is posed as a collaborative work between the student and the tutor. So, if Paco believes that the student has sufficient knowledge to do the next action, it will expect him/her to do it. Otherwise, Paco will intervene and explain to the student what to do next.

Normit-SE supports self-explanation for students to explain themselves while solving the problem. More precisely, this explanation is requested for each action that is performed for the first time and in the case that an error is made. The student can also get a hint for every committed error. The correct solution is available only on request.

d: ASK QUESTIONS-ANSWER

This approach is the simplest to implement because the two ITS of this group pose questions based only on the fragments of information provided to the student before the dialog and then wait for short answers. If the student makes a mistake, these two ITSs react in a different way. On the one hand, Abdullah gives another chance to the student to correct his/her answer. If the student fails again to provide a correct answer, Abdullah shows it to the student and continues with the next question. On the other hand, Lana directly explains the right answer to the student and continues with the next question.

Additionally, Abdullah can classify students' answers into different levels of correction (highly correct, partially correct or near miss). In this way, it can reply to students with different types of encouragement.

Before closing this section, it is worth mentioning some ITSs that incorporate remarkable mechanisms to enhance the adaptability to the student. In this regard, we can highlight DeepTutor, Lana and Oscar.

DeepTutor extended the tutorial capability of AutoTutor with macro-adaptability. This macro adaptation system relies on learning progressions, to model and organize knowledge based on what is known about how students truly progress through the content that is being taught. Macro-adaptivity is particularly important when tutoring over longer periods than a few sessions with the tutor and when students start with unequal knowledge. Finally, Lana and Oscar consider learning styles to deliver personalized content during the dialog.

2) SUPPORT RESOURCES

Support resources refer to the types of resources with which each ITS supports its teaching process either during the dialogue, before or after. In any case, these resources serve to motivate the dialogue (e.g. simulating a physical phenomenon), to help students build utterances, or to describe concepts to be addressed in the conversation.

This feature can take some of the following eight values: animated agent, audio and video, conceptual maps, images, option menu, simulation, table and virtual slides. Some papers do not report the use of any specific material to support the teaching process.

It is worth clarifying that the "option menu" value is assigned to systems that provide menu options to help students build their utterances.

The results of this category are shown in Table 5 and allow us to see that animated agents and images are clearly the resources most used by ITSs because they facilitate the learning process and can easily be incorporated into the system. Almost half of the ITSs included in this SLR use animated agents. Among them, we find Autotutor and its descendants and other ITSs such as Jacob, Lana, My Science Tutor and Oscar.

Several ITSs employ simulations in different ways. In AutoTutor 3D, Beetle II, My Science Tutor, VCAEST and SCoT, the simulation is used to motivate the tutorial dialogue. On the other hand, in the case of Jacob and Paco, simulations are used to support the resolution of the problem guided by the tutor.

3) INPUT TEXT FORMAT

We distinguished two input text formats that were encouraged for student responses: long and short answer. A long answer or short essay is typically composed of at least 2 sentences and is the form of evaluation of most ITSs. Instead, short answers normally consist of at most two short sentences.

It is worth noting that even though Ms. Lindquist and Rimac admit only short answers in natural language, they support the construction of explanations by means of option menus.

D. RQ4. NLU APPROACH

This research question was proposed to determine the NLU approaches that the ITSs apply to understand the statements that a user can make during a conversation and relate them to a task that a user wants to perform. Table 6 shows the classification of each ITS according to the three types of approaches mentioned in Section II.B.

1) SYMBOLIC APPROACH

Table 6 shows that the symbolic approach is the most widely used by the ITSs.

In this group are Abdullah and Lana, which apply pattern matching for Arabic language treatment. Lana also resorts to short text similarity if the user input cannot be correctly recognized.

Jacob and the first versions of CIRCSIM-Tutor analyze students' comments through superficial semantic grammars. Subsequently, the last version of CIRCSIM-Tutor incorporated finite-state transducers, which permits the identification of key concepts included in one- or two-word answers [83].

Atlas-Andes and ProPL are supported by Atlas, which has an engine for NLU called CARMEL [84], which is in charge of analyzing the comments of the students. CARMEL consists of a comprehensive syntactic analyzer and robust and efficient algorithms necessary for semantic analysis and interpretation.

In Beetle II, students' utterances are analyzed by means of a process that includes two stages: in the first, the TRIPS dialog analyzer [85] generates a semantic representation that is domain independent; and in the second, the contextual interpreter applies a reference resolution approach and a set of rules to obtain a representation in terms of the Beetle II domain. Later, in [86], the group of researchers responsible for Beetle II presented a study of how to improve the robustness of the semantic interpreter. Basically, the improvement consisted of implementing a classifier based on lexical similarity within the symbolic approach.

Dialog interprets a mixture of natural language and mathematical expressions through a domain reasoner that requires deep syntactic and semantic analysis to obtain a formal representation of the steps followed by the student. As part of this process, the system processes the students' utterances by using the OpenCCG parser [87], which is grounded on a combinatorial grammar with a lexical base for German.

My Science Tutor incorporates the Phoenix analyzer [88] to extract an expression from the student's statement through a semantic grammar that works with entities, events and their relationships. Then, this expression is compared with patterns of system elements to assess its correctness.

Oscar and ReportTutor rely on pattern matching to address the grammatically incomplete or incorrect expressions that students usually introduce.

Paco interprets the students' utterances through an interpretation algorithm of Collagen's speech [89]. Collagen represents utterances using an artificial discourse language. This language is intended to include the types of utterances that people use when collaborating on tasks. For example, it includes utterance types for agreeing; asking or proposing how a task should be accomplished; asking what should be done next; etc. Thus, the available menu options in Paco for building utterances are defined according to these types of utterances.

Rimac is built on the TuTalk tutorial dialog toolkit [68]. Hence, it can employ different NLU methods because TuTalk supports the use of different NLU modules, implementing approaches such as minimum-distance (by default), LSA, and Naïve Bayes. As minimum-distance approach is symbolic, we classified Rimac as symbolic, but the use of TuTalk opens the possibility of adopting also statistical approaches.

SCoT uses a bidirectional unification grammar called Gemini [90].

2) STATISTICAL APPROACH

This approach has been applied by AutoTutor, many of its descendants and RMT.

As explained above, when the student provides an utterance, AutoTutor and its versions must compare this utterance with a set of anticipated expectations and misconceptions. To accomplish this goal, they apply LSA to determine whether two excerpts of text are conceptually similar [91]. LSA, for its part, creates two vectors, a vector that represents the semantic content of the student's utterance, and another vector that represents the semantic content of an expectation or misconception. Afterward, it employs the cosine function to calculate the conceptual similarity.

The LSA technique represents the meaning of text based on latent concepts that are automatically derived from an extensive collection of text (*corpus*). Nevertheless, even though it can detect that the student's response does not fit an ideal response, it cannot point out exactly which concepts and/or relations between concepts are wrong or missing in the student's response because LSA processes the response as a whole. The underlying limitation of LSA is that it ignores the sentence syntax and word ordering. As a result, LSA cannot handle negations or resolve term references (e.g., personal pronouns). Over time, some subsequent versions of AutoTutor have added significant improvements to the statistical approach to mitigate this limitation. In this sense, they incorporated different semantic evaluation algorithms, and therefore, these descendants of Autotutor became hybrid systems. In the next section, we will mention these systems (Aries, DeepTutor, Guru and Gaze Tutor).

On the other hand, we have AutoTutor Lite. This version of AutoTutor does not employ the full range of semantic analysis methods used in other Autotutor descendants but is limited to LSA and extended weighted keyword matching. As both BRCA and VCAEST are based on AutoTutor Lite, they do not take advantage of all of the semantics methods implemented in other descendants of AutoTutor.

3) HYBRID APPROACH

This approach is used by the ITSs Aries, DeepTutor, Gaze Tutor, Geometry Explanation Tutor, Guru, Why2-Altas and ITSpoke.

Aries applies both LSA and regular expressions to assess students' inputs [92]. Regular expressions were used in this system to describe the patterns of short answers and word variations.

DeepTutor designers adopted some of the semantic similarity methods implemented on a toolkit called SEMILAR [93]. This toolkit includes an optimal lexical matching solution to give a more sensitive treatment to the structure and semantic decomposition. The optimal lexical matching aims at finding an optimal global assignment of words in one sentence (student input) to words in the other sentence (expectation/misconception) based on word-to-word similarity while simultaneously maximizing the match between the syntactic dependencies.

Geometry Explanation Tutor relies primarily on a knowledge-based approach to recognize if the statements of the students are correct or partially correct, with which a semantic representation based on LCFlex [84] and a unifier of characteristic structures are created. When the knowledge-based approach fails, it resorts to a Naïve Bayes text classifier to classify the students' explanations with respect to a subset of categories. Thus, this system determines if the student is focusing on the correct geometry rule.

Guru and its descendant, Gaze Tutor, employ LSA but overcome the limitation of this approach mentioned in the previous section by relying on conceptual maps. By means of this technique, the domain model is expressed as a set of triples *(key_term, pedagogical_relation, proposition)*. Then, a set of triples are derived from students' utterances and compared with the triples that represent the domain model to identify different types of students' errors [94].

Why2-Atlas analyzes the contributions (essays) of students with two modules, a sentence-level understander (SLU) and a discourse-level understander (DLU). The SLU module works in the same way as in Geometry Explanation Tutor, and the DLU module receives the logical representations provided by the SLU and generates proofs using abductive reasoning [95].

To perfect the interpretation of the students' explanations in Why2-Atlas [32], researchers presented a method to heuristically combine multiple natural language understanding approaches. In [96], they proposed a new mechanism to analyze the explanations of the students. First, this new mechanism classifies the student's utterance into either an explanation or a short answer, and then, each type of utterance is processed in a different fashion.

The explanation is processed in two stages: 1) An analysis of the sentence, which generates a representation in first-order predicate logic, and 2) An evaluation of the accuracy and integrity of said representation.

The analysis of the sentence is performed following three different approaches: a) CARMEL [84], which provides syntactic and semantic analysis through LCFlex; b) RAINBOW [97], which provides text classifiers based on a bag-of-words model; and c) RAPPEL, obtained through MINIPAR [98], which uses syntactic dependency characteristics derived symbolically through templates that represent each proposition in the language. Next, to evaluate the logical representations obtained with each approach, they are matched with the nodes of an Assumptions-based Truth Maintenance System (ATMS) [99].

ITSpoke relies on the text-based Why2-Atlas dialog system in the back-end, with the novel approach that ITSpoke recognizes the emotions and attitudes of the students in the spoken dialog input.

E. RQ5. EMPIRICAL EVALUATION

The purpose of this question was to know what kind of evidence exists with regard to evaluations of ITSs with natural language dialogue. Next, we will briefly describe the evaluations found in the studies included in this review. However, it could happen that some of the referenced ITSs could have been evaluated in other ways (e.g., student surveys), but have not been reported in any published paper.

Additionally, we want to note that a detailed description of these studies would require another systematic review focused on this topic. Hence, in this section, we will only mention the main objective and some highlighted results for each study.

As shown in Table 7, we decided to classify the found evaluations into five types. Some of the referenced papers include evaluations of more than one type. In fact, all of the studies on the student impressions are published in papers that also comprise studies that compare different tutoring strategies or compare the learning gains provided by ITS and traditional learning methods, as shown in Table 8.

1) IMPACTS OF DIFFERENT TUTORING STRATEGIES

In this group, we included the experiments that have studied the performance of each ITS while working with different tutoring strategies. We distinguished three types of tutoring strategies in these experiments:

1. Tutoring feedback supported by dialog: these tutoring strategies represent the behavior of dialog-supported versions of the systems.
2. Limited tutoring feedback: these tutoring strategies present the operation of cut-down versions of the evaluated systems, which, for example, cannot reply to students' answers, or provide the complete solution without analyzing the student's response.
3. Random feedback: this strategy consists of providing random feedback regardless of the student input.

TABLE 8. Evaluation of ITSs versus traditional learning method.

ITS	A/B test	Pre/Post test	Control group	Effect size	Highlighted results
Short-term Learning					
Abdullah		X			The results revealed the learning gain and suggested a strong statistically significant relation between the student's scores before the tutoring (pretest) and after the tutoring (posttest).
Aries		X		1.3 *	The posttest scores were significantly higher than the overall pretest scores.
AutoTutor	X	X	TextBook	0.8	[12] AutoTutor produced learning gains that are quite higher than produced by control.
AutoTutor	X	X	Expert human tutor		[12] Learning gains are on par.
AutoTutor	X	X	Read text and nothing	0.31 and 1.23	[107] The results indicated that AutoTutor did not improve the performance of the students in evaluations of superficial learning, but it did in the performance ones of deep learning.
AutoTutor 3D		X			The results revealed that the students that utilized the simulations more effectively (by manipulating the relevant parameters) experienced more learning gains.
Beetle II	X	X	Distractor task (lesson) on another topic	1.72	Beetle II (ELICIT) provided high learning gains in comparison with the control group, but the control group did not receive specific training on the target topic.
BRCA Gist	X	X	Web pages and irrelevant tutorial with another ITS on a different topic	0.26 *	The BRCA Gist group scored significantly higher than the web pages group, and both scored significantly higher than the group that received an irrelevant tutorial. The difference between the BRCA Gist group and the web pages group was greater in the participants with a lower educational level.
CIRCSIM-Tutor (1)	X	X	Textual material	1.24 *	The result showed that the use of the system for one hour produces significant learning gains; additionally, even this brief usage improves more the student's ability to solve problems than reading textual material on the topic.
CIRCSIM-Tutor (2)		X			[108] The results indicated learning gain and showed that the scores were statistically significant.
Guru	X	X	Human tutoring and classroom	0.72 (Guru vs classroom)	[109] There were significant learning gains in the students with Guru and the human tutors in comparison with class control. On the other hand, although no difference was found between the groups of Guru and novice human tutors, the authors pointed out several limitations to this comparison.
Ms. Lindquist	X	X	CAI and classroom teachers	0.5 (Ms. Lindquist vs classroom)	Ms. Lindquist and a CAI solution outperformed the classroom teachers. The authors hypothesized that this finding was mainly due to the benefit of immediate feedback. Ms. Lindquist provided a slightly higher learning gain than CAI solution.
ProPL	X	X	Read text	0.58 (in composition tasks)	The experimental group demonstrated greater ability than the control group at composition tasks in terms of algorithm correction. Instead, in the case of the decomposition skill, no differences were found between the two groups.
RMT	X	X	Classroom instruction and CAI	0.71 and 0.34	This study revealed that the condition of dialog generated greater educational benefits than control conditions.
VCAEST	X	X	Live-action training		No significant differences in learning gains were found.
Why2-Atlas	X	X	Human tutoring and reading of a short text		[96] Why2-Atlas students scored higher than the short text and human tutoring students on one of three posttests and the same in the other two.
Why2-AutoTutor	X	X	Read textbook and nothing	1.02 (Why2-AutoTutor vs Read textbook)	The results showed that Why2-AutoTutor outperformed the other two approaches in the learning gain.
Long-term Learning					
Why2-AutoTutor	X	X	Minilesson (textbook style)	0.97 (post-test) 0.93 (retention test)	[110] The results of the multiple-choice tests and the trials showed learning gain from the pre-tests to the post-tests. This learning gain was maintained both in the retention test and the far transfer test, which was performed 1 week later.
Why2-AutoTutor	X	X	Canned text remediation		[111] In experiment 3 of 7, the essay results show that there was a very small but unreliable advantage for Why2-AutoTutor over the canned text remediation condition. This advantage was maintained in both the retention test and the far transfer test, which were performed 1 week later.

* They used a different formula than Cohen's to calculate the effect size.

Next, we will group the studies with regard to the types of strategies that were compared in each study.

a: COMPARISON OF DIFFERENT TUTORING STRATEGIES SUPPORTED BY DIALOG

Aries researchers studied how the different types of trialogs affected learning (the types of trialogs were explained in Section IV.C.1). They found that teaching trialog outperformed vicarious trialog. In addition, they observed that the trialogs had little impact on immediate testing but did have a significant impact after a two-day delay.

AutoTutor Affective-Sensitive was evaluated under three different conditions: regular AutoTutor, Support affective, and Shakeup affective. The difference between the Support and Shakeup approaches lies in the fact that while Support was designed assuming that the origin of the emotion is in the material to be learned, Shakeup was designed assuming that the source of the emotion is in the student. The experiment did not see a significant main effect for the tutor type but actually did show a slight tendency in favor of Support.

DeepTutor [100] was evaluated with college students under two training conditions: only micro-adaptive versus fully adaptive (macro and micro-adaptive). In only micro-adaptive condition, the ITS used a fixed, predefined set of instructional tasks for all students. In contrast, the ITS used in the fully adaptive condition could categorize students to different levels of understanding based on their pre-test score and then select appropriate tasks that were deemed most conducive of learning at that level of understanding. After comparing the results, they drew the conclusion that the learning gains from the fully adaptive condition were significantly greater than the gains from the other condition.

In [101], EER-Tutor researchers compared two versions of the system: with adaptive support and without adaptive support. The non-adaptive version provides the same dialog to two different students with different knowledge levels. Instead, the adaptive version can select the dialogues while considering previous student's errors. The learning gain of the group who received adaptive dialogues was significantly higher than the gain of the non-adaptive group.

In Gaze Tutor, which is an improved version of Guru, the effectiveness of the system was evaluated in two conditions: reactive to the gaze (using Gaze tutor) and non-reactive to the gaze (using Guru ITS). The gaze-reactive condition is related to a tutoring strategy in which the tutor monitors the student's gaze to detect when the student is bored, disengaged, or zoning out, and then, it attempts to reengage the student with dialog moves. The authors reported that the dialogs sensitive to the gaze were successful in reorienting the students and that the gaze-reactive approach was more effective than the non-reactive approach in promoting learning gains in the more gifted students.

Oscar was assessed in two experiments. The results unveiled that tutoring with personalization based on the learning style performed better than tutoring without this personalization.

In ReportTutor, the effect of the feedback timing was evaluated using two interfaces, one immediate and the other delayed. The analysis showed that a significant improvement was found in the writing of the reports under both conditions, but there was no effect of the feedback timing on the performance gains.

Rimac researchers conducted an experiment to compare two versions of the system. The first version was defined by using only direct lines of reasoning and remediation dialogs, whereas the second version also incorporated some decision rules that were triggered by different types of situations that can arise during the dialog. The results revealed that the second version outperformed the first version.

b: TUTORING FEEDBACK SUPPORTED BY DIALOG VS LIMITED TUTORING FEEDBACK

The learning effectiveness of Atlas-Andes was evaluated in a small comparative evaluation with Andes (without a dialog capacity). This comparison showed a significant effect in favor of the tutorial dialog of Atlas-Andes.

In an evaluation of Beetle II, researchers compared two versions of Beetle, TELL (tell the right answer without analyzing the student's response) and ELICIT (guide toward the right answer over several dialog turns). It was found that the TELL version was as effective as the ELICIT version without a significant difference between them. Previously, in [102], researchers had achieved similar results.

The Beetle II researchers in [103] studied the effect of different types of interpretation errors in learning gain using two tutoring policies similar to TELL and ELICIT. The results indicated that most of the interpretation problems are not significantly correlated with the learning gain. However, errors related to the misuse of domain terminology appeared to be particularly significant.

In another study with Beetle II [104], researchers compared three conditions: human-human tutoring and two human-computer tutoring conditions similar to TELL and ELICIT. The results indicated that even though the students produced the same percentage of content talk (statements including domain concepts that pertain to the lesson) under the two human-computer conditions, the proportion of content talk was correlated only with learning gain in the condition similar to ELICIT.

An evaluation of Geometry Explanation Tutor showed that students did not learn more using explanations than using menu options (to select the name of a geometry definition or theorem that justifies a problem-solving step). However, the students who used explanations performed better at stating explanations in the post-tests. Additionally, the results indicated that good quality feedback (rated by two human experts) correlates with students' progress through the dialogues (rated by two human experts) and with learning. This finding suggests that students do utilize the system's feedback and be able to extract the information they need to improve their explanations.

For Kermit-SE, researchers conducted an experiment to compare the learning gains under two different conditions, Kermit (previous version without dialog capacity) with only detailed hints (on how to correct a mistake) and Kermit-SE (with self-explanation support). After working with Kermit-SE, students who used Kermit-SE were divided into two groups, self-explainers and non-self-explainers (did not use the self-explanation support). They found that the students who used Kermit improved more than the students who used Kermit-SE; and the difference in the learning gain between the self-explainers and non-self-explainers was not statistically significant.

An experiment with Ms. Lindquist compared its regular tutoring feedback with limited tutoring feedback in which the system tells students the right answers as soon as they commit a mistake. The results of this experiment revealed that the tutoring supported by the dialog provided a greater learning gain than the limited approach.

Normit-SE researchers conducted a study and showed that despite not having a significant difference, the students who utilized self-explanation (using the Normit-SE) learned to handle the restrictions faster than those who were not asked to explain themselves (using its basic version).

The effectiveness of SCoT as a learning tool was measured in a study. The results showed that the tutorial dialogs with SCoT were more effective than a mere simulation without tutoring.

c: TUTORING FEEDBACK SUPPORTED BY DIALOG VS LIMITED TUTORING FEEDBACK VS RANDOM FEEDBACK

AutoTutor Lite researchers conducted a study to compare a new feedback system (based on learner's characteristic curves) with two other different feedback generators: random feedback and no feedback. They found that the feedback supported by the student model based on the characteristic curves led to greater learning of the participants than the other two generators.

2) ITS VERSUS TRADITIONAL LEARNING METHOD

In this group, we included the experiments aimed at comparing the learning gain provided by the ITS with the one provided by a traditional learning method such as textbooks, (expert or novice) human tutoring or a lesson/tutorial on a different topic to the one to be learnt, etc. To conduct these comparisons, these experiments applied A/B testing and pre/posttest.

A/B testing is the process of comparing two versions of a product/tool/method by testing the subject's response to version A against version B and determining which of the two versions is more effective. Typically, two different groups of users utilize the two versions, respectively, and then the users' responses are compared.

In the experiments mentioned in this section involving A/B testing, researchers compared the learning gains of two groups, the experimental group and the control group. While the experimental group used the evaluated ITS (version A), the control group used a traditional learning method (version B). In some of the experiments, researchers employed more than one control group.

On the other hand, there are four experiments, as can be seen in Table 8, in which the experiments only assessed the learning gain of an experimental group by using only pre/posttest. In such cases, we can roughly assume that the traditional leaning method was "nothing".

Table 8 shows the classification of the evaluations with regard to the addressed learning retention into two learning types: short-term learning and long-term learning. For each study, Table 8 specifies which type of evaluation was conducted (pre/posttest and/or A/B test), the control group, and a highlighted result. For most of the ITSs, we describe the learning gain by using the standardized effect size calculated as the difference in the posttest means for the experimental and control groups, divided by the within-group population standard deviation (0.2 small, 0.5 medium, 0.8 large) [105]. For reference, students working one-on-one with expert human tutors often score 2.0 higher than students working on the same topic in classrooms [106].

3) STUDENT IMPRESSIONS

This group contains some user surveys in which researchers employed questionnaires to collect the students' impressions on the ITSs. Table 9 shows the evaluated ITSs together with the most highlighted evaluated aspects and the most relevant results obtained in each study. Moreover, in the field "Related experiments", we indicate which experiment of the two previous subsections (E.1 or E.2) was performed prior to the student survey and presented in the same paper.

4) COMPONENT EVALUATION

This group encompasses the validations of certain properties of ITSs components. Table 10 shows the referenced studies grouped into three types with regard to the type of evaluated component.

5) POSITIVE ARGUMENTATION

In this group, we included Paco, which was evaluated through a qualitative study with seven users. After being trained with Paco, they were interviewed about their impressions. Most of them praised the Paco's overall teaching skills and were able to complete the entire task without major errors.

6) NOT REPORTED

We could not find any reported evaluation for Lana or Jacob. However, in the case of Lana, the authors presented an experimental methodology in which they indicated how they would conduct the evaluation and the hypotheses they would want to prove.

V. DISCUSSION

According to the results shown above, we can state that most ITSs belong to the EMT type and are aimed at teaching topics in STEM domains at the university level.

TABLE 9. Evaluation of student impressions of ITSs.

ITS	Related experiment	Highlighted evaluated aspects	Highlighted results
Abdullah	Section E.2	Does Tutor Abdullah overload you with information?	The majority (58.6%) of the students had a neutral feeling, 25.9% of the students were happy, and 15.5% of the students were not happy.
Beetle II	Section E.1	The tutoring quality	TELL scored higher than ELICIT in tutor satisfaction questions.
CIRCSIM-Tutor (1)	Section E.2	Satisfaction with the use	The results of the survey were quite positive. Students found the program easy to use; they felt that they learned a lot; and they would recommend the program to other students.
Kermit-SE	Section E.1	Time to learn from the interface, the amount learned, and the enjoyment	There was no significant difference between the time needed by the two groups to learn the interface. The difference in the mean responses on the amount learned and the enjoyment were not significant.
Oscar	Section E.1	Satisfaction with the use	They welcomed the use of Oscar, with 94% and 90% agreeing that Oscar helped them to revise. Here, 92% of the participants would use Oscar if it were available, with 78% saying they would prefer Oscar over learning from a book. Unexpectedly, 46% of the sample said they would use Oscar rather than attending a face-to-face tutorial.
ProPL	Section E.2	Satisfaction with the use	Most students indicated a desire to use the system again in the future.
ReportTutor	Section E.1	Satisfaction with the two interfaces (immediate versus delayed)	Although the scores did not reveal a statistical difference in the overall preference, the residents felt that the delayed interface was easier to use and more flexible than the other. First year residents had a significantly higher total survey attitude score toward the immediate interface than other residents.
VCAEST	Section E.2	Do you feel a sense of efficacy/adequacy/self-confidence regarding the simulation?	The participants in the live action training condition felt that their live action training was effective and adequate (94.73%), and those who received virtual training felt that their training was not effective, where 56.25% of the participants responded with either “somewhat” or “not at all”.

TABLE 10. Evaluation of the components of ITSs.

ITS	Evaluated component	Highlighted result
Atlas-Andes	NLU	The study found that the NLU component was able to correctly classified student responses into one of the authored categories 96.3% of the time with a precision of 99.5% and a recall of 94.4%.
AutoTutor	NLU	The study in [112] presented an evaluation that concerns LSA as a technique to interpret the statements of the student. This evaluation served to validate its efficiency to track the coverage of an expectation during the dialogue and obtained excellent results.
My Science Tutor	NLU	The results showed that using a global language model, the baseline system had a word error rate (WER) of 30.9% with an overall recall of 84% and precision of 89% for the extraction of frame elements.
My Science Tutor	Voice recognizer	The results indicated a WER of 27.4% with a recall of 86% and a precision of 90% after adapting the unsupervised speaker in batches for the ASR system.
SCoT	Voice recognizer	The evaluation concluded that voice recognizer it is mature enough to be used in tutorial dialog systems.
ITSpoke	Emotion recognizer	The study investigated the possibility of recognizing the emotional state of the students (negative, positive and neutral) in two conditions, with a human tutor and a computer tutor. The results showed significant improvements in the prediction precision over the relevant baselines.

Additionally, ITSs mostly enable a dialog aimed at supporting problem solving (the apply level of Bloom’s taxonomy). Another significant group of ITSs is aimed at facilitating the building of connections between new knowledge and prior knowledge in the students’ mind (the understand level in Bloom’s taxonomy) by supporting the generation of explanations.

Concerning the instructional approach, most of the ITSs help the student to actively elaborate explanations and justifications of a previous student input. This previous input could be a previous student’s answer, an erroneous prediction related to a simulation (e.g., Beetle II), a wrong action in problem-solving (e.g., SCoT), etc. Moreover, there are ITSs, such as SCoT and Rimac, whose tutorial dialog does not begin until students have executed a task completely or have provided a (right or wrong) solution.

AutoTutor and its descendants address students’ explanations through a wide variety of pedagogical strategies. Nevertheless, all of them except for DeepTutor rely on the same statistical technique for NLU called LSA. Over time, to improve its effectiveness, LSA has been improved and complemented with other techniques in different AutoTutor versions, as explained above.

On the other hand, the majority of ITSs included in this SLR use a symbolic approach based on lexicons, grammars and pattern matching. Although this approach was used in the design of many of the early ITSs, over the past decade (2010-2020), symbolic and statistics approaches have been adopted at a similar frequency and, in some cases, in the same system. The structured semantic representations that symbolic systems produce offer advantages for integrating tutoring with simulation-based environments or with environments in which problems are dynamically generated. In contrast, the statistics approach tends to be more robust than the symbolic approach against unexpected student inputs or linguistic errors. Furthermore, the symbolic approach requires more upfront investment in parsing and interpretation infrastructure to be developed and deployed in new domains [56].

Thus, some researchers sought to leverage the strengths of each approach by developing hybrid systems. As mentioned below, we believe that this approach will continue to be a future trend.

We found empirical evaluations for 90.91% of the considered ITSs that measure learning gains and/or assess the impacts of different tutoring strategies. Here, it is worth noting that some of the early ITSs have not had as many rigorous evaluations as those of the ITSs developed in the past decade. In most of these studies, the ITSs outperformed the control conditions (static contents, classroom teaching, etc.) or provided learning gains in pre- and posttest evaluations. Additionally, some of these studies included surveys on student impressions, which in most of the cases revealed a high satisfaction with the ITS.

VI. THREATS TO VALIDITY

This section describes concerns that must be improved in future replications of this study and other aspects that must be accounted for to extend the results of the SLR performed in this work. In this section, according to the classification defined in [113], we will address the following threats to validity: construct, internal, external, and conclusion.

The main constructs in this review are the two concepts of the intelligent tutoring system and dialog system. However, different closely related keywords are used to refer to these concepts in the literature. Therefore, to mitigate this threat, the search query was appropriately defined to include these different keywords. Additionally, we performed snowballing to identify missing studies.

As threats to the internal validity, some subjective decisions could have occurred during the paper selection and data extraction. For example, as classifying some of the ITSs into the EMT category, because many of the primary studies do not explicitly indicate that the systems belong to this category. To alleviate this threat, we performed the selection process in an iterative way, and we reviewed and discussed carefully the data extraction with respect to the categories considered in this work. In addition, before performing the data extraction, we asked an external expert to validate the coverage of the research questions.

External validity is concerned with establishing the generalizability of the SLR results, which is related to the degree to which the primary studies are representative for the review topic. To ease this threat, we grounded the study on the most well-known digital libraries in computer science.

With regard to the conclusion validity, it is possible that some excluded studies in this review should have been included.

VII. FUTURE TRENDS

The findings of this SLR allow us to foresee three future trends in the research and development of ITSs with natural language dialogue.

First, we expect there to be work on more reusable solutions that take advantage of already existing ITSs as building blocks. An example of this is the work being done by the Office of Naval Research through two projects: 1) SKOPE-IT [114], a tutor for the mathematics domain that teaches how to solve algebra problems by integrating AutoTutor technologies and ALEKS ITS, and 2) ElectronixTutor [115], a tutor for the electronics domain that teaches navigators to learn about electronic circuits integrating as components using AutoTutor, Dragoon, LearnForm, ASSISSTments and Beetle II.

Second, we expect there to be improvements in natural language processing by means of hybrid systems. Thus far, some effort has been conducted to optimize the NLU techniques. On the one hand, the AutoTutor family has opted to improve and obtain the best performance out of its LSA statistical technique by incorporating enhancements based on symbolic methods. On the other hand, other ITSs (e.g., Why2-Atlas) have adopted hybrid approaches from the beginning by combining symbolic with statistics approaches.

Finally, we expect there to be integration of ITSs with dialog in rich learning environments, for example, 3D interactive simulations in which the users can find (student or tutor) animated agents with which to interact. As a result of this type of interaction, nonverbal communication would gain relevance.

VIII. CONCLUSIONS AND FUTURE WORK

In this SLR, we have gained some valuable insights into the ITSs that provide tutoring in natural language developed over the past twenty years. Our objective was to know, apart from their essential and pedagogical characteristics, the purpose of the dialog that they provide, the NLU techniques that they rely on, and the results of empirical evaluations of the ITSs.

This SLR eventually selected 49 primary studies, in which 33 relevant ITSs could be identified. Thereupon, we analyzed these ITSs from the perspective of five research questions. As a result, we classified the ITSs into different categories and briefly described their most relevant characteristics with respect to each research question.

We think that this systematic review will be useful for the e-learning community since it gathers evidence of ITSs with dialog that have been implemented thus far and the techniques used to provide tutoring in natural language. These ITSs vary in the way that they simulate the mechanisms of human dialogue. However, all of them attempt to understand natural language, formulate adaptive responses and implement pedagogical strategies that help students to learn.

In future work, we will aim to explore the educational possibilities of the currently available platforms for developing dialog systems in the cloud. Furthermore, we plan to work on the development of an ITS with dialog for procedural training that is fully integrated with a 3D virtual learning environment (i.e., context-aware). Despite procedural training has a key role in many application domains, as shown in this systematic review, it has been poorly covered in the literature to date, and therefore, this development should serve to begin to bridge this gap.

REFERENCES

[1] D. Sleeman and J. S. Brown, *Intelligent Tutoring Systems*. New York, NY, USA: Academic, 1982.

[2] B. D. Boulay, ''Recent meta-reviews and meta–analyses of AIED systems,'' *Int. J. Artif. Intell. Educ.*, vol. 26, no. 1, pp. 536–537, Mar. 2016.

[3] J. A. Kulik and J. D. Fletcher, ''Effectiveness of intelligent tutoring systems: A meta-analytic review,'' *Rev. Educ. Res.*, vol. 86, no. 1, pp. 42–78, Mar. 2016.

[4] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu, ''Intelligent tutoring systems and learning outcomes: A meta-analysis.,'' *J. Educ. Psychol.*, vol. 106, no. 4, pp. 901–918, 2014.

[5] S. Steenbergen-Hu and H. Cooper, ''A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning,'' *J. Educ. Psychol.*, vol. 106, no. 2, pp. 331–347, 2014.

[6] K. VanLEHN, ''The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems,'' *Educ. Psychol.*, vol. 46, no. 4, pp. 197–221, Oct. 2011.

[7] V. Aleven, B. McLaren, I. Roll, and K. Koedinger, ''Toward tutoring help seeking applying cognitive modeling to meta-cognitive skills,'' in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2004, pp. 227–239.

[8] M. Prince, ''Does active learning work? A review of the research,'' *J. Eng. Educ.*, vol. 93, no. 3, pp. 223–231, Jul. 2004.

[9] V. J. Shute, ''Focus on formative feedback,'' *Rev. Educ. Res.*, vol. 78, no. 1, pp. 153–189, Mar. 2008.

[10] M. Chi, M. Roy, and R. Hausmann, ''Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning,'' *Cognit. Sci., Multidisciplinary J.*, vol. 32, no. 2, pp. 301–341, Mar. 2008.

[11] K. VanLehn, S. Siler, C. Murray, T. Yamauchi, and W. B. Baggett, ''Why do only some events cause learning during human tutoring?'' *Cognition Instruct.*, vol. 21, no. 3, pp. 209–249, Sep. 2003.

[12] B. D. Nye, A. C. Graesser, and X. Hu, ''AutoTutor and family: A review of 17 years of natural language tutoring,'' *Int. J. Artif. Intell. Educ.*, vol. 24, no. 4, pp. 427–469, Dec. 2014.

[13] A. Mitrovic, ''Fifteen years of constraint-based tutors: What we have achieved and where we are going,'' *User Model. User-Adapted Interact.*, vol. 22, nos. 1–2, pp. 39–72, Apr. 2012.

[14] M. A. Emran and K. Shaalan, ''A survey of intelligent language tutoring systems,'' in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2014, pp. 393–399.

[15] B. Kitchenham and S. Charters, ''Guidelines for performing systematic literature reviews in software engineering,'' Keele Univ., Keele, U.K., Tech. Rep. EBSE-2007-01, 2007.

[16] J. Carbonell, ''AI in CAI: An artificial-intelligence approach to computer-assisted instruction,'' *IEEE Trans. Man Mach. Syst.*, vol. 11, no. 4, pp. 190–202, Dec. 1970.

[17] E. Wenger, *Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge*. San Francisco, CA, USA: Morgan Kaufmann, 1987.

[18] A. S. Gertner and K. VanLehn, ''Andes: A coached problem solving environment for Physics,'' in *Proc. 5th Int. Conf. Intell. Tutoring Syst.*, 2000, pp. 133–142.

[19] S. Ritter, J. R. Anderson, K. R. Koedinger, and A. Corbett, ''Cognitive tutor: Applied research in mathematics education,'' *Psychonomic Bull. Rev.*, vol. 14, no. 2, pp. 249–255, Apr. 2007.

[20] A. Mitrovic and S. Ohlsson, ''Evaluation of a constraint-based tutor for a database language,'' *Int. J. Artif. Intell. Educ.*, vol. 10, nos. 3–4, pp. 238–256, 1999.

[21] E. Mousavinasab, N. Zarifsanaiey, S. R. Niakan Kalhori, M. Rakhshan, L. Keikha, and M. Ghazi Saeedi, ''Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods,'' *Interact. Learn. Environ.*, pp. 1–22, Dec. 2018, doi: 10.1080/10494820.2018.1558257.

[22] B. D. Nye, ''Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context,'' *Int. J. Artif. Intell. Educ.*, vol. 25, no. 2, pp. 177–203, Jun. 2015.

[23] B. D. Nye, ''Barriers to ITS adoption: A systematic mapping study,'' in *Intelligent Tutoring Systems*. Cham, Switzerland: Springer, 2014, pp. 583–590.

[24] M. O. Dzikovska *et al.*, ''Intelligent tutoring with natural language support in the Beetle II system,'' in *Sustaining TEL: From Innovation to Learning and Practice*. Berlin, Germany: Springer, 2010, pp. 620–625.

[25] M. Chi, ''Eliciting self-explanations improves understanding,'' *Cognit. Sci.*, vol. 18, no. 3, pp. 439–477, Sep. 1994.

[26] A. Weerasinghe and A. Mitrovic, ''Enhancing learning through self-explanation,'' in *Proc. Int. Conf. Comput. Educ.*, 2002, pp. 244–248.

[27] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark, ''Intelligent tutoring goes to school in the big city,'' *Int. J. Artif. Intell. Educ.*, vol. 8, no. 1, pp. 30–43, 1997.

[28] A. Mitrovic, B. Martin, and P. Suraweera, ''Intelligent tutors for all: Constraint-based modeling methodology, systems and authoring,'' *IEEE Intell. Syst.*, vol. 22, no. 4, pp. 38–45, 2007.

[29] C. Conati, A. Gertner, and K. VanLehn, ''Using Bayesian networks to manage uncertainty in student modeling,'' *User Model. User-Adapt. Interact.*, vol. 12, no. 4, pp. 371–417, Nov. 2002.

[30] A. Graesser, A. Olney, M. Ventura, and G. T. Jackson, ''AutoTutor's coverage of expectations during tutorial dialogue,'' in *Proc. FLAIRS Conf.*, 2005, pp. 518–523.

[31] J. Hirschberg and C. D. Manning, ''Advances in natural language processing,'' *Science*, vol. 349, no. 6245, pp. 261–266, Jul. 2015,

[32] P. W. Jordan, M. Makatchev, and K. VanLehn, ''Combining competing language understanding approaches in an intelligent tutoring system,'' in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2004, pp. 346–357.

[33] J. Klavans and P. Resnik, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, MA, USA: MIT Press. 1996.

[34] S. D'Mello and A. Graesser, ''Design of dialog-based intelligent tutoring systems to simulate human-to-human tutoring,'' in *Where Humans Meet Machines: Innovative Solutions for Knotty Natural-Language Problems*, A. Neustein and J. A. Markowitz, Eds. New York, NY, USA: Springer, 2013, pp. 233–269.

[35] M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann, ''Learning from human tutoring,'' *Cognit. Sci.*, vol. 25, no. 4, pp. 471–533, Jul. 2001.

[36] J. O'Shea, Z. Bandar, and K. Crockett, ''Systems engineering and conversational agents,'' in *Intelligence-Based Systems Engineering*, A. Tolk and L. C. Jain, Eds. Berlin, Germany: Springer, 2011, pp. 201–232.

[37] M. Owda, Z. Bandar, and K. Crockett, ''Information extraction for SQL query generation in the conversation-based interfaces to relational databases (C-BIRD),'' in *Agent and Multi-Agent Systems: Technologies and Applications*. Berlin, Germany: Springer, 2011, pp. 44–53.

[38] J. Cassell, ''Embodied conversational agents: Representation and intelligence in user interfaces,'' *AI Mag.*, vol. 22, no. 4, pp. 67–83, 2001.

[39] A. C. Graesser, ''Conversations with AutoTutor help students learn,'' *Int. J. Artif. Intell. Educ.*, vol. 26, no. 1, pp. 124–132, Mar. 2016.

[40] A. C. Graesser, S. Lu, G. T. Jackson, H. H. Mitchell, M. Ventura, A. Olney, and M. M. Louwerse, ''AutoTutor: A tutor with dialogue in natural language,'' *Behav. Res. Methods, Instrum., Comput.*, vol. 36, no. 2, pp. 180–192, May 2004.

[41] A. C. Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, and R. Kreuz, ''AutoTutor: A simulation of a human tutor,'' *Cognit. Syst. Res.*, vol. 1, no. 1, pp. 35–51, Dec. 1999.

[42] K. Millis, C. Forsyth, H. Butler, P. Wallace, A. Graesser, and D. Halpern, ''Operation ARIES!: A serious game for teaching scientific inquiry,'' in *Serious Games and Edutainment Applications*, M. Ma, A. Oikonomou, and L. C. Jain, Eds. London, U.K.: Springer, 2011, pp. 169–195.

[43] G. T. Jackson, A. Olney, A. C. Graesser, and H.-J. J. Kim, ''AutoTutor 3-D simulations: Analyzing users' actions and learning trends,'' in *Proc. Annu. Meeting Cognit. Sci. Soc.*, vol. 28, 2006, pp. 1557–1562.

[44] S. D'mello and A. Graesser, ''AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back,'' *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 4, pp. 1–39, Dec. 2012.

[45] J. Sullins, S. Craig, and X. Hu, ''Exploring the effectiveness of a novel feedback mechanism within an intelligent tutoring system,'' *Int. J. Learn. Technol.*, vol. 10, no. 3, pp. 220–236, 2015.

[46] V. Rus, S. D'Mello, X. Hu, and A. C. Graesser, "Recent advances in conversational intelligent tutoring systems," *AI Mag.*, vol. 34, no. 3, pp. 42–54, 2013.

[47] S. D'Mello, A. Olney, C. Williams, and P. Hays, "Gaze tutor: A gaze-reactive intelligent tutoring system," *Int. J. Hum.-Comput. Stud.*, vol. 70, no. 5, pp. 377–398, May 2012.

[48] A. M. Olney, A. C. Graesser, and N. K. Person, "Tutorial dialog in natural language," in *Advances in Intelligent Tutoring Systems*, R. Nkambou, J. Bourdeau, and R. Mizoguchi, Eds. Berlin, Germany: Springer, 2010, pp. 181–206.

[49] A. C. Graesser, G. T. Jackson, E. C. Matthews, H. H. Mitchell, A. Olney, M. Ventura, P. Chipman, D. Franceschetti, X. Hu, M. M. Louwerse, and N. K. Person, "Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog," in *Proc. 25th Annu. Conf. Cognit. Sci. Soc.*, 2003, pp. 1–6.

[50] D. M. Morrison, B. Nye, and X. Hu, "Where in the data stream are we?: Analyzing the flow of text in dialogue-based systems for learning," in *Design Recommendations for ITS: Instructional Management*, vol. 2, R. A. Sottilare, A. C. Graesser, X. Hu, and B. S. Goldberg, Eds. New York, NY, USA: Instructional Management, 2014, pp. 237–247.

[51] B. D. Nye, "Integrating GIFT and autotutor with sharable knowledge objects (SKO)," in *Proc. Artif. Intell. Educ. (AIED) Workshop Generalized Intell. Framework Tutoring (GIFT)*, 2013, pp. 54–61.

[52] C. R. Wolfe, V. F. Reyna, C. L. Widmer, E. M. Cedillos-Whynott, P. G. Brust-Renck, A. M. Weil, and X. Hu, "Understanding genetic breast cancer risk: Processing loci of the BRCA gist intelligent tutoring system," *Learn. Individual Differences*, vol. 49, pp. 178–189, Jul. 2016.

[53] K. T. Shubeck, S. D. Craig, and X. Hu, "Live-action mass-casualty training and virtual world training: A comparison," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2016, pp. 2103–2107.

[54] O. G. Alobaidi, K. Crockett, J. D. O'Shea, and T. M. Jarad, "The application of learning theories into abdullah: An intelligent arabic conversational agent tutor," in *Proc. Int. Conf. Agents Artif. Intell.*, 2015, pp. 361–369.

[55] S. S. Aljameel, J. D. O'Shea, K. A. Crockett, A. Latham, and M. Kaleem, "Development of an arabic conversational intelligent tutoring system for education of children with ASD," in *Proc. IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. (CIVEMSA)*, Jun. 2017, pp. 24–29.

[56] M. Dzikovska, N. Steinhauser, E. Farrow, J. Moore, and G. Campbell, "BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics," *Int. J. Artif. Intell. Educ.*, vol. 24, no. 3, pp. 284–332, Sep. 2014.

[57] C. W. Woo, M. W. Evens, R. Freedman, M. Glass, L. S. Shim, Y. Zhang, Y. Zhou, and J. Michael, "An intelligent tutoring system that generates a natural language dialogue using dynamic multi-level planning," *Artif. Intell. Med.*, vol. 38, no. 1, pp. 25–46, Sep. 2006.

[58] C. Benzmüller, H. Horacek, I. Kruijff-Korbayová, M. Pinkal, J. Siekmann, and M. Wolska, "Natural language dialog with a tutor system for mathematical proofs," in *Cognitive Systems*. Berlin, Germany: Springer, 2007, pp. 1–14.

[59] V. Aleven, A. Ogan, O. Popescu, C. Torrey, and K. Koedinger, "Evaluating the effectiveness of a tutorial dialogue system for self-explanation," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2004, pp. 443–454.

[60] D. J. Litman and K. Forbes-Riley, "Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors," *Speech Commun.*, vol. 48, no. 5, pp. 559–590, May 2006.

[61] W. Ward, R. Cole, D. Bolaños, C. Buchenroth-Martin, E. Svirsky, S. V. Vuuren, T. Weston, J. Zheng, and L. Becker, "My science tutor: A conversational multimedia virtual tutor for elementary school science," *ACM Trans. Speech Lang. Process.*, vol. 7, no. 4, pp. 18:1–18:29, 2011.

[62] A. Latham, K. Crockett, and D. McLean, "An adaptation algorithm for an intelligent natural language tutoring system," *Comput. Educ.*, vol. 71, pp. 97–110, Feb. 2014.

[63] H. C. Lane and K. VanLehn, "Teaching the tacit knowledge of programming to noviceswith natural language tutoring," *Comput. Sci. Educ.*, vol. 15, no. 3, pp. 183–201, Sep. 2005.

[64] G. M. El Saadawi, E. Tseytlin, E. Legowski, D. Jukic, M. Castine, J. Fine, R. Gormley, and R. S. Crowley, "A natural language intelligent tutoring system for training pathologists: Implementation and evaluation," *Adv. Health Sci. Educ.*, vol. 13, no. 5, pp. 709–722, Dec. 2008.

[65] P. Albacete, P. Jordan, and S. Katz, "Is a dialogue-based tutoring system that emulates helpful co-constructed relations during human tutoring effective?" in *Artificial Intelligence in Education*. Cham, Switzerland: Springer, 2015, pp. 3–12.

[66] E. Arnott, P. Hastings, and D. Allbritton, "Research methods tutor: Evaluation of a dialogue-based tutoring system in the classroom," *Behav. Res. Methods*, vol. 40, no. 3, pp. 694–698, Aug. 2008.

[67] K. VanLehn, P. W. Jordan, C. P. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, and S. Siler, "The architecture of Why2-Atlas: A coach for qualitative physics essay writing," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2002, pp. 158–167.

[68] P. W. Jordan, B. Hall, M. Ringenberg, Y. Cue, and C. Rosé, "Tools for authoring a dialogue agent that participates in learning studies," in *Proc. Conf. Artif. Intell. Educ., Building Techn. Rich Learn. Contexts Work*, 2007, pp. 43–50.

[69] R. Freedman, C. P. Rosé, M. A. Ringenberg, and K. VanLehn, "ITS tools for natural language dialogue: A domain-independent parser and planner," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2000, pp. 433–442.

[70] C. Rosé, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein, "Interactive conceptual tutoring in atlas-andes," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2001, pp. 256–266.

[71] M. Evers and A. Nijholt, "Jacob—An animated instruction agent in virtual reality," in *Proc. Adv. Multimodal Interfaces (ICMI)*, 2000, pp. 526–533.

[72] J. Rickel, N. Lesh, C. Rich, C. Sidner, and A. Gertner, "Collaborative discourse theory as a foundation for tutorial dialogue," in *Proc. 6th Int. Conf. Intell. Tutoring Syst.*, 2002, pp. 542–551.

[73] N. T. Heffernan, K. R. Koedinger, and L. Razzaq, "Expanding the model-tracing architecture: A 3rd generation intelligent tutor for algebra symbolization," *Int. J. Artif. Intell. Educ.*, vol. 18, no. 2, pp. 153–178, 2008.

[74] A. Mitrovic, "The effect of explaining on learning: A case study with a data normalization tutor," in *Proc. Int. Conf. Artif. Intell. Educ. (AIED)*, 2005, pp. 499–506.

[75] A. Weerasinghe and A. Mitrovic, "Facilitating deep learning through self-explanation in an open-ended domain," *Int. J. Knowl.-Based Intell. Eng. Syst.*, vol. 10, no. 1, pp. 3–19, Feb. 2006.

[76] A. Weerasinghe and A. Mitrovic, "Individualizing self-explanation support for ill-defined tasks in constraint-based tutors," in *Proc. Workshop ITS Ill-Defined Domains*, 2006, pp. 56–64.

[77] H. Pon-Barry, B. Clark, E. O. Bratt, K. Schultz, and S. Peters, "Evaluating the effectiveness of SCoT: A spoken conversational tutor," in *Proc. ITS Workshop Dialogue-Based Intell.*, 2004, pp. 23–32.

[78] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objetives*. London, U.K.: Longman, 2001.

[79] V. V. Bulitko and D. C. Wilkins, "Automated instructor assistant for ship damage control," in *Proc. Amer. Assoc. Artif. Intell.*, 1999, pp. 778–785.

[80] S. Katz, G. O'Donnel, and H. Kay, "An approach to analyzing the role and structure of reflective dialogue," *Int. J. Artif. Intell. Educ.*, vol. 11, no. 3, pp. 320–343, 2000.

[81] A. C. Graesser and N. K. Person, "Question asking during tutoring," *Amer. Educ. Res. J.*, vol. 31, no. 1, pp. 104–137, Mar. 1994.

[82] N. K. Person, A. C. Graesser, D. Harter, E. Mathews, and T. R. Group, "Dialog move generation and conversation management in AutoTutor," in *Proc. AAAI Fall Symp. Building Dialogue Syst. Tutorial Appl.*, 2000, pp. 87–94.

[83] M. Glass, "Processing language input in the CIRCSIM-Tutor intelligent tutoring system," in *Proc. Artif. Intell. Educ.*, 2001, pp. 210–221.

[84] C. P. Rosé, "A framework for robust semantic interpretation," in *Proc. 1st North Amer. Chapter Assoc. Comput. Linguistics Conf.*, 2000, pp. 311–318.

[85] J. Allen, M. Manshadi, M. Dzikovska, and M. Swift, "Deep linguistic processing for spoken dialogue systems," in *Proc. Workshop Deep Linguistic Process. (DeepLP)*, 2007, pp. 49–56.

[86] M. O. Dzikovska, E. Farrow, and J. D. Moore, "Combining semantic interpretation and statistical classification for improved explanation processing in a tutorial dialogue system," in *Proc. Int. Conf. Artif. Intell. Educ. (AIED)*, 2013, pp. 279–288.

[87] J. Baldridge and G.-J.-M. Kruijff, "Multi-modal combinatory categorial grammar," in *Proc. 10th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2003, pp. 211–218.

[88] W. Ward and B. Pellon, "The CU communicator system," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop*, 1999, pp. 1–4.

[89] C. Rich, C. L. Sidner, and N. Lesh, "Collagen: Applying collaborative discourse theory to human-computer interaction," *AI Mag.*, vol. 22, no. 4, p. 15, 2001.

[90] J. Dowding, J. M. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran, "Gemini: A natural language system for spoken-language understanding," in *Proc. Workshop Human Lang. Technol. (HLT)*, 1993, pp. 54–61.

[91] A. Graesser, P. Penumatsa, M. Ventura, Z. Cai, and X. Hu, "Using LSA in AutoTutor: Learning through mixed-initiative dialogue in natural language," in *Handbook of Latent Semantic Analysis*, T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, Eds. Abingdon, U.K.: Routledge, 2007, pp. 243–262.

[92] Z. Cai, A. C. Graesser, C. Forsyth, C. Burkett, K. Millis, P. Wallace, D. Halpern, and H. Butler, "Trialog in ARIES: User input assessment in an intelligent tutoring system," in *Proc. 3rd IEEE Int. Conf. Comput. Intell. Syst.*, 2011, pp. 429–433.

[93] V. Rus, R. Banjade, M. Lintean, N. Niraula, and D. Stefanescu, "SEMILAR: A semantic similarity toolkit for assessing students' natural language inputs," in *Proc. Educ. Data Mining*, 2013, pp. 402–403.

[94] A. M. Olney, N. K. Person, and A. C. Graesser, "Guru: Designing a conversational expert intelligent tutoring system," in *Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches*, C. Boonthum-Denecke, P. M. McCarthy, and T. A. Lamkin, Eds. Hershey, PA, USA: IGI Global, 2012, pp. 156–171.

[95] M. Makatchev, P. W. Jordan, and K. VanLehn, "Abductive theorem proving for analyzing student explanations to guide feedback in intelligent tutoring systems," *J. Automated Reasoning*, vol. 32, no. 3, pp. 187–226, Feb. 2004.

[96] P. Jordan, M. Makatchev, U. Pappuswamy, K. VanLehn, and P. Albacete, "A natural language tutorial dialogue system for physics," in *Proc. 19th Int. FLAIRS Conf.*, 2006, pp. 521–526.

[97] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *Proc. AAAI/ICML-98 Workshop Learn. Text Categorization*, 1998, pp. 41–48.

[98] D. Lin and P. Pantel, "Discovery of inference rules for question-answering," *Natural Lang. Eng.*, vol. 7, no. 4, pp. 343–360, Dec. 2001.

[99] M. Makatchev and K. VanLehn, "Analyzing completeness and correctness of utterances using an ATMS," in *Proc. Conf. Artif. Intell. Educ., Supporting Learn. Through Intell. Socially Informed Technol.*, 2005, pp. 403–410.

[100] V. Rus, D. Stefanescu, W. Baggett, N. Niraula, D. Franceschetti, and A. C. Graesser, "Macro-adaptation in conversational intelligent tutoring matters," in *Proc. Intell. Tutoring Syst.*, 2014, pp. 242–247.

[101] A. Weerasinghe, A. Mitrovic, M. Van Zijl, and B. Martin, "Evaluating the effectiveness of adaptive tutorial dialogues in EER-tutor," in *Proc. 18th Interfaces Conf. Comput. Educ.*, 2010, pp. 33–40.

[102] M. O. Dzikovska, J. D. Moore, N. Steinhauser, G. Campbell, E. Farrow, and C. Callaway, "BEETLE II: A system for tutoring and computational linguistics experimentation," in *Proc. ACL Syst. Demostrations*, 2010, pp. 13–18.

[103] M. O. Dzikovska, J. D. Moore, N. Steinhauser, and G. Campbell, "The impact of interpretation problems on tutorial dialogue," in *Proc. ACL Conf. Short Papers*, 2010, pp. 43–48.

[104] M. O. Dzikovska, N. B. Steinhauser, J. D. Moore, G. E. Campbell, K. M. Harrison, and L. S. Taylor, "Content, social, and metacognitive statements: An empirical study comparing human-human and human-computer tutorial dialogue," in *Sustaining TEL: From Innovation to Learning and Practice*. Berlin, Germany: Springer, 2010, pp. 93–108.

[105] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York, NY, USA: Academic, 1988.

[106] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as One-to-One tutoring," *Educ. Researcher*, vol. 13, no. 6, pp. 4–16, Jun. 1984.

[107] A. C. Graesser, K. N. Moreno, J. C. Marineau, A. B. Adcock, A. M. Olney, and N. K. Person, "AutoTutor improves deep learning of Computer Literacy: Is it the dialog or the talking head?" in *Proc. Artif. Intell. Educ.*, 2003, pp. 47–54.

[108] B. Mills, M. Evens, and R. Freedman, "Implementing directed lines of reasoning in an intelligent tutoring system using the atlas planning environment," in *Proc. Int. Conf. Inf. Technol., Coding Comput. (ITCC)*, 2004, pp. 729–733.

[109] A. M. Olney, S. D'Mello, N. Person, W. Cade, P. Hays, C. Williams, B. Lehman, and A. Graesser "Guru: A computer tutor that models expert human tutors," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2012, pp. 256–261.

[110] G. T. Jackson, M. Ventura, P. Chewle, and A. Graesser, "The impact of Why/AutoTutor on learning and retention of conceptual Physics," in *Proc. Intell. Tutoring Syst.*, 2004, pp. 501–510.

[111] K. VanLehn, A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney, and C. P. Rosé, "When are tutorial dialogues more effective than reading?" *Cognit. Sci.*, vol. 31, no. 1, pp. 3–62, Feb. 2007.

[112] A. C. Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, D. Harter, Tutoring Research Group, and N. Person, "Using latent semantic analysis to evaluate the contributions of students in AutoTutor," *Interact. Learn. Environ.*, vol. 8, no. 2, pp. 129–147, Aug. 2000.

[113] C. Wohlin, P. Runeson, M. Host, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Berlin, Germany: Springer-Verlag, 2012.

[114] B. D. Nye, P. I. Pavlik, A. Windsor, A. M. Olney, M. Hajeer, and X. Hu, "SKOPE-IT (shareable knowledge objects as portable intelligent tutors): Overlaying natural language tutoring on an adaptive learning system for mathematics," *Int. J. STEM Educ.*, vol. 5, no. 1, p. 12, Dec. 2018.

[115] A. C. Graesser *et al.*, "ElectronixTutor: An intelligent tutoring system with multiple learning resources for electronics," *Int. J. STEM Educ.*, vol. 5, no. 1, p. 15, Dec. 2018.

JOSÉ PALADINES received the M.Sc. degree in educational research from UNL, Ecuador, in 2006, and the M.Sc. degree in computer science from UNIANDES, Ecuador, in 2010. He is currently pursuing the Ph.D. degree in computer science with UPM. His main research interests include computer science education, intelligent tutoring systems, and natural language processing.

JAIME RAMÍREZ received the M.Sc. degree in computer science and the Ph.D. degree in computer science from UPM, in 1996 and 2002, respectively. He is currently an Associate Professor with the Computer Science School, UPM. His main research interests include ontology engineering and adaptive systems with a special focus on intelligent tutoring systems combined with 3D virtual environments.