Doctoral Thesis

Intelligent Corrective Feedback for Communicative Computer-Assisted Language Learning

Sungjin Lee (이 성 진)

Division of Electrical and Computer Engineering
(Computer Science and Engineering)
Pohang University of Science and Technology

2011

외국어 회화 교육을 위한 지능적인 교정적 피드백 방법

Intelligent Corrective Feedback for Communicative Computer-Assisted Language Learning

Intelligent Corrective Feedback for Communicative Computer-Assisted Language Learning

by

Sungjin Lee
Division of Electrical and Computer Engineering
(Computer Science and Engineering)
Pohang University of Science and Technology

A dissertation submitted to the faculty of the Pohang University of Science and Technology in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Division of Electrical and Computer Engineering (Computer Science and Engineering).

Pohang, Korea
May 31, 2011
Approved by

Prof. Gary Geunbae Lee

Academic Advisor

Intelligent Corrective Feedback for Communicative Computer-Assisted Language Learning

Sungjin Lee

The undersigned have examined this dissertation and hereby certify that it is worthy of acceptance for a doctoral degree from POSTECH.

May 31, 2011

Committee	Chair	Gary Geunbae Lee	(Seal)
	Member	Jonghyeok Lee	(Seal)
	Member	Seungjin Choi	(Seal)
	Member	Seungwon Hwang	(Seal)
	Member	Minhwa Chung	(Seal)

DECE 20075572 Sungjin Lee, 이 성 진, Intelligent Corrective Feedback for Communicative Computer-Assisted Language Learning. 외국어 회화 교육을 위한 지능적인 교정적 피드백 방법, Division of Electrical and Computer Engineering (Computer Science and Engineering), 2011, 165P, Advisors: Gary Geunbae Lee. Text in English.

ABSTRACT

Although there have been enormous investments into English education all around the world, not many differences have been made to change the English instruction style. Considering the shortcomings for the current teaching-learning methodology, we have been investigating Dialog-Based Computer-Assisted Language Learning (DB-CALL) systems. This thesis contains a set of our approaches to DB-CALL including theories, technologies, systems, and field studies. On top of the state-of-the-art technologies of spoken dialog system, a variety of adaptations have been applied to overcome some problems caused by numerous errors and variations naturally produced by non-native speakers so that learners can engage in free conversations. While free conversation is invaluable to the acquisition process, it is not sufficient for learners to fully develop their second language proficiency. Corrective feedback to learners' grammatical errors is necessary for improving accuracy in their interlanguage. The main contribution of this thesis lies in the research on intelligent corrective feedback that handles learners' errors and helps learners to use more appropriate words and expressions. Many studies on detecting grammatical errors and generating corrective feedback have been done in written mode, but there has been a dearth of research in oral mode. This thesis proposes an effective strategy of corrective feedback in oral communication based on second language acquisition theory and practice, and presents efficient methods that enable the strategy. Integrating these efforts resulted in intelligent educational robots and virtual 3D language learning

games. To verify the educational effects of our approaches on students' communicative abilities, we have conducted a field study at an elementary school in Korea. The results showed that our DB-CALL approaches can be enjoyable and fruitful activities for students.

Contents

List of Figures

List of Tables

Chapter I

INTRODUCTION

It is a fact that the private English education fee in Korea, reaching up to 16 trillion won annually, adds a great burden to Korean economy, resulting in countless articles overflowing in the media on strengthening the public education system that focuses on enhancing students' speaking ability to straighten out their hunchbacked English ability compared with the excessive grammar knowledge. This shows clear evidence for the necessity for changing our current foreign language education system in public schools which mainly focuses on vocabulary memorization and grammar-translation methodology.

Although there have been enormous investments into English education all around the world, not many differences have been made to change the rote learning style in English instruction. In addition, computer-based English learning is in the center of interest, however, this method also fails to provide the opportunity for free conversation and stays at the level of simple repetition of the given text. These teaching-learning methods cannot provide persistent motivation for learners to reach the high proficiency levels in foreign language learning.

Recent development of spoken dialog systems has enabled Computer-Assisted Language Learning (CALL) systems to bear a closer resemblance to oral conversation than the earlier CALL applications. We call such a communicative CALL system a Dialog-based CALL (DB-CALL) system that allows the user to engage in some form of meaningful dialog with embodied or disembodied agents. But studies on DB-CALL system are still relatively new and most are in the early stages. Many research projects have focused on adapting spoken dialog systems which were developed for native speakers to the use of language learning by incorporating non-native speakers' data. While free conversation is invaluable to the acquisition process, it is not sufficient for learners to fully develop their second language (L2) proficiency. Corrective feedback to learners' grammatical errors is necessary for improving accuracy in their interlanguage.

Considering the shortcomings for the current teaching-learning methodology, we have been investigating English learning systems using natural language processing technology in immersion context based on the assumptions of Second Language Acquisition (SLA) theory and practice. Through the systems, foreign language learners practice English conversation in natural contexts and are provided with corrective feedback based on the error correction procedures. POSTECH and KIST's Center for Intelligent Robotics have been cooperating in developing robots as educational assistants, called Mero and Engkey (Fig. I.1). These robots were designed with expressive faces, and have typical face recognition and speech functions allowing learners to have a more realistic and active context.

Another system, Pomy (POstech iMmersive English studY) (Fig. I.2), presents

Figure I.1: Mero and Engkey

a virtual reality immersion environment, where learners experience the visual, aural and tactual senses to help them develop into independent learners and increase their memory and concentration abilities to a greatest extent. Pomy also provides corrective feedback to learner's error by creating a ghost tutor character. The ghost tutor casts himself in the role of English tutor and helps learners to use more appropriate words and expressions. When a learner produces ungrammatical utterances, he provides both implicit and explicit feedback. Further, he can provide a full sentence for learners' fragmental responses. Even when the learner speaks without errors, the ghost tutor can give alternative examples of expressions to help the learner learn various new forms.

1.1 This Thesis

This thesis considers intelligent corrective feedback for language learning that handles learners' errors and helps learners to use more appropriate words and expressions. While many studies on detecting grammatical errors and generating corrective feedback have been done in written mode, there has been a dearth of research in

Figure I.2: Pomy

oral mode. Our ultimate goal is to design an effective strategy of corrective feedback based on SLA theory and practice, and to develop efficient methods that enable the strategy. On the basis of the SLA theory and practice, we have distinguished global error (nearly unintelligible) from local error (intelligible but partly incorrect) and designed different methods for each type of error. When a learner speaks, we judge whether the utterance has a global error or local errors. In case of global error, it would be better that the system provides a complete sentence that realizes the learner's intention, whereas, for local error, it would be wiser to pinpoint and correct the erroneous part with keeping other parts untouched.

In the remainder of this dissertation, we discuss how to handle global and local errors, introduce our approaches to this problem, and, finally, presents the educational effectiveness of our methods through comprehensive field studies.

1.2 Overview

Chapter II aims to give a literature review on SLA theory and previous CALL systems.

Chapter III considers global error handling. When language learners make global errors, a DB-CALL system should provide matching fluent utterances by inferring the actual learner's intention both from the utterance itself and from the dialog context as human tutors do. We propose an hybrid inference model that allows a practical and principled way of separating the utterance model and the dialog context model so that only the utterance model needs to be adjusted for each fluency level. Also, we propose a feedback generation method that provides native-like utterances by searching the example expression database using the inferred intention.

Chapter IV decribes our method for local error handling. It is not a trivial task to detect grammatical errors in oral conversations because of the unavoidable errors of automatic speech recognition systems. To provide corrective feedback, a novel method to detect grammatical errors in speaking performance is proposed. The proposed method consists of two sub-models: the grammaticality-checking model and the error-type classification model. We automatically generate grammatical errors that learners are likely to commit and construct error patterns based on the articulated errors. When a particular speech pattern is recognized, the grammaticality-checking model performs a binary classification based on the similarity between the error patterns and the recognition result using the confidence score. The error-type classification model chooses the error type based on the most similar error pattern and the error frequency extracted from a learner corpus.

Chapter V presents an automated method to generate realistic grammatical errors that can perform crucial functions for advanced technologies in CALL, including generating corrective feedback in DB-CALL systems, simulating a language learner to optimize tutoring strategies, and generating context-dependent grammar quizzes as educational materials. Our goal is to make grammatical errors generated by automatic simulators more realistic. To generate realistic errors, expert knowledge of language learners' error characteristics was imported into a statistical modeling system that uses Markov logic, which provides a theoretically sound way to encode knowledge into probabilistic first-order logic. We learned the weights of first-order formulas from a learner corpus.

Chapter VI introduces the educational assistant robots that we developed for foreign language learning and explores the effectiveness of DB-CALL which is in its early stages. To achieve this purpose, a course was designed in which students have meaningful interactions with intelligent robots in an immersive environment. A total of 24 elementary students, ranging in age from ten to twelve, were enrolled in English lessons. A pre-test/post-test design was used to investigate the cognitive effects of the DB-CALL approach on the students' oral skills. Descriptive statistics and the pre-test/post-test design were used to investigate the affective effects of DB-CALL approach.

Chapter VII gives conclusions and some thoughts on future work.

Chapter II

RELATED WORK

2.1 Second Language Acquisition Theory

There has been tremendous worldwide growth in using computer-based methods for learning different language skills and components. One of the ultimate goals of CALL is to provide learners with a good environment that facilitates acquiring communicative competence in the L2. Since the advent of SLA theories, a number of crucial factors have been revealed for improving students' productive conversational skills: (1) comprehensible input [1], (2) comprehensible output [2], (3) corrective feedback [3], and (4) motivation and attitude [4].

In relation to oral understanding, accumulated work on the process of listening suggests that comprehension can only occur when listeners place what they hear in context, i.e. the knowledge of who the participants are (sex, age, personality, relationship), the setting (where the linguistic situation takes place), the topic (the thing being talked about) and even the purpose (what language is used for) [5, 6].

What is really retained after understanding is not the literal meaning but some mental representation mainly provided by contextual information [7]. Hence it has become quite clear that in giving students comprehension activities out of context we set them a difficult task [8]. By placing learners in virtual worlds that imitate real life situations or in real worlds in which robots act as conversational agents, our systems are able to provide comprehensive inputs.

While comprehensible input is invaluable to the acquisition process, it is not sufficient for students to fully develop their L2 proficiency. The output hypothesis claims that production makes the learner move from 'semantic processing' prevalent in comprehension to more 'syntactic processing' that is necessary for improving accuracy in their interlanguage [2]. Specifically, producing output is one way of testing one's hypotheses about the L2. Learners can judge the comprehensibility and linguistic well-formedness of their interlanguage utterances against feedback obtained from their interlocutors, leading them to recognize what they do not know, or know only partially. The recognition of problems may then prompt the learners to attend to the relevant information in the input, which Schmidt [9] claims to be "the first step in language building." Additionally, output processes enable learners not only to reveal their hypotheses, but also to reflect on them using language. Reflection on language may deepen the learners' awareness of forms, rules, and form-function relationships if the context of production is communicative in nature. By adapting spoken dialog systmes for non-native speakers, our systems allow learners to speak freely so that learners can produce arbitrary erroneous output that reflects their interlanguage. In the aftermath, learners can judge whether their output is comprehensible or not based on system responses.

On the other hand, it has been argued that corrective feedback plays a beneficial role in facilitating the acquisition of certain L2 forms which may be difficult to learn through input alone, including forms that are rare, are low in perceptual salience, are semantically redundant, do not typically lead to communication breakdown, or lack a clear form-meaning relationship. Johnson [10] contends that if there is no concern for feedback in terms of linguistic correctness, meaning-based activities per se may accelerate language progress but in the long term lead to ‘‘fluent but fossilised students.’’ Our main contributions lie here, our systems can provide corrective feedback to learners' errors by virtue of the ability to detect global and local errors, which will be descried in more detail in the remainder of this thesis.

Motivation and attitude is another crucial factor in L2 achievement [4]. For this reason it is important to identify both the types and combinations of motivation that assist in the successful acquisition of a foreign language. In order to make the language learning process a more motivating experience, instructors need to put a great deal of thought into developing programs which maintain student interest and have obtainable short term goals. The use of an interesting computer-based method can help to increase the motivation level of students, and computer-based learning has an advantage over human-based learning in that it seems to be a more relaxed atmosphere for language learning [11, 12, 13].

2.2 CALL Systems

Computers have been viewed as a potentially beneficial tool for second language learning for several decades. With the explosion of Internet communication tools, several Computer-Mediated Communication (CMC) contexts have emerged such as instant messages, e-mails, chat rooms and discussion boards. CMC is widely discussed in language learning because CMC provides opportunities for language learners to practise their language. Early CMC research qualified and quantified language production from a mainly socio-cultural perspective (learner-learner and learner-teacher interactions). In recent years, a number of studies have investigated the role of written feedback for L2 development and have found a positive relationship between feedback and L2 development [14, 15, 16, 17, 18, 19].

With the advances in Natural Language Processing (NLP) technologies, a number of Intelligent CALL (ICALL) applications have emerged which employ sophisticated NLP techniques to provide dynamic, individualized feedback to learners' errors. Contrary to CMC applications which rely solely on human-human interactions, ICALL applications play crucial roles of both interlocutors and teachers. For example, BANZAI [20] employs artificial intelligence and NLP technology to enable learners to freely produce Japanese sentences and to provide detailed feedback concerning the specific nature of the learner's errors. E-tutor [21, 22] also provides error-specific and individualized feedback by performing a linguistic analysis of student input and adjusting feedback messages suited to learner expertise. Many studies have shown that students learn better with feedback that explains the particular error they are making and that considers their knowledge of the language.

However, the systems employed in this line of investigation can be described as non-communicative in that the primary focus of task interaction was on linguistic form, since no model of knowledge representation was present to facilitate meaning-focused exchanges. In recent years, there has been a shift in CALL research towards conversational interaction. This trend has been motivated by rapid globalization and great emphasis on communicative competence in the target language in a variety of situations. Recent development of spoken dialog systems has enabled CALL systems to bear a closer resemblance to oral conversation than the earlier CALL applications. We call such a communicative ICALL system a DB-CALL system. Many research projects have provided pronunciation training for oral skills using a speech recognizer in a forced recognition mode [23, 24], but a few systems exist that allow the user to engage in some form of meaningful dialog with embodied or disembodied agents in virtual words.

DEAL [25] is a spoken dialog system for providing a multidisciplinary research platform, particularly in the areas of human-like utterance generation, game dialogue, and language learning. The domain is the trade domain, specifically a flea market situation. DEAL provides hints about things the user might try to say if he or she is having difficulties remembering the names of things, or if the conversation has stalled for other reasons.

SCILL [26] was developed based on the spoken dialogue system of MIT. This system covers the topics of weather information and hotel booking. Researchers also implemented the simulated user to produce example dialogs to expose language learners to language use and to expand the training corpus for the system.

SPELL [27] provides opportunities for learning languages in functional situations such as going to a restaurant, expressing (dis-)likes, etc. Recast feedback is provided if the learner's response is semantically correct but has some grammatical errors. This system combines semantic interpretation and error checking in the speech recognition process. Thus, it uses a special speech recognition grammar to cover both normal speech and erroneous speech.

Let's Go [28] is a spoken dialog system that provides a bus schedule for the area around Pittsburgh, PA, USA. The researchers modified an extant system for the native speaker to adapt non-native speakers' data for the use of language learning. Modifications include the addition of new words, new constructs and the relaxation of some syntactic constraints to accept ungrammatical sentences. Based on the recognition result for the user utterance, they compute its distance to each target sentence and select the closest target. If words were deleted, inserted or substituted by the non-native, they generate a prompt that is both confirmation and correction.

In Japan, the educational use of robots has been studied, mostly with Robovie [29] in elementary schools, focusing on English language learning. Robovie has one hundred behaviors. Seventy of them are interactive behaviors such as hugging, shaking hands, playing paper–scissors–rock, exercising, greeting, kissing, singing, briefly conversing. For the purpose of English education, the robot could only speak and recognize English. In total, the robot could utter more than 300 sentences and recognize about 50 words. Robovie has tended to be extremely restrictive in the number of words it can recognize so that the conversations have been confined to a chain of short-time interactions.

IROBI [30] was recently introduced by Yujin Robotics in Korea. IROBI was specifically designed and trialled for tutoring and educational services. IROBI, which has a sitting child-like appearance, is designed with an LCD panel on its chest to support easy communication with children, allowing voice and touch screen input without face and gesture recognition. IROBI was used to compare the effects of non-computer-based media and web-based instruction with the effects of robot-assisted learning for children.

Studies on DB-CALL are still relatively new and most are in the early stages in a starting phase. Therefore, many attempts need to be made to investigate the effects of their use.

Chapter III

GLOBAL ERROR HANDLING

3.1 Introduction

SLA researchers have claimed that feedback provided during conversational interaction facilitates the acquisition process [31, 32]. When language learners make global errors (nearly unintelligible), helpful interactional processes include the negotiation of meaning and provision of recasts, both of which can supply corrective feedback to let learners know that their utterances were problematic. A further interactional process that can result from feedback is known as modified output. For example, consider the interactional processes, in which the system negotiates to determine the meaning using a clarification request in response to the learner's unnatural expression (Table III.1). The language learner modified the original utterance to convey the intended meaning by referring the recast provided by the system.

To achieve this goal, rule-based systems usually anticipate error types and handcraft a large number of error rules but this approach makes these methods fragile to

Table III.1: An example dialog in which the DB-CALL system returns a feedback recommending use of a native-like utterance

Speaker	Intention	Dialog
System:	wh-question(trip-purpose)	What is the purpose of your trip?
User:	inform(trip-purpose)	It's . . . I . . . purpose business
System:	clarify(understanding)	Sorry, I don't understand. What did you say? ← **Clarification request** ***On screen:*** *try this expression* *"I am here on business"* ← **Recast**
User:	inform(trip-purpose)	I am here on business ← **Modified output**

unexpected errors and diverse error combinations [33, 27, 28]. A more serious problem is that just correcting grammatical errors cannot guarantee that the utterance is fluent and meaningful. Therefore, we argue that the proper language tutoring methodology is not to correct specific errors but to provide native-like utterance examples which realize the user's intention.

To accomplish this purpose, as human tutors do, we first infer the actual learners' intention from the erroneous utterances by taking not only the utterance itself but also the dialog context into consideration, and then generate a corrective feedback based on the inferred intention.

The remainder of this chapter is structured as follows. Section 3.2 briefly describes related studies. Section 3.3 introduces the system architecture and operation. Section 3.4 presents the detailed description of our hybrid intention recognition model. Section 3.5 describes the experimental setup. Section 3.6 shows the experimental results to assess the method's potential usefulness. Finally, Section 3.7 gives our conclusion.

3.2 Related Works

There are several studies on general dialog systems which have examined incorporating the dialog context into recognizing dialog acts. Due to the difficulties of extracting and employing rich dialog context, most of them included just a few types of context such as previous dialog act [34], or dialog state in finite-state model [35]. Recently, Ai et. al [36] investigated the effect of using rich dialog context and showed promising results. The ways to incorporate the dialog context mostly involved just combining all features both from the utterance and the context into one feature set which was then used to train inference models. For DB-CALL, however, such approaches can be problematic, because distinct handling for each of fluency levels is important in a language learning setting. Given a dialog scenario, the dialog context model is relatively invariant; thus we propose a hybrid model that combines the utterance model and the dialog context model in a factored form. This approach allows us to adjust the hybrid model to a required fluency level by replacing only the utterance model.

3.3 System Architecture and Operation

The whole system consists of the intention recognizer and the dialog manager (Fig. III.1). The intention recognizer is a hybrid model of the dialog state model and one of the utterance models. A specific utterance model is chosen according to a learner's proficiency level. When the learner utters, the utterance model elicits n-best hypotheses of the learner's intention, and then they are re-ranked by the results of the dialog state model. The detailed algorithm will be described at the next section.

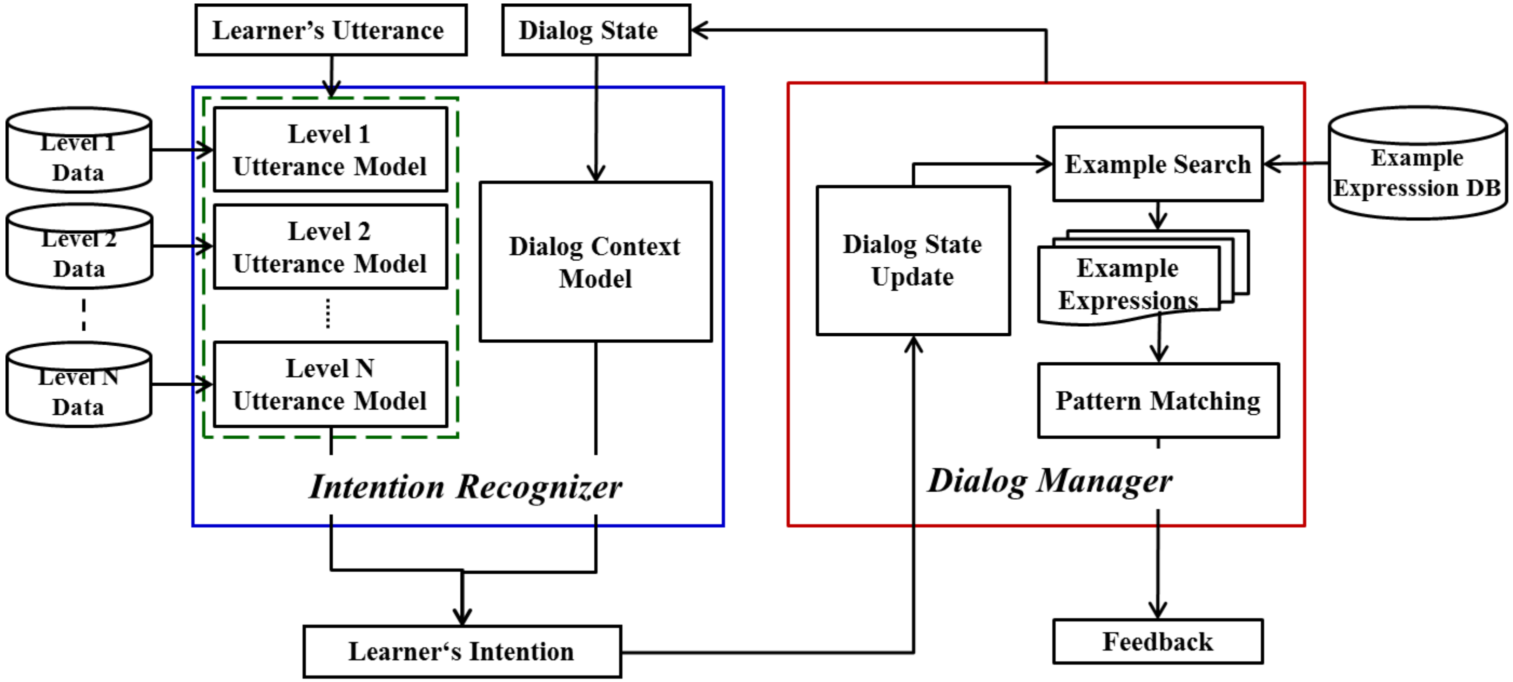

Figure III.1: System architecture

The role of the dialog manager is generating system responses according to the learner's intention and generating corrective feedback if needed. Corrective feedback generation takes two steps: 1) Example Search: the dialog manager retrieves example expressions by querying Example Expression DB (EEDB) using the learner's intention as the search key. 2) Example Selection: The dialog manager selects the best example which maximizes the similarity to the learner's utterance based on lexico-semantic pattern matching.

If the example expression is distant from the learner's utterance more than a pre-defined threshold, the dialog manager shows the example as recast feedback and conduct a clarification request to induce learners to modify their utterance (Table III.1). Otherwise, the dialog manager shows one of the retrieved examples as paraphrase feedback so that learners may acquire another expression with the same meaning. Sometimes, students have no idea about what to say and they cannot continue the dialog. In such a case, time out occurs and the utterance model does not generate hypotheses. Hence, the dialog system searches EEDB with only the result of the dialog state model and shows the retrieved expression as suggestion feedback so that learners can use it to keep a conversation.

3.4 Hybrid Intention Recognition Model

Our representation of user intention consists of dialog act and type of subtask as shown in Table III.1. For example, the first system utterance "What is the purpose of your trip?" can be abstracted by the intention wh-question (trip-purpose).

The hybrid model merges hypotheses from the utterance model with hypotheses from the dialog context model to find the best overall matching user intention. In the language production process, user intentions are first derived from the dialog context; subsequently the user intentions determine utterances [37]. By using this dependency and the chain rule, the most likely expected user's intention $I(U, D)$ given the dialog context D and the utterance U can be stated as follows:

$$I(U, D) = argmax_I P(I|U, D) \tag{III.1}$$

$$I(U, D) = argmax_I \frac{P(I, U, D)}{P(U, D)} \tag{III.2}$$

$$I(U, D) = argmax_I \frac{P(U|I)P(I|D)P(D)}{P(U, D)} \tag{III.3}$$

By using Bayes' rule, Eq. III.3 can be reformulated as:

$$I(U, D) = argmax_I \frac{P(U)P(U|I)P(I|D)P(D)}{P(U, D)P(I)} \tag{III.4}$$

$P(U)$, $P(D)$, and $P(U, D)$ can be ignored, because they are constant for all I (Eq. III.5):

$$I(U, D) = argmax_I \frac{P(I|U)P(I|D)}{P(I)} \tag{III.5}$$

In this formula, $P(I|U)$ represents the utterance model and $P(I|D)$ represents the dialog-context model. The next two subsections discuss each sub-model in detail.

3.4.1 Utterance model

To predict the user intention from the utterance itself, we use maximum entropy model [38] trained on linguistically-motivated features. This model offers a clean way to combine diverse pieces of linguistic information. We use the following linguistic features for the utterance model.

- **Lexical word features:** Lexical word features consist of lexical trigrams using current, previous, and next lexical words. They are important features, but the lexical words appearing in training data are limited, so data sparseness problems can arise.

- **Part-of-speech (POS) tag features:** POS tag features also include POS tag trigrams matching the lexical features. POS tag features provide generalization power over the lexical features.

The objective of this modeling is to find the I that maximizes the conditional probability, $P(I|U)$ in Eq. III.5, which is estimated using Eq. III.6:

$$P(I|U) = \frac{1}{Z} exp\left(\sum_{k=1}^{K} \lambda_k f_k(I,U)\right), \tag{III.6}$$

where K is the number of features, f_k denotes the features, λ_k the weighted parameters for features, and Z is a normalization factor. We use a limited memory version of the quasi-Newton method (L-BFGS) to optimize the objective function.

3.4.2 Dialog-context model

Our representation of a dialog context consists of diverse pieces of discourse and subtask information as shown in Table III.2.

The task of predicting the probable user intention in a given dialog context can be viewed as searching for dialog contexts that are similar to the current one in

Table III.2: Representation of dialog context and an example for immigration domain

Dialog Context Features	
PREV_SYS_INT	Previous user intention Ex) PREV_SYS_INT = wh-question(job)
PREV_USR_INT	Previous system intention Ex) PREV_USR_INT = inform(job)
SYS_INT	A list of exchanged information states which is essential to successful task completion; (c) denotes confirmed, (u) unconfirmed Ex) INFO_EX_STAT = [nationality(c), job(u)]
INFO_EX_STAT	Previous system intention Ex) PREV_SYS_INT = wh-question(job)
DB_RES_NUM	Number of database query results Ex) DB_RES_NUM = 0

dialog context space and then inferring to the expected user intention from the user intentions of the dialog contexts found. Therefore, we can formulate the task as the k-nearest neighbors (KNN) problem [39]. We had a number of reasons for choosing instance-based learning methodology. First, instance-based learning provides high controllability for tuning the model incrementally during operation, which is practically very desirable property. Second, an elaborate similarity function can be applied. Many of other approaches, e.g. maximum entropy model used in the utterance model, express the similarity between states in a simplest manner through the features that the states share, losing elaborate regularities between features. For the dialog context model, we can easily predict which features become important features to measure similarity conditioning on certain values of other features using general discourse knowledge. For example, if the current system dialog act is “inform”, the number of database query results becomes an important feature. If the number of results is greater than one, the most likely expected user intention would be “select”. If the number of results equals one, "ack" would be the most probable intention. Otherwise, the users might want to modify their requirements. Another example, if all exchanges of information are confirmed and the current system intention is "wh-question", the current system intention itself becomes the most important feature to determine the next user intention.

However, the conventional KNN model has two drawbacks. First, it considers no longer the degree of similarity after selecting k nearest contexts, hence intentions that occur rarely cannot have a chance to be chosen regardless of how close they are to the given dialog context. The second drawback is that if dialog contexts with, say, intention A, are locally condensed rather than widely distributed, then A is specifically fitted intention to the local region of the dialog context. So the intention A should be given greater preference than other intentions. To cope with these drawbacks, we introduce a new concept, locality, and take both similarity and locality into account in estimating the probability distribution of the dialog context model (Eq. III.8, III.9).

The similarity function is defined as the following equation:

$$Similarity(D, D') = \sum_{k=1}^{K} \lambda_k f_k(D, D'), \tag{III.7}$$

where D and D' are dialog states, K is the number of features, f_k denotes the feature functions, λ_k the weighted parameters for features. Our feature functions first include the simplest tests, whether a feature is shared or not, for each feature of a dialog context (Table III.2). For composite features, individual tests are also included for each constituent to alleviate data sparseness problems. For example, we include feature functions not only for system intention but also for its constituents, system dialog act and type of subtask. In addition, we include a number of feature functions which test the elaborate rules as illustrated in the examples above. The weighted parameters are given initial values based on general discourse and task knowledge and optimized on the development data set with minimum error rate criteria.

The locality function is the ratio between the number of elements of the set $S_{I,D}$, and the number of elements of the set S_I:

$$Locality(I, D) = \frac{n(S_{I,D})}{n(S_I)}, \tag{III.8}$$

where $S_I = \{D' | \forall D' \, intention(D') = I\}$, and $S_{I,D} = \{D' | \forall D' \in S_D \, and \, intention(D') = I\}$, and S_D is the set of k nearest neighbors of the given dialog context D.

The score function calculates the score of the intention I based on the set of k nearest dialog contexts using both similarity and locality:

$$Score(I, D) = Locality(I, D) \sum_{D' \in S_{I,D}} Similarity(D, D'), \tag{III.9}$$

To let the dialog-context model be a probability distribution, the score function is divided by the normalization factor:

$$P(I|D) = \frac{Score(I, D)}{\sum_{I'} Score(I', D)} \tag{III.10}$$

3.5 Experiments

To verify the proposed model, we conducted a case study for dialogs in a system devoted to immigration into an English-speaking country. We used the example based dialog management method [40] to build the dialog system. We used the dialog corpus in an immigration domain which consists of 192 dialogs (3,517 utterances, 18.32 utterances per dialog). The number of types of dialog act is seven and the number of types of subtask is eleven. We annotated each utterance in dialogs by two steps. In the first step, we manually annotated each utterance with the speaker's intention and component slot-values. In the second step, we automatically annotated each utterance with the discourse information (Table III.2) using in-house software.

3.5.1 Example dialog

The example dialog gives a flavor of the system in real action. A brief commentary is given at key points in the dialog and is printed in italics.

S1: Hi. This is a dialogue system for English learning. May I ask your name for authentication?

U1: My name is Hong Kil Dong

S2: You are a new user. We have four themes such as immigrant, transportation, hotel, restaurant theme. Which theme do you want to learn?

U2: I'd like the immigrant theme

S3: The immigrant theme is selected. Let's get started

S4: Hello

U3: Hi

S5: Your passport please

U4: Here this

On screen: *[Recast] Here you are [Recast] is an example utterance provided by recast feedback. "Here this", the student's utterance, is understandable based on context but it is not a good expression. The system recommends a better expression so that the student becomes better in a next dialogue.*

S6: Sorry, I don't understand. What did you say?

U5: Here you are

S7: What's your nationality?

U6: I am a Korean

On screen: *[Paraphrase] I come from Korea When the user utterance is correct the system shows another possible expression by paraphrase feedback.*

S8: What's the purpose of your visit?

U7: Business

On screen: *[Paraphrase] I am here on business If a user input is simple words or a phrase, then the system shows a complete sentence.*

S9: How long are you staying?

U8: I stay for three weeks

On screen: *[Recast] I will stay for three weeks The user utterance is semantically understandable but it is grammatically wrong. The system shows a grammatically correct sentence on screen which is obtained by searching Example Expression DB.*

S10: Sorry, I don't understand. What did you say?

U9: I will stay for three weeks

S11: Anything to declare?

U10: Pardon me?

S12: Anything to declare?

U11: Nothing

On screen: *[Paraphrase] Just my personal things*

S13: What kind of job?

U12: [TIMEOUT]

On screen: *[Suggestion] I am a student [Suggestion] is an example utterance provided by suggestion feedback. The system knows that it asked the user about a job in the previous question. And it also knows that it already got information about nationality, purpose of visit and period through the dialog history. Therefore by using such dialogue history the system can find out a proper example from Example Expression DB.*

U13: I am a computer programmer

S14: Ok, no problem

U14: Thank you. Bye

S15: The immigrant theme has been done. Your total score is B. In next time, you better use more comprehensive and complete expressions, and try to say more quickly. Would you like to continue learning another theme?

The system provides assessment feedback which includes total score and specific messages that inform the student which part they need to improve.

U15: No.

S16: Study session ends.

3.5.2 Simulated language learner

In this experiment, instead of involving real language learners, we simulated them by injecting grammatical errors into clear utterances generated using the user simulation method described in [41]. We did this because we want to freely adjust the fluency level of users for intensive tests in a variety of settings, and because the cost of engaging large numbers of language learners with diverse fluency levels during development is particularly high. Employing a simulated user will greatly accelerate the development cycles. To simulate various fluency levels, we examined the Korean Student English Essay (KSEE) corpus which is a collection of essays by Korean university students. The corpus consists of 119 essays (915 sentences). We annotated the corpus with the error tags defined in [42]. The frequencies of error

Table III.3: Three categories of error types and the top 5 error types in each category

Category	**Error type with category**
Substitution	Spell (71%)
(47%)	Plural Form (14%)
	Subject Verb Agreement (10%)
	Incorrect Preposition (3%)
	Incorrect Determiner (2%)
Deletion	Missing Determiner (62%)
(36%)	Missing Preposition (18%)
	Missing Conjunction (13%)
	Missing Verb (4%)
	Missing Subject (3%)
Insertion	Extra Preposition (36%)
(17%)	Extra Determiner (26%)
	Extra Conjunction (20%)
	Extra Verb (15%)
	Extra Intensifier (3%)

types were measured. In total, 65 error types and 2,404 instantiated errors were discovered. We classified error types into three categories: substitution, insertion, and deletion. For each category, we listed the five most common error types (Table III.3) which account for 73% of the errors. As Foster [43, 44] and Lee [45] generated a treebank of ungrammatical English, we also produced artificial grammatical errors systemically. The error generation procedure takes as input a part-of-speech tagged sentence which is assumed to be well-formed, and outputs a part-of-speech tagged ungrammatical sentence. In the first step of the error generation procedure, we set the Grammar Error Rate (GER) between 0%   100% and determined error counts to be produced based on the GER. Then, we distributed the errors among categories and error types according to the percentages in the error types list (Table III.3).

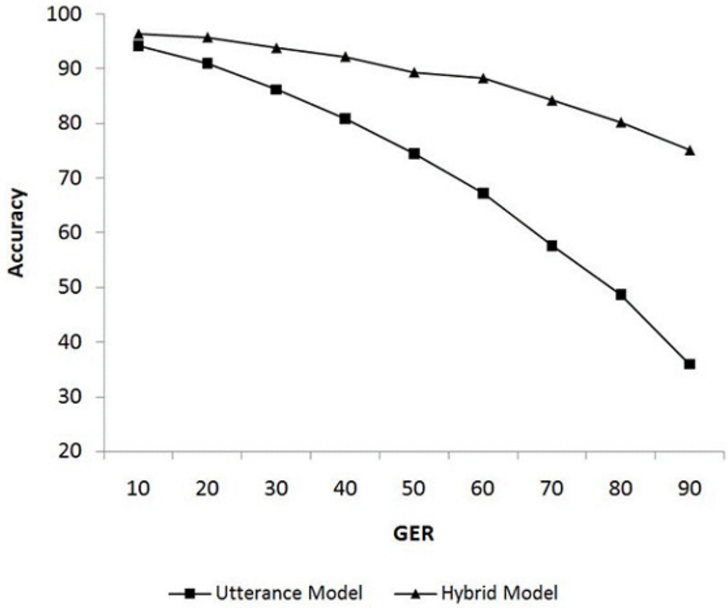

Figure III.2: Comparison between the hybrid model and the utterance only model

3.6 Results

3.6.1 Hybrid model vs. Utterance model

To verify the effectiveness of the dialog state-awareness, we compared the hybrid model with the utterance model. The utterance model just omits the dialog context model from the hybrid model. We conducted 200 dialogs for each model per 10% GER intervals. The hybrid model significantly outperformed the utterance model for overall range of GER. As the GER increased, the performance of the utterance model decreased dramatically, whereas the performance of the hybrid model decreased smoothly (Fig. III.2). It verifies the effectiveness of dialog state-awareness through our hybrid approach.

3.6.2 Appropriateness of feedback

On the contrary to the task oriented dialogs, language tutoring systems do not need to exactly recognize the learner's intention. Even if the inferred intention is not the same as the actual one, the feedback can be valuable for language acquisition as long as the feedback is appropriate to the dialog context. In fact, human tutors also generate feedback relying on only the dialog context when the learners' utterances are highly incomprehensible. Often, without such feedback, the conversation even can be stuck with a learner's problematic utterance, thereby cannot successfully finish. As McClelland [46] noted the role of success motivation in learning, the completion of a dialog itself is undoubtedly important rewards in foreign language learning. Therefore, we want our hybrid model to provide feedback appropriate to the dialog context regardless of whether the inferred intention exactly equals the original one or not.

To evaluate the appropriateness of the feedback, we conducted 200 dialogs per 10% GER intervals from 10% to 90%, and observed the Dialog Completion Rate (DCR) as the GER increased. As the GER increased, the performance of the hybrid model decreased, whereas the DCR decreased very slightly (Fig. III.3). Because of the clarification sub-dialogs, the average dialog length increased as the GER increased. Based on this result, we can conclude that our method is suitable to produce appropriate feedback even when the inferred intention is not the same as the actual one. This is because the dialog context model effectively confines candidate intentions within the given context.

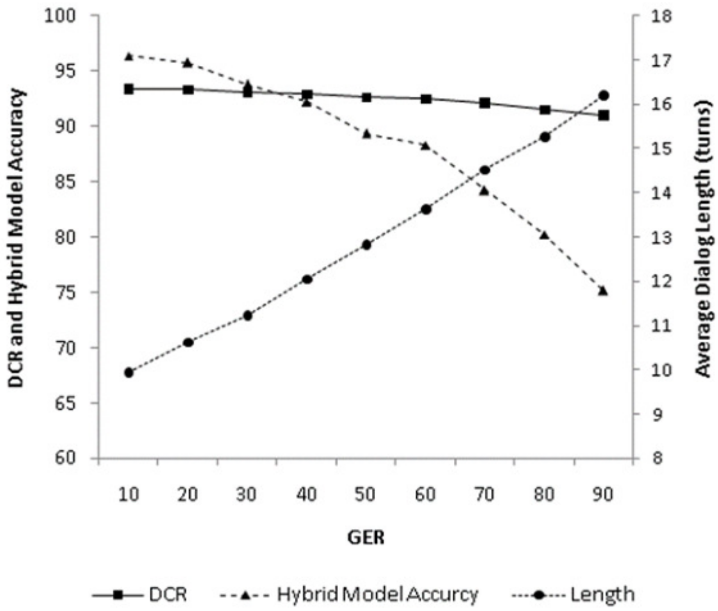

Figure III.3: The relation between Dialog Completion Rate and the performance of the hybrid model and the Average Dialog Length

3.7 Conclusion

When language learners speak incomprehensible utterances, a DB-CALL system should provide matching fluent utterances. We proposed a novel hybrid model that allows natural decomposition between the utterance model and the dialog context model in terms of the psychological language production process. It led to an efficient way for adjusting to diverse fluency levels with minimal efforts. In addition, our elaborate dialog context model using enhanced k-nearest neighbors algorithm gave rise to more accurate inference of the language learners' intention. Also, it proved to be effective to provide appropriate context-aware feedback so that the learners can obtain positive rewards by successfully completing dialogs.

Chapter IV

LOCAL ERROR HANDLING

4.1 Introduction

Since free conversation is only good for raising fluency in speaking a second language, a learner must also acquire other important aspects of the spoken language, such as morphology and syntax. Therefore, when language learners make local errors (intelligible but partly incorrect), it would be wiser to pinpoint and correct the erroneous part with keeping other parts untouched. To account for this fact, CALL systems have been developed to detect grammatical errors in speaking performance, provide learners with corrective feedback, and allow learners to try repeatedly until they manage to produce the correct form.

However, it is not a trivial task to detect grammatical errors in oral conversations because of the unavoidable errors of Automatic Speech Recognition (ASR) systems. The ASR errors make it mostly impossible to employ parser-based methods which have usually been developed to detect grammatical errors in learners writings [22].

As grammatical error detection in speaking performance is in a relatively early stage, only a few reports have been published. In addition, most previous studies have lacked proper evaluations to judge the usefulness for language tutoring. In this chapter, we propose a novel method capable of handling ASR errors and we provide several evaluation results that are helpful in considering the practicality of the method.

The remainder of this chapter is structured as follows. Section 4.2 briefly describes related studies. Section 4.3 presents a detailed description of the methods. Section 4.4 outlines the experimental setup. Section 4.5 shows the results and discusses their meaning. Finally, Section 4.6 offers our conclusion.

4.2 Related Work

Many research projects have tested the idea of providing corrective feedback to grammatical errors in writing, but few systems exist that detect grammatical errors in speaking performance and provide learners with corrective feedback.

The Let's Go system [28] is a spoken dialog system that provides bus schedules. The researchers adapted non-native speakers speech data and modified the semantic-parsing grammar that originally was developed for the native speaker. Modifications include the addition of new words, new constructs and the relaxation of some syntactic constraints to accept ungrammatical sentences. Based on the recognition result for the user utterance, the system computes its distance to each target sentence using dynamic programming and selects the closest target. If the two match exactly, no correction is produced and the dialogue continues normally. If words were deleted, inserted or substituted by the non-native speaker, they generate both confirmation and correction. The idea is that whenever a non-native speaker utters an ungrammatical utterance, the speakers goal was actually to utter one of the target sentences, but the speaker made a mistake by inserting, deleting or substituting a word. The evaluation results, however, showed numerous false positives (i.e., the user utterance was judged as ungrammatical although it was grammatical), and most of them were caused by ASR errors. This result clearly shows that we need to take into consideration ASR errors when we judge grammaticality to reduce false positives.

The Spoken Electronic Language Learning (SPELL) system [27] provides opportunities for learning languages in functional situations such as going to a restaurant or expressing (dis-)likes. Recast feedback is provided if the learner's response is semantically correct but has some grammatical errors. To reduce the confusion between ASR errors and grammatical errors, the system embeds error checking into the speech recognition process. Within the constrained environment defined for SPELL, it is readily possible to predict to a reasonable degree what learners might say at any given stage; similarly, it is then possible to predict certain grammatical errors that they might make. The aim is to develop finite-state network (FSN)-based recognition grammars specifically for non- native speakers that take into account both grammatical and ungrammatical predicted responses (Fig. IV.1). However, this study did not conduct experiments on the performance of the error detection component. Therefore, we implemented the method as the baseline system and performed comparative experiments with our method.

The Development and Integration of Speech technology into COurseware for language learning (DISCO) system [47] is under development and supposed to extend

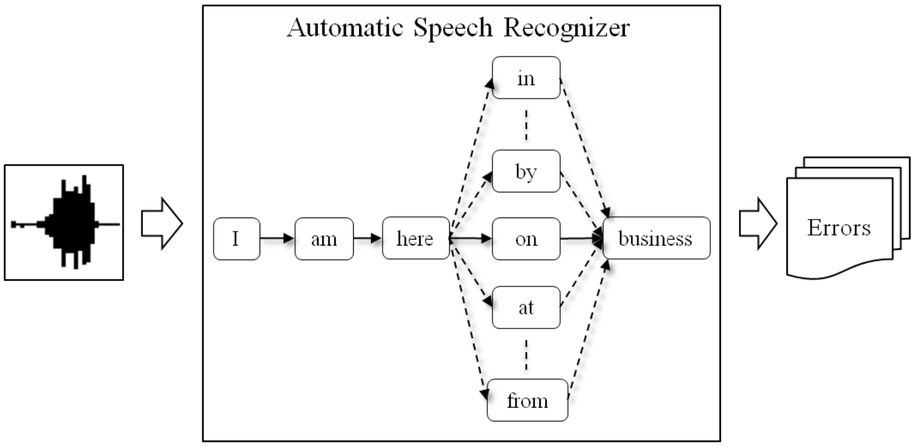

Figure IV.1: An example of FSN-based recognition grammar to detect possible preposition errors for the correct response "I am here on business"

the previous pronunciation training project to morphology and syntax training in well-designed exercises. The aim of the DISCO project is to optimize Dutch learning through interaction in realistic communication situations and provide intelligent feedback on important aspects of speaking. For detecting morphological and syntactic errors, grammatical error simulation software can be used. This software takes appropriate responses as input and expands them to form pools of correct and incorrect responses. Similar to SPELL, the speech recognition module determines which utterance was spoken and the system determines whether errors have been made depending on which of the possible utterances has been recognized. The evaluation on the DISCO system has not yet been performed because the system is currently under development.

4.3 Grammatical Error Detection

The simplest way to detect grammatical errors in speaking performance while reducing the hindrance of ASR errors is the method employed in SPELL and DISCO. The system takes appropriate responses as input and expands them using a grammatical error simulator to form FSN-based recognition grammars that include both correct and incorrect responses. The system determines whether errors have been made depending on which of the possible utterances has been recognized.

However, this approach has severe drawbacks. As the grammar size exponentially increases because of the numerous ungrammatical responses, the recognition performance sharply decreases. Often the recognized hypothesis could be a totally different utterance because the FSN-based Viterbi-decoding searches for the hypothesis at a nearly utterance level. Moreover, even if the recognized hypothesis is similar to the learners speech, it could be useless for error detection. Because of the grammatical error simulation, we have many similar ungrammatical variants of a correct response. When the learners utterance is one of the variants, it is highly likely that these similar variants are placed on the N-best hypotheses. However if the right hypothesis is not the top hypothesis, the system would produce wrong feedback because the system takes only the top hypothesis.

Therefore, we investigate a method that uses ASR systems with an N-gram language model to not get a totally different hypothesis and that considers multiple hypotheses based on confidence scores at a word level by exploiting a confusion network (CN) [48]. According to [48], the posterior probability of a word hypothesis can serve as a confidence score for the word to occur at the position. Unlike previous

Grammatical Error Detection	I	am	here	at	business
1) Grammaticality Checking	0	0	0	1	0
2) Error Type Classification	None	None	None	PRP_LXC	None

Figure IV.2: The grammatical error detection model consists of two sub-models

methods that just rely on the Viterbi- decoding process of ASR systems, this approach allows us to use machine learning techniques that we can try various useful features and have more opportunity to optimize a sophisticated objective function such as a low false positive rate and a high F-score.

Besides ASR errors, there are several factors that make it hard to detect grammatical errors. Because there are far more grammatical words than ungrammatical words in the data, the grammatical error detection model, which is implemented as a classifier, must be constructed to effectively learn from the imbalanced data distribution. When accuracy is the performance measure, using the classifier trained on the highly imbalanced data simply produces the majority class for all test data to achieve the best performance. In addition, the number of error types to classify is relatively large. This can make the model learning and selection procedure vastly complicated. Therefore, to cope with these difficulties, we divide the grammatical error detection model into two sub-models: the grammaticality-checking model and the error-type classification model (Fig. IV.2).

4.3.1 Grammaticality Checking Model

The grammaticality-checking task takes the recognized hypothesis in the form of a CN and determines the grammaticality at each word position in sequence. Even without error type information, the grammaticality- checking function may be very useful for some applications, e.g., categorizing learners proficiency level and generating implicit corrective feedback such as repetition, elicitation, and recast feedback.

Feature Extraction

To judge the grammaticality, we first extract error patterns from the simulated ungrammatical responses. The error pattern is a 5-tuple consisting of the erroneous word and its two left and two right neighbor words. For example, the error pattern for the proposition error at 'at' for the utterance 'I am here at business' will be a tuple ⟨'am', 'here', 'at', 'business', '-'[1]⟩. The error pattern is also tagged with the error type and structural deviation (e.g., deletion or substitution) for the error-type classification task.

When a speech is recognized, at each position in the CN, we extract a feature vector by comparing the error patterns with the segment of the CN, consisting of the target position and the two left and right neighboring positions. We extracted seven features (Table IV.1) for each error pattern. For example, if the first word in the error pattern exists among the competing word hypotheses at the first position in the CN, then we take the confidence score of the matched word hypothesis as the S1 feature. If there is no matched word hypothesis, we simply set the feature to zero. The higher the matching scores an error pattern has, the more likely the

[1] - is a blank symbol.

Table IV.1: Description of features extracted from each error pattern to train the grammaticality checking model

Feature	**Description**
S1	Confidence score of the word hypothesis matching the first word in the error pattern
S2	Confidence score for the second word in the error pattern
S3	Confidence score for the third word in the error pattern
S4	Confidence score for the fourth word in the error pattern
S5	Confidence score for the fifth word in the error pattern
TS	Total score of L2, L1, TW, R1, and R2.
SD	Indicator of structural error type: 1 for Deletion and 0 for Substitution

recognized result has the relevant error in it. Because the number of error patterns is very large and likely uninformative, only the features extracted from top 10 error patterns ranked by the TS feature are used. In addition, we perform a similar feature extraction process at the parts-of-speech (POS) level. We apply POS tagging to both the recognition result and the error patterns to get additional features from the top 10 POS-level error patterns. The POS-level features contribute to raising the recall rate by alleviating the data sparseness problem of lexical-level features.

Fig. IV.3 depicts the aforementioned feature extraction process.

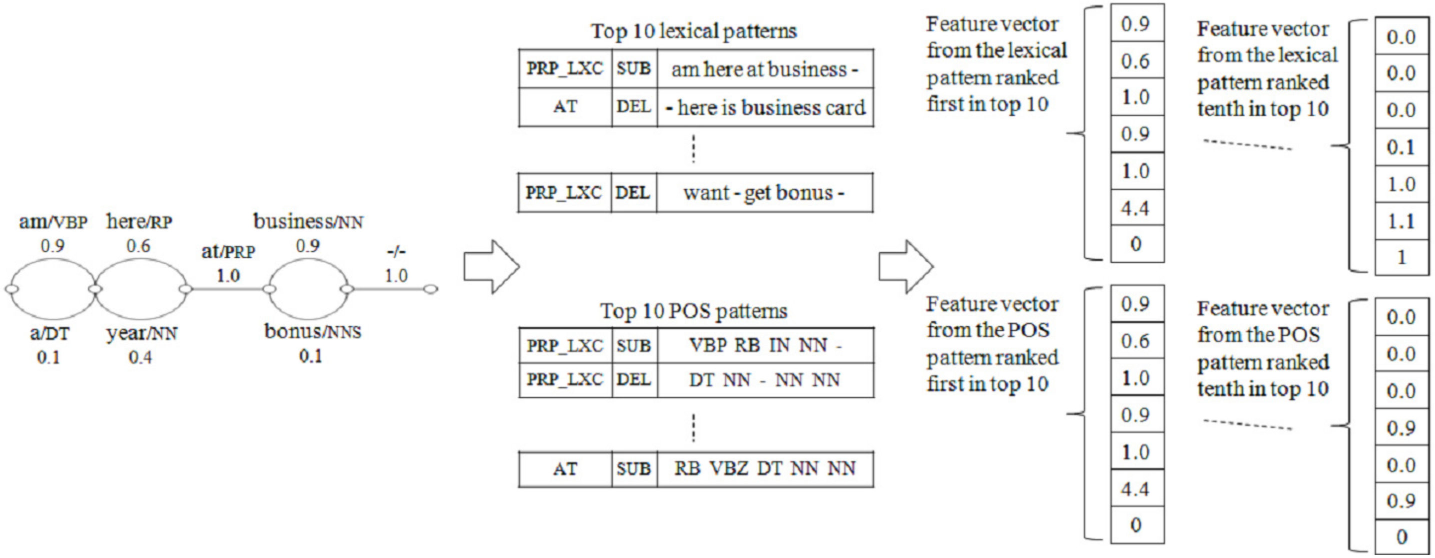

Figure IV.3: An illustration of feature extraction process. PRP_LXC and AT denote proposition lexical error and article error. SUB and DEL mean substitution and deletion

Model Selection and Parameter Learning

We use the LIBSVM [49] Support Vector Machine (SVM) classifier to produce a model that predicts grammaticality. We use a radial basis function (RBF) as the kernel because unlike linear kernels, an RBF kernel allows us to handle nonlinear interactions between attributes (e.g., dependency between the feature SD and the other features in Table IV.1) and relationship between class labels and attributes. We conduct simple scaling on the data to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. We linearly scale each attribute to the range [0, 1].

As mentioned before, the grammaticality-checking task is non-trivial because of the highly imbalanced data distribution. To address this problem, we can oversample the minority class or undersample the majority class to make the data balanced. Also we can implement cost-sensitive learning to assign a larger penalty value to false negatives versus false positives. Without loss of generality, we will assume that the positive class is the minority class (i.e., ungrammatical), and the negative class is the majority class. However, these approaches do not explicitly optimize the objective function that is important for a tutoring setting. For example, false positives (i.e., the user utterance was judged as ungrammatical although it was grammatical) are more detrimental than false negatives. Furthermore, precision is more important than recall. Therefore, in this study, we solve the problem by using a custom objective function instead of using accuracy as a performance measure. The objective function to optimize for this study is:

$$\begin{aligned} &Maximize\ F - score \\ &Subject\, to\ Precision > 0.90, \\ &\quad False\, positive\, rate < 0.01 \end{aligned} \tag{IV.1}$$

There are two parameters for an RBF kernel: C and γ. To find the best parameters C and γ that optimize the objective function, we perform a grid-search using 5-fold cross-validation.

4.3.2 Error Type Classification Model

To provide meta-linguistic feedback (i.e., detailed explanations about the grammatical error), we need to identify the error type. Identifying the error type is also beneficial to construct the learner model. Thus, we perform error-type classification for the words that are determined as ungrammatical by the grammaticality-checking model. The simplest way to classify the error type is to choose the error type associated with the top ranked error pattern. But this approach has two flaws: it does not have a principled way to break tied error patterns, and it does not consider the error frequency. Therefore, to solve both problems at the same time, we reorder error patterns by weighting more heavily errors that occur more frequently:

$$Score(e) = TS(e) + \alpha * EF(e), \tag{IV.2}$$

where TS returns the TS feature of the error pattern e and EF returns the error frequency of the relevant error, normalized summing to one. We set the constant α

as 0.1 for this study.

4.4 Experimental Setup

To evaluate the proposed method, we apply the method to detect grammatical errors of Korean learners of English.

4.4.1 Grammatical Error Simulation

One of the key elements to the development of ASR-based CALL systems for morphology and syntax training is to expand the recognition grammar to include not only grammatical responses but also ungrammatical responses. In SPELL, as each new scenario is developed, it is essential to create ungrammatical responses by hand. However, using human experts to anticipate various types of grammatical errors and list all possible realizations of the errors is too laborious and costly. Thus, as in DISCO, automatic generation of realistic grammatical errors to create recognition grammars is crucial to the development of such systems.

We developed a grammatical error simulator that generates errors that Korean learners of English usually make[2]. To generate realistic errors, expert knowledge of language learners error characteristics was imported into a statistical modeling system that uses Markov logic [50]. A Markov logic network can be seen as a first-order knowledge base with weights attached to each of the formulas. A total of 119 Markov logic network formulas were written. For example, English learners often commit pluralization errors with irregular nouns. These errors result because they

[2] In this section, we only give a brief account of grammatical error simulation, please refer to Chapter V for a detailed exposition.

over-generalize the pluralization rule, e.g., attaching 's/es' to the end of a singular noun, so that they apply the rule even to irregular nouns such as 'mice' and 'feet'. This characteristic is captured by the simple formula:

$$\begin{aligned} & IrregularPluralNoun(s,i) \wedge PosTag(s,i,NNS) \\ & \implies ErrorType(s,i,N\_NUM), \end{aligned} \tag{IV.3}$$

where $IrregularPluralNoun(s,i)$ is true if and only if the ith word of the sentence s is an irregular plural, NNS stands for plural noun, and N_NUM is the abbreviation for noun number error. We learned the weights of first-order formulas from the NICT JLE corpus [51]. This is a speech corpus of Japanese speakers learning English[3] . The corpus data were obtained from 1,281 audio- recorded speech samples, 167 of which are error-annotated, from an English oral proficiency interview test. The current version of the error tagset targets morphological, grammatical, and lexical errors and can describe diverse grammatical errors. The error tagset currently includes 46 tags. Because of the space limitation, please refer to [51] for the full list of error types. Lexis errors related to open-word class (i.e., n_lxc, v_lxc, aj_lxc, and av_lxc), were excluded in this experiment because realizing such errors without encountering the data sparseness problem requires a huge amount of learner data. Some other errors (i.e., o_je, o_lxc, o_odr, o_uk, and o_uit) were also excluded because these error categories have not yet been clearly analyzed for practical applications. Error categories that occurred less than five times were also excluded to improve reliability. This results in a total of 23 error types. In addition, we did not explicitly generate

[3]Unfortunately, there is no Korean learners corpus. But Korean and Japanese speakers learning English have very similar error characteristics because the two languages have very similar grammatical structures.

insertion errors because many insertion errors appear implicitly as replacement errors in the NICT JLE corpus. The insertion errors, which are not covered in this model, usually relate to vocabularies in open-word classes or are highly unpredictable even when linguistic context is taken into account.

4.4.2 Data Preparation and Setup of Grammatical Error Detection Models

We took 100 utterances from the NICT JLE corpus and expanded it to form the pool of 5000 utterances using the grammatical error simulator. For the training data, we randomly chose 250 utterances from the utterance pool and ten male Korean speakers to each read 50 utterances, resulting in 500 recordings. For the test data, we randomly chose 50 utterances from the utterance pool and ten male Korean speakers to each read the utterances, resulting in 500 recordings.

We developed our own English ASR system to recognize Korean learners English more reliably. The acoustic model is based on 3-state left-to-right, context- dependent, 8-mixture, and cross-word tri-phone models, trained on the Korean-Spoken English Corpus [52] using the HTK version 3.4.1 toolkit [53]. A backed-off bigram trained on the 5000 utterances is used as a language model to cover both grammatical and ungrammatical utterances. The lattice output of the speech recognizer is converted to the CN using the lattice-tool [54]. After constructing the English ASR system, the recognition performance of the ASR system was evaluated on both the traininig and test speech data. The word error rate was 15.20% at the vocabulary size of 530.

For a comparative evaluation on the grammaticality- checking task, we developed the proposed method and also implemented the FSN-based ASR system (FSN in Fig. IV.4) which was employed in SPELL and DISCO as the baseline system. In addition, to verify the effect of confidence score-based soft match, we developed an exact pattern match-based method (EPM in Fig. IV.4) that judges the recognition result as ungrammatical only when there is an error pattern that exactly matches the sequence formed by picking the word with highest confidence score at each position in the CN. For a comparative evaluation on the error-type classification task, we developed the proposed method with the error frequency estimated from the NICT JLE corpus and implemented the method that takes the error type of the top ranked error pattern as the baseline system.

4.5 Results and Discussion

The results showed that the proposed model largely outperformed the baseline FSN model for all metrics (Fig. IV.4). It is because the FSN-based Viterbi-decoding exhibited a very low sentence-level recognition performance due to the relatively large size of the recognition grammar consisting of many similar variants for various grammatical errors. This affects not only the precision and recall but also the false-positive rate, where it can be detrimental for language tutoring by frustrating learners. The proposed method also surpasses the EPM model in F-score. It is attributed to the large gain in the recall rate. The proposed method achieves a far higher recall rate than that of the EPM model by exploiting a soft pattern match based on the confidence score. Furthermore, the proposed method lost little precision by virtue of the SVM model optimization to satisfying the constraints on the precision and false

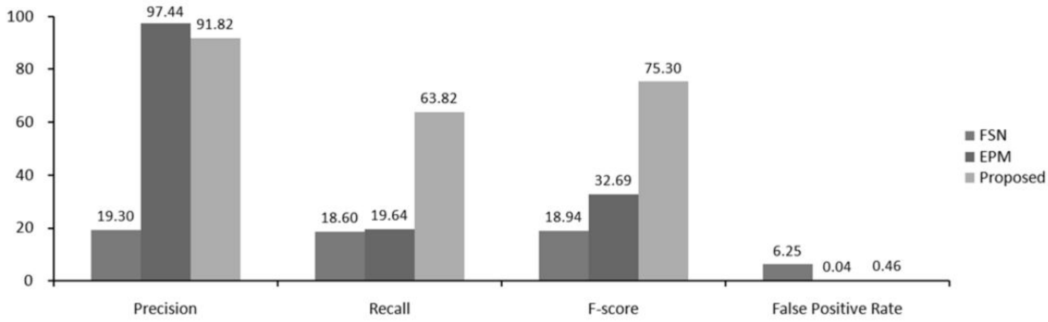

Figure IV.4: Experimental results on the grammaticality checking task

positive rate. Both the EPM model and proposed model showed a very low false positive rate. This implies that the proposed method is very suitable for educational applications. For the error-type classification task, the baseline method that does not consider the error frequency showed an accuracy of 95.55%. The proposed method improved the baseline performance by 4.05%. The result of the baseline model is quite good already, but the incorporation of error frequency into the model gives us an additional performance gain.

4.6 Conclusion

This chapter proposed a novel method to detect grammatical errors in speaking performance to provide corrective feedback on grammatical errors. The results showed that for the grammaticality-checking task, the proposed method largely outperformed the two comparative models respectively by 56.36% and 42.61% in F-score while keeping the false positive rate very low. For the error-type classification task, the proposed method exhibited very high performance with a 99.6% accuracy rate. Because high precision and a low false positive rate are important criteria for the language tutoring setting, the proposed method will be helpful for intelligent CALL systems.

Chapter V

GRAMMATICAL ERROR SIMULATION

5.1 Introduction

Computer-based methods for learning language skills and components have been used to address several advanced research topics: to generate corrective feedback in DB-CALL systems for developing oral skills, to simulate language learners to optimize tutoring strategies based on reinforcement learning, and to generate context-dependent grammar quizzes as educational game-play in language-learning games in virtual worlds. A primary purpose of this study is to investigate grammatical error simulation, which is a common component of these technologies. To provide further rationale for this study, brief discussions of the aforementioned technologies and the role of grammatical error simulation within them are presented in the following subsections.

5.1.1 Corrective feedback in DB-CALL systems

One of the ultimate goals of CALL is to provide learners with an environment that facilitates the acquisition of communicative competence, especially oral skills. As a result, the demand for CALL systems that help language learners develop oral skills has increased, and numerous CALL systems for pronunciation training have been developed to meet this demand [23, 24]. However, pronunciation is only one of the skills required for proficiency in speaking a second language; a learner must also learn other important aspects of the spoken language, such as morphology and syntax.

To account for this fact, DB-CALL systems, e.g., SPELL [27], DISCO [47], and CSIEC [55] have been developed to detect grammatical errors in speaking performance, provide learners with corrective feedback, and give learners an opportunity to try repeatedly until they manage to produce the correct form. The key element of such systems is the development of speech recognition grammars, which predict not only grammatical responses but also ungrammatical responses to deliver relevant and immediate feedback to the learner's utterance. In these systems, recognition grammars to anticipate the particular grammatical errors for each predicted response were usually written by hand. However, using human experts to anticipate various types of grammatical errors and list all possible realizations of the errors is too laborious and costly. Thus, automatic generation of realistic grammatical errors to create recognition grammars is crucial to the development of DB-CALL systems.

5.1.2 Language learner simulation to learn effective tutoring strategies

It is widely acknowledged that careless tutoring can result in embarrassment and demotivation of the language learner [56]. However, designing effective tutoring strategies from learners' data requires numerous language learners with a variety of cognitive and affective factors (e.g., proficiency level, age, gender, mood, and concentration) [57], which is usually very expensive and time consuming even with design support systems [58].

Recently, user simulation has become widely used in the development of spoken dialog systems to develop dialog strategies that use reinforcement learning [59, 60]. For educational purposes, Ai et al. [36] used user simulation to study how to manipulate the strength of a tutor's feedback to maximize a student's overall certainty in the entire dialog. In contrast to experiments with human subjects, user simulation generates a large corpus of user behaviors in a low-cost and time-efficient manner. These studies have developed essential components of user simulation (e.g., intention-level, utterance-level, and speech recognition-level simulation), and have not yet simulated grammatical errors because it has been assumed that the systems were to be used by native speakers, who normally produce few grammatical errors in their utterances. Because language learners naturally commit numerous grammatical errors, grammatical error simulation should be embedded into user simulation to find effective tutoring strategies and to evaluate DB-CALL systems.

5.1.3 Grammar quiz generation for language learning games

Conventional teaching–learning methods cannot provide persistent motivation for learners to achieve high proficiency in foreign language learning. To encourage learners to continue learning, educational systems should manage stress levels and provide motivation by establishing interesting game components, with achievable goals and attractive rewards, based on different students' proficiency levels. Multiple-choice questions about grammaticality could provide good game-play. While a conversation is going on, a grammar quiz related to the conversational situation could also be occurring (Fig. V.1). If students successfully pass the test, they get a reward such as game money that they can use to purchase game items, including fancy costumes and character skills (e.g., flying, healing). Grammatical error simulation allows educational systems to immediately generate grammar quizzes from situation-related sentences.

The remainder of this chapter is structured as follows: Section 5.2 briefly describes related studies. Section 5.3 presents a detailed description of the methods. The experimental setup follows in Section 5.4. Section 5.5 introduces the performance metric used in this study. Section 5.6 shows the results and discusses their meaning. Finally, Section 5.7 gives our conclusion.

5.2 Related Work

Grammatical error simulation is a relatively novel research area, and only a few reports in this area have been published. In Foster's [43] pioneering work, she described a procedure that automatically introduces frequently-occurring grammatical

Figure V.1: A snapshot of Pomy (A language learning game that is under development by the authors) with a grammar quiz

errors into sentences to generate ungrammatical training data for a robust parser. Wagner et al. [61, 62] used the same method to automatically generate test sets for the problem of classifying a sentence as grammatically well-formed or ill-formed.

However, there are several reasons that the algorithm cannot be directly applied to grammatical error generation for our purpose. First, it either introduces one error per sentence or none, regardless of how many words in the sentence are likely to generate errors. Second, it determines the type of error it will create using only the relative frequencies of error types and their relevant POSs. These constraints often make the approach generate unrealistic errors. For example, when the algorithm tries to create an error by deleting a word given the input sentence in Table V.1,it would probably omit the word 'go', because verbs are among the most frequently omitted POS. The result is an unrealistic error like the first simulated output, "He wants to

Table V.1: Examples of simulated outputs

Input sentence	**He wants to go to a movie theater**
Unrealistic simulated output	He wants to to a movie theater
Realistic simulated output	He want go to movie theater

to a movie theater''. Third, this algorithm does not consider the different interlingual errors made by learners with different native languages. Native Japanese speakers learning English, for example, tend to make errors with subject–verb agreement, omission of the preposition of prepositional verbs, and omission of articles because their first language does not have similar grammatical rules. Thus, these students often commit errors like the second simulated output, ''He want go to movie theater.''

In a later study [63], the method was improved to consider more contextual information, i.e., the POS of the words immediately to the left or right of the target word. However, the method still does not employ other contextual information such as parse trees, syntactic roles, semantic categories, or irregularities, which are essential to describing the error characteristics of language learners. This discrepancy ismainly due to the lack of a flexible language that is powerful enough to describe complicated dependencies in contextual information.

A major purpose of this investigation is to develop an effective approach to grammatical error simulation that generates realistic errors by encoding the complicated characteristics of intralingual and interlingual errors using Markov logic [50], which is a powerful and flexible approach to soft constraint satisfaction problems.

5.3 Methods

In this section, we give a brief introduction to Markov logic, which is a high-level language to describe soft constraint satisfaction problems and present an implementation of the Markov Logic Network (MLN) for grammatical error simulation.

5.3.1 Language for soft constraints satisfaction problems: Markov logic

Error characteristics can be taken as constraints that the grammatical error simulation method should satisfy when generating errors. However, errors do not occur every time the constraints are met, so we cast this problem as a soft (weighted) constraint satisfaction problem. The probability of an error occurring is represented as a weight attached to the constraints of the error.

Many real-world problems can be formalized as constraint satisfaction problems, so several high-level languages are available in the field of constraint programming, such as OPL [64],Oz [65], Eclipse [66], Zinc [67], and Minizinc [68]. However, these methods do not allow probabilistic inference.

In contrast, popular probabilistic techniques such as graphical models (e.g., Maximum Entropy and Bayesian Networks) are usually limited to expressing a simple binary property of an observation, because complex features [1] that express combinations of properties require a huge amount of manual work [69]. For example, a word starting with a capital letter (like the word Day) is more likely to be a proper noun (NNP) than a common noun (e.g., in the expression United Nations Day). However,

[1] In graphical models, features play the same role of constraints.

a word that is capitalized but occurs at the beginning of a sentence (the previous word is hsi [2]), as in "Day after day ...", is not more likely to be a proper noun. To capture this tendency, this complex feature in Eq. V.1 should be defined manually

$$f(c,x) = \begin{cases} 1, & if\ word_{i-1} = \langle s \rangle \wedge\ isupperfirst(word_i)\ \wedge c = NN, \\ 0, & otherwise \end{cases} \tag{V.1}$$

In addition, researchers must write specific handling to determine the boolean value of each complex feature because graphical models do not perform logical inference. This is a serious issue in that the amount of code to be written could grow without bound, as numerous logical formulas with various combinations of many logical operators would be required (e.g., and, or, not, imply, equal). Furthermore, because graphical models can only handle propositional universes, researchers should manually instantiate first-order predicates (e.g., instantiating isupperfirst(x) to isupperfirst('a'), isupperfirst('b'), etc.). It should also be noted that there are computational problems as well. In principle, probabilistic inference can be performed using belief propagation, Markov Chain Monte Carlo (MCMC), the variational approximation, and other methods, but these can be extremely slow in practice when the weights of complex features are large and can break down when the weights are infinite [50].

Thus, neither constraint programming nor graphical models can provide a proper solution for grammatical error simulation. However, in recent years, there have been increasingly frequent attempts to combine constraint satisfaction problems and probabilistic inference. One of the most expressive formalisms is Markov logic, which combines a first-order logic representation with the semantics of graphical models

[2] ⟨ s ⟩ is a symbol to represent beginning of sentences.

to define a family of probability distributions in relational domains. Markov logic can solve the problems mentioned above by providing automatic logical reasoning and transforming first-order clauses into propositions.Most toolkits for Markov logic support a development interface that is a sufficiently flexible high-level language to describe target problems [70, 71, 72]. To deal with computational issues, Poon and Domingos [73] introduced MC-SAT, which handles determinism and achieves very large decreases in execution time by combining MCMC with satisfiability testing. Therefore, we used the learning and inference algorithms provided in the Alchemy package, which is an efficient open-source software package for Markov logic.

5.3.2 MLN formulas for grammatical error simulation

An MLN can be seen as a first-order knowledge base with weights attached to each of the formulas. The MLN is a template for the construction of probabilistic graphical models, namely Markov random fields. A total of 119 MLN formulas were obtained by analyzing the NICT JLE corpus, which consists of three components: (1) basic formulas based on POS, which are comparable to the methods used in previous work; (2) analytic formulas drawn from expert knowledge obtained through error analysis on a learner corpus; and (3) error-limiting formulas that penalize a statistical model's over-generation of nonsense errors.

Basic formulas

Error patterns obtained by error analysis cannot explain every error that learners commit, as error patterns might capture a lack or an over-generalization of knowledge about a particular construction. Furthermore, an error can take the form of a performance slip, which can occur randomly due to carelessness or tiredness; therefore, more general formulas are needed as a default case. The basic formulas are represented by the following simple rules:

$$PosTag(s, i, +pt) \implies ErrorType(s, i, +et), \tag{V.2}$$

$$PosTag(s, i-1, +ppt) \wedge PosTag(s, i, +pt) \implies ErrorType(s, i, +et), \tag{V.3}$$

$$PosTag(s, i, +pt) \wedge PosTag(s, i+1, +npt) \implies ErrorType(s, i, +et), \tag{V.4}$$

in which all free variables are implicitly universally quantified. In Eq. V.2, the "$+pt, +et$" notation signifies that the MLN contains an instance of this rule for each (part of speech, error type) pair. The evidence predicate in this case is $PosTag(s, i, pt)$, which is true if and only if the ith position of the sentence s has the part of speech pt. The query predicate is $ErrorType(s, i, et)$, which is true if and only if the ith position of the sentence s has the error type et, and querying it returns the probability that the word at position i would generate an error of type et. In Eq. V.3, the evidence predicate represents the POS bigram consisting of the POSs of the previous and current words. In Eq. V.4, the evidence predicate represents the POS bigram consisting of the POSs of the current and next words.

Analytic formulas

In addition to the basic formulas, analytic formulas provide concrete knowledge about the real error made by language learners. Various error sources for each error type can be identified by inspecting the grammar of the target language and the linguistic differences between the native and target languages. We roughly categorize error sources into three groups: (1) over-generalization of the rules of the second language; (2) lack of knowledge of some rules of the second language; and (3) applying rules and forms of the native language to the second language.

English learners often commit pluralization errors with irregular nouns. These errors result because they over-generalize the pluralization rule, i.e., attaching 's/es' to the end of a singular noun, so that they apply the rule even to irregular nouns such as 'mice' and 'feet'. This characteristic is captured by the simple formula:

$$\begin{aligned} &IrregularPluralNoun(s,i) \wedge PosTag(s,i,NNS) \\ &\quad \Longrightarrow\; ErrorType(s,i,N\_NUM), \end{aligned} \tag{V.5}$$

where $IrregularPluralNoun(s,i)$ is true if and only if the ith word of the sentence s is an irregular plural, NNS stands for plural noun, and N_NUM is the abbreviation for noun number error.

One trivial error caused by a lack of knowledge of the second language is using a singular noun form for weekly events:

$$\begin{aligned} &Word(s,i-1,on) \wedge DayNoun(s,i) \wedge PosTag(s,i,NNS) \\ &\quad \Longrightarrow\; ErrorType(s,i,N\_NUM), \end{aligned} \tag{V.6}$$

where $Word(s,i-1,on)$ is true if and only if the $i-1$th word is 'on' and $DayNoun(s,i)$ is true if and only if the ith word of the sentence s is a noun describing a day, such as Sunday(s). Another example is use of plurals after 'every' due to ignorance that a noun modified by 'every' should be singular:

$$\begin{aligned} &Word(s, i_d, every) \wedge Departminer0f(s, i_d, i_n) \wedge PosTag(s, i_n, NNS) \\ &\quad \implies ErrorType(s, i_n, N\_NUM), \end{aligned} \tag{V.7}$$

where $Departminer0f(s, i_d, i_n)$ is true if and only if the i_dth word is the determiner of the i_nth word.

Japanese often allows omission of the subject of a sentence, and native Japanese speakers learning English often commit errors of subject omission, providing one example of errors that result from applying the rules of the first language to the second language. The following formula is for the case:

$$Subject(s, i) \wedge PosTag(s, i, PRP) \implies ErrorType(s, i, PRP\_LXC), \tag{V.8}$$

where $Subject(s, i)$ is true if and only if the ith word is the subject, PRP stands for pronoun, and PRP_LXC is the abbreviation for pronoun lexis error.

Error limiting formulas

A number of elementary formulas explicitly stated as hard formulas prevent the MLN from generating improbable errors that might result from over-generalizations of the statistical model. For example, a verb complement error should not assign a probability to words that are not complements of a verb:

$$!VerbComplements(s, i_v, i_c) \implies !ErrorType(s, i_v, V\_CMP)., \tag{V.9}$$

where "!" denotes logically 'not', and "." at the end signifies that this is a hard formula. Hard formulas are given maximum weight during inference. $VerbComplements(s, i_v, i_c)$ is true if and only if the i_cth word is a complement of the verb at the i_vth position, and V_CMP is the abbreviation for verb complement error.

5.4 Experimental Setup

In this section, we give a brief introduction to the NICT JLE corpus [74] from which we learned the weights of the constraints of the MLN for grammatical error simulation. In addition, we present how the baseline model and the proposed model were set up for comparative experiments. Finally, we describe the overall procedure of grammatical error simulation.

5.4.1 NICT JLE corpus

The learner corpus is a collection of spoken or written language data from non-native speakers. Several learner corpora [75, 74, 76] have been developed recently. Some of the existing learner corpora are annotated with errors from which interlingual and intralingual error characteristics for various native and learned languages can be extracted using computational techniques. The NICT JLE corpus is one of them. This is a two-million-word speech corpus of native Japanese speakers learning English. The corpus data were obtained from 1281 audio-recorded speech samples, 167 of which are error annotated, from an English oral proficiency interview test, i.e. the ACTFL-ALC Standard Speaking Test. The current version of the error tagset targets morphological, grammatical, and lexical errors and can describe

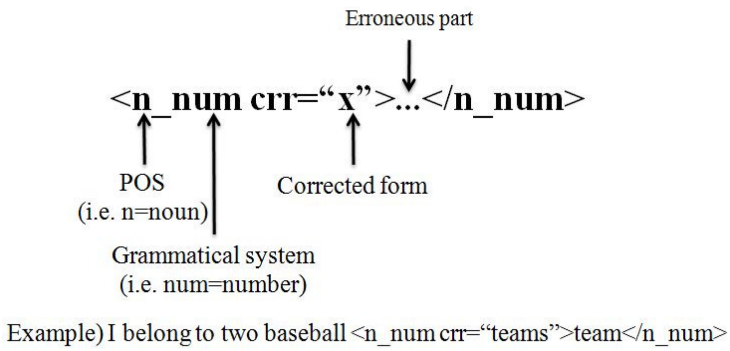

Figure V.2: Structure of an error tag and an example of an error-tagged sentence

diverse grammatical errors [51]. The error tags contain three pieces of information: POS, morphological/grammatical/lexical rules, and a corrected form (Fig. V.2). The error tagset currently includes 46 tags (Table V.2).

5.4.2 Data preparation and setup of grammatical error simulation models

To verify the quality of the simulated grammatical errors, the proposed grammatical error simulation method was compared against the real learners' errors and the baseline model using only the basic formulas of Section 5.3.2. This method is comparable to the method of Foster and Andersen [63]. According to Sumner and Domingos [77], the most commonly used probabilistic models can be succinctly formulated as MLNs, including the Maximum Entropy technique, Bayesian Networks, a Hidden Markov Model, Conditional Random Field, etc. In fact, the baseline model is equal to the Maximum Entropy approach, which is a representative graphical model. This implies that the baseline model is good enough to show the difference between the results of the Markov logic approach and the results of previous research based on

Table V.2: Error category

Tag	**Error category**
Noun	
<n_inf>...</n_inf>	Noun inflection
<n_num>...</n_num>	Noun number
<n_cs>...</n_cs>	Noun case
<n_cnt>...</n_cnt>	Countability of noun
<n_cmp>...</n_cmp>	Complement of noun
<n_lxc>...</n_lxc>	Lexis
Verb	
<v_inf>...</v_inf>	Verb inflection
<v_agr>...</v_agr>	Subject verb disagreement
<v_fml>...</v_fml>	Verb form
<v_tns>...</v_tns>	Verb tense
<v_asp>...</v_asp>	Verb aspect
<v_vo>...</v_vo>	Verb voice
<v_fin>...</v_fin>	Usage of finite/infinite verb
<v_ng>...</v_ng>	Verb negation
<v_qst>...</v_qst>	Question
<v_cmp>...</v_cmp>	Complement of verb
<v_lxc>...</v_lxc>	Lexis
Modal verb	
<mo_lxc>...</mo_lxc>	Lexis
Adjective	
<aj_inf>...</aj_inf>	Adjective inflection
<aj_us>...</aj_us>	Usage of positive/comparative/superlative of adjective
<aj_num>...</aj_num>	Adjective number
<aj_agr>...</aj_agr>	Number disagreement of adjective
<aj_qnt>...</aj_qnt>	Quantitative adjective
<aj_cmp>...</aj_cmp>	Complement of adjective
<aj_lxc>...</aj_lxc>	Lexis

Adverb	
<av_inf>...</av_inf>	Adverb inflection
<av_us>...</av_us>	Usage of positive/comparative/superlative of adverb
<av_pst>...</av_pst>	Adverb position
<av_lxc>...</av_lxc>	Lexis
Preposition	
<prp_cmp>...</prp_cmp>	Complement of preposition
<prp_lxc1>...</prp_lxc1>	Normal preposition
<prp_lxc2>...</prp_lxc2>	Dependent preposition
Article	
<at>...</at>	Article
Pronoun	
<pn_inf>...</pn_inf>	Pronoun inflection
<pn_agr>...</pn_agr>	Number/sex disagreement of pronoun
<pn_cs>...</pn_cs>	Pronoun case
<pn_lxc>...</pn_lxc>	Lexis
Conjunction	
<con_lxc>...</con_lxc>	Lexis
Relative pronoun	
<rel_cs>...</rel_cs>	Case of relative pronoun
<rel_lxc>...</rel_lxc>	Lexis
Interrogative	
<itr_lxc>...</itr_lxc>	Lexis
Others	
<o_je>...</o_je>	Japanese English
<o_lxc>...</o_lxc>	Collocation
<o_odr>...</o_odr>	Misordered words
<o_uk>...</o_uk>	Unknown type errors
<o_uit>...</o_uit>	Unintelligible utterance

other graphical models, which are usually limited to simple binary descriptions.

The advantage of using the Standard Speaking Test data as a source is that each speaker's data includes his or her proficiency level based on the Standard Speaking Test scoring method; this information facilitates analysis and comparison of interlanguage characteristics at each developmental stage. We exploited this property, accomplishing level-specific error simulation by dividing the 167 error annotated files into three groups: Beginner (levels 1–4), Intermediate (levels 5–6), and Advanced (levels 7–9).

Two sets of three different MLNs (one set for baseline models and one set for proposed models) were trained from these three level-specific data sets. To determine the truth value of the numerous predicates of the MLN formulas, a number of features were extracted: words, lemmas, POS, irregularity, the positions of words in sentences, grammatical relationships (e.g., a subject of a verb, an object of a verb, an object of a preposition), and semantic information (e.g., countability of noun, day noun, binary noun). The morphological analyzer of LanguageTool[3] was used to lemmatize words and synthesize the erroneous word forms in the final step of grammatical error simulation. The Stanford parser [78] was utilized to extract syntactic features. For semantic categories, we made corresponding lists of words using English dictionaries.

Some error categories were excluded. Lexis errors related to open-word class (i.e., n_lxc, v_lxc, aj_lxc, and av_lxc), were excluded in this experiment because realizing such errors without encountering the data sparseness problem requires a huge amount of learner data. Some other errors (i.e., o_je, o_lxc, o_odr, o_uk, and o_uit) were also excluded because these error categories have not yet been clearly ana-

[3] http://www.languagetool.org

lyzed for practical applications. Error categories that occurred less than five times were also excluded to improve reliability. In addition, we did not explicitly repair insertion errors because many insertion errors appear implicitly as replacement errors in the NICT JLE corpus. The insertion errors, which are not covered in this model, usually relate to vocabularies in open-word classes or are highly unpredictable even when linguistic context is taken into account.When multiple errors overlap in the same or nearly-same position, we took only the outermost error type. For example, in the case of "They are ⟨v_lxc crr = "watching"⟩looking⟨prp_lxc2 crr = "at"⟩⟨/prp_lxc2⟩⟨/v_lxc ⟩ monkeys", the error type v_lxc is taken.

5.4.3 Overall grammatical error simulation procedure

The purpose of grammatical error simulation is to generate an ill-formed sentence when given a well-formed input sentence. The generation procedure involves four steps: (1) generating probabilities of error types for each word in the well-formed input sentence through MLN inference; (2) determining an error type by sampling the generated probability for each word; and (3) determining how errors structurally deviate from correct usage (omission or replacement) by sampling according to the ratio of the number of omission errors to the number of replacement errors in the NICT JLE corpus; (4) an ill-formed output sentence is created by realizing the chosen error types (Fig V.3).

Phase	Error type	He	wants	to	go	to	a	movie	theater	
Inference	**v_agr**	0.000	**0.371**	0.000	0.000	0.000	0.000	0.000	0.000	*1 step*
	prp_lex	0.000	0.000	**0.284**	0.000	0.269	0.000	0.000	0.000	
	at	0.000	0.000	0.000	0.000	0.000	**0.355**	0. 000	0.000	
	...	...	...	...	...	...	...	...	...	
	none	0.921	0.449	0.604	0.866	0.605	0.506	0.781	0.798	*2 step*
Sampling		none	v_agr	prp_lex	none	none	at	none	none	*3 step*
Structual Deviation			replace	omit			omit			*4 step*
Realization		He	***want***		go	to		movie	theater	

Figure V.3: An example grammatical error simulation process

5.5 Performance Metric

Evaluating the quality of grammatical error simulation helps users judge whether the simulated errors are useful for educational applications. This research employs several automatic evaluation metrics and human evaluation to measure the quality of grammatical error simulation.

5.5.1 Precision and recall rate

There are no generally accepted criteria regarding what constitutes a good grammatical error simulation. A good learner model should be able to generate learner-like behavior and errors. Precision and recall rate are common measures of quality in user modeling [79] and were first used to evaluate user simulationmodels of dialog systems by Schatzmann et al. [60]. Precision is a measure of the proportion of correct errors to all of the predicted errors. Recall measures how many of the errors in the real responses are predicted correctly. It is not possible to specify what levels of the precision and recall rates are necessary to claim that simulated errors are realistic. Nevertheless, the precision and recall rates offer reliable methods to compare simulated and real user responses. For the experiments, we divided the NICT JLE corpus into training and test sets and used 10-fold cross-validations for each group.

5.5.2 Comparison of error distributions

Previous studies of CALL systems have found that the absolute and relative error frequencies can be useful for selecting pronunciation and syntactic errors and developing a CALL system [80, 81]. Therefore, it is important to develop a grammatical error simulation method that can generate errors with realistic frequencies to evaluate such CALL systems and to produce development data for them.

To verify the quality of the error distributions of the simulated grammatical errors, the proposed grammatical error simulation method was compared with the real learners' errors and the baseline model, with 10-fold cross-validations performed for each group. The validation results were added across the rounds to compare the number of simulated errors with the number of real errors. The Kullback-Leibler (KL) divergences, which measure the difference between two probability distributions, were also measured to provide a quantitative comparison. A KL divergence value close to zero indicates that the distribution of simulated grammars is similar to that of real human errors.

5.5.3 Human evaluation on quality of error generation

The overall quality of the simulated errors was assessed by experienced researchers in English as a foreign language. Ninety sentences (30 sentences for each proficiency level) were randomly selected from the erroneous sentences in the NICT JLE corpus. For each erroneous sentence, we took the corresponding correct sentence from the corpus and then generated two sets of three sentences, one set using the baseline model and one set using the proposed model. The evaluators were given a correct sentence and a set of seven corresponding erroneous sentences (one real error and six simulated errors) for each of the 90 sentences. They were required to accomplish two tasks: (1) identify the sentence produced by a human, (2) give a score on the five-point Likert scale, which uses a scale of 1–5 to indicate "strongly disagree" to "strongly agree" with the statement "The error sentence seems to be generated by

Table V.3: An example of evaluation items.

Human	**Score**	**Sometimes, I feel like hitting them**
		Sometimes, I feel like them hitting
		Sometimes, I felt hitting them
		Sometimes, I feel like hit them
		Sometimes, I feel hitting them
		Sometimes, I feels like hitting them
		Sometimes, I feel hitting them
		Sometimes, I feel like hitting

real language learners, not randomly generated by machine" (Table V.3). Descriptive statistics, analysis of variance (ANOVA), and multiple post-hoc comparisons were used to find the effects of different error sources on the human evaluation. All analyses were conducted using SAS, version 9.1.3.

5.5.4 Effects on labor-saving

The amount of labor-savings, i.e., the predicted time and cost associated with this method compared with othermethods of error generation, could be of particular interest to those who plan to exploit grammatical error simulation for various applications. Because no standard method to evaluate the cost of error generation exists, the labor-saving benefits of grammatical error simulation was validated by using the elapsed time necessary to generate erroneous sentences. The relative improvement was measured by calculating the time performance gain rate $TPGR = \frac{T_N - T_G}{T_N}$,where T_N is the elapsed time when not using the simulation, and T_G is the elapsed time when using it.

We asked four researchers of English as a foreign language to generate three erroneous sentences for each correct sentence and to annotate what errors they made in two different situations: using grammatical error simulation and not using grammatical error simulation. When not using grammatical error simulation, the evaluators created errors through their own effort. When using grammatical error simulation, they chose one of the simulated sentences or created erroneous sentences themselves only when they could not choose one of the simulated sentences. To remove any learning effects, two sets of dialogs (with two dialogs in each set) regarding shopping in a grocery store were used: two of the evaluators worked on the first set of dialogs without grammatical error simulation and then on the second set with grammatical error simulation, while another two evaluators worked on the second set without grammatical error simulation and the first set with grammatical error simulation.

5.6 Results and discussion

5.6.1 Precision and recall rate

To assess whether grammatical error simulation models produce learner-like errors, we need to compare their output with real responses in the same contexts. For this purpose, we used the precision and recall metrics. The precision and recall rates of the proposed model were higher than those of the baseline model at all learner levels (Figs. V.4, V.5, V.6). The proposed method increased the precision by 6% and the recall by 8.33% averaged across all proficiency levels.

5.6.2 Comparison of error distributions

Quality in the error distributions of the simulation was examined by comparing the number of simulated errors with the number of real learners' errors. The distribution

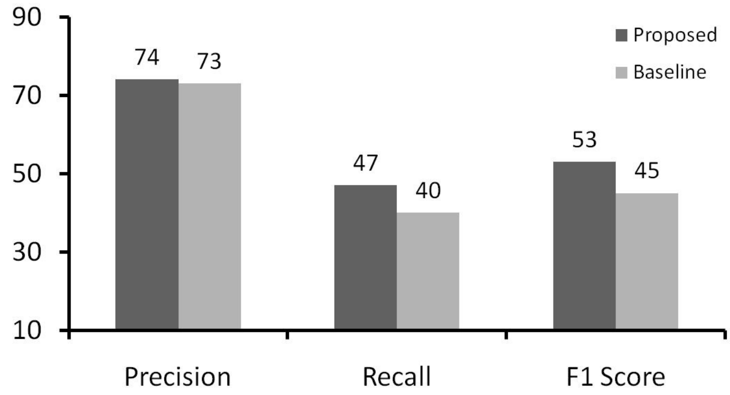

Figure V.4: Comparison between the precision and recall rates of the proposed and baseline methods for the beginner level

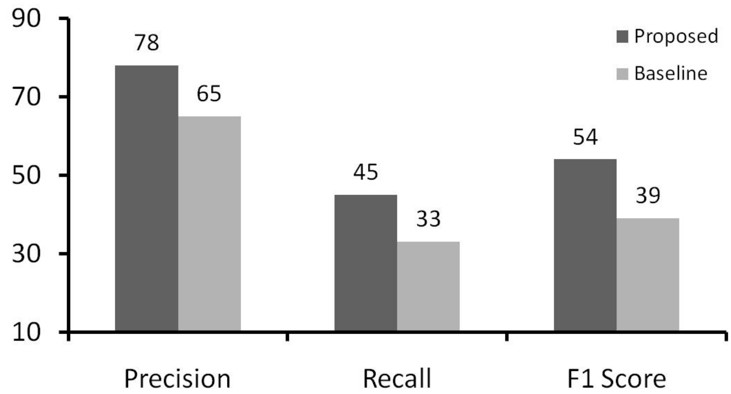

Figure V.5: Comparison between the precision and recall rates of the proposed and baseline methods for the intermediate level

of the simulated grammatical errors generated by the proposed model was similar to that of real learners at all learner levels (Figs. V.7, V.8, V.9). These findings were further quantitatively confirmed by the KL divergence.

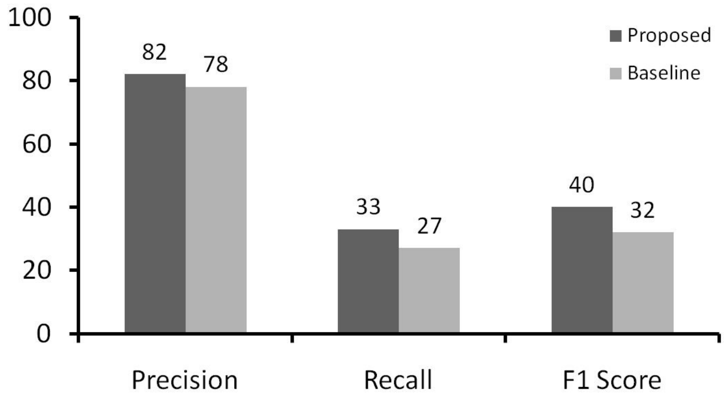

Figure V.6: Comparison between the precision and recall rates of the proposed and baseline methods for the advanced level

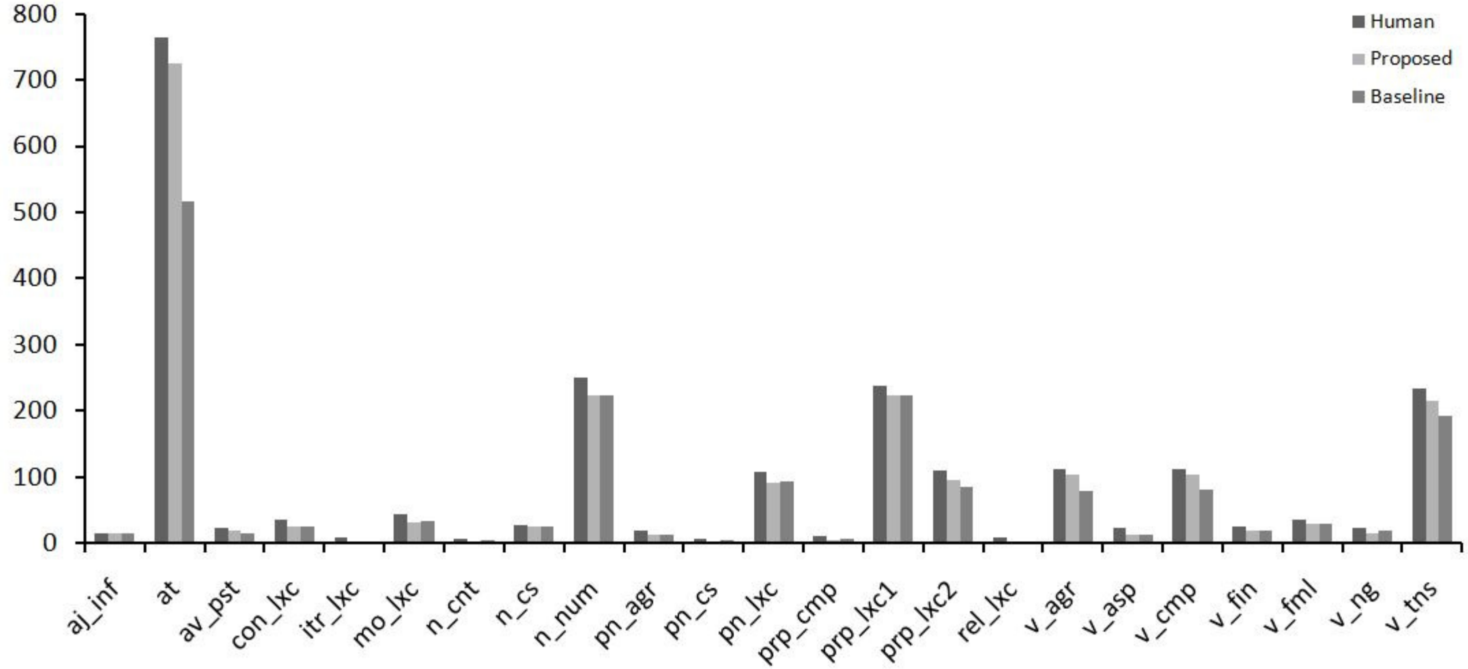

Figure V.7: Comparison between the distributions generated by humans, the proposed method, and the baseline method at the beginner level

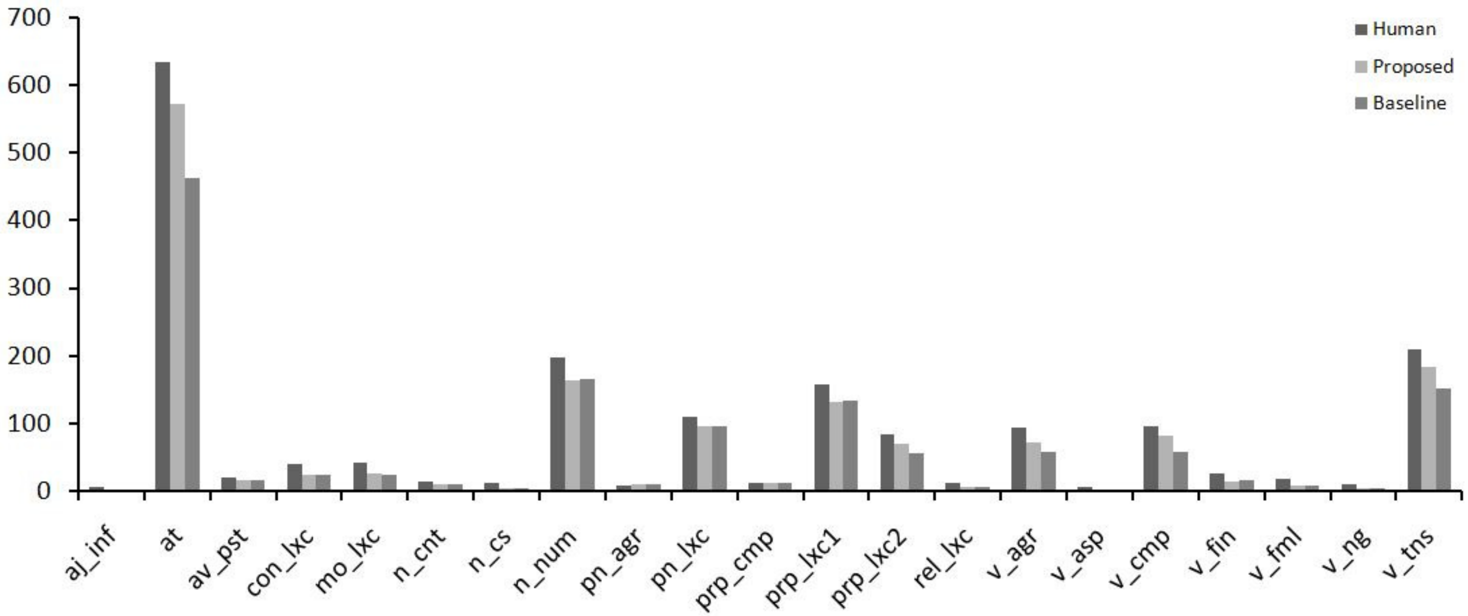

Figure V.8: Comparison between the distributions generated by humans, the proposed method, and the baseline method at the intermediate level

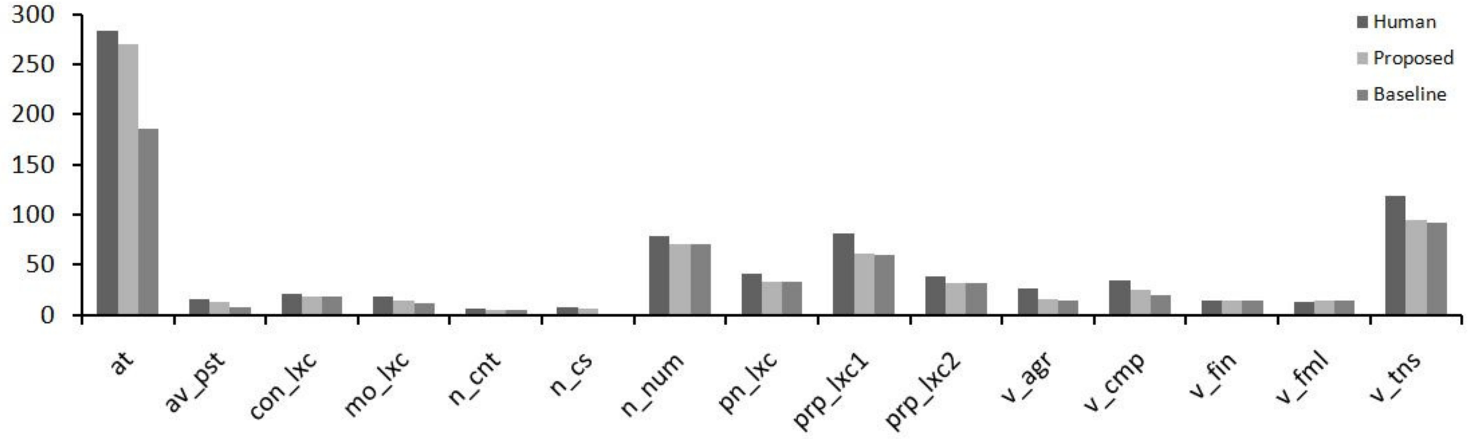

Figure V.9: Comparison between the distributions generated by humans, the proposed method, and the baseline method at the advanced level

Table V.4: KL divergences of the error distributions of the proposed and baseline models from the human error distribution

Proficiency level	**Model**	
	Baseline	**Proposed**
Beginner	0.012	0.010
Intermediate	0.015	0.013
Advanced	0.022	0.008

For all learner levels, the KL divergence of the error distribution of the proposed model from the human error distribution is less than that of the baseline model (Table V.4). The proposed method produced a relative improvement in the KL divergence of 37.5% averaged across all proficiency levels.

In most error categories, the number of simulated errors was smaller than the number of real learners' errors. This difference might be caused by the more constraints brought into the generation procedure. Another reason might be that the grammatical error simulation procedure sometimes tried to produce unrealizable errors in the given context, which result from several factors including insufficient contextual information, POS tagging errors, and parsing errors. The lack of contextual information is a more severe problem for the baseline model than the other issues in that it considers only the POS of the words immediately to the left or right of the target word. Therefore, the proposed model is more realistic than the baseline model when generating errors that are inherently sensitive to their contexts (e.g., v_agr, v_cmp, v_tns, prp_lxc2) or errors for which the POS coverage is not well matched to the coverage of the error category (e.g., DT - POS vs. at - error category).

Table V.5: The number of times chosen as human

	Human	**Proposed**	**Baseline**
Count	86	102	82
Weighted count	86	34	27

5.6.3 Human evaluation on quality of error generation

The overall quality of the simulation was assessed by experienced researchers in English as a foreign language. In the task of identifying the sentence produced by a human, the total counts for the proposed and baseline models were divided by three (weighted) for fair comparison, as there were three times more simulated sentences than human-generated sentences. The human-generated sentences obtained the highest weighted count, the proposed model received the second highest, and the baseline model had the lowest weighted count (Table V.5). These results indicate that the proposed model is more likely to produce realistic errors than the baseline model.

In the task of scoring the error on a five-point Likert scale, the human-generated errors also recorded the highest score, and the proposed model obtained a higher score than the baseline model. This finding is consistent with the previous result. The descriptive statistics and ANOVA results indicate that significant differences exist between the error sources (Table V.6). To identify the differences in detail, a Scheffé post-hoc analysis was performed. Significant mean differences were found between all pairs of models (Table V.7).

Table V.6: Mean, standard deviation, and ANOVA results of human evaluation

Model	N	Mean	SD	F	Pr > F
Human	270	4.644	0.920	39.95	<0.0001
Proposed	810	4.285	1.245		
Baseline	810	3.874	1.524		

Table V.7: Results of Scheffé post-hoc analysis

Model	Human	Proposed	Baseline
Human	–	0.359[a]	0.770[a]
Proposed	-0.359[a]	–	0.411[a]
Baseline	-0.770[a]	-0.411[a]	–

[a] Significant difference ($p < 0.05$)

Table V.8: Time performance gain rates

	Not using (min)	Using (min)	TPGR (%)
Dialog1	63	29	54
Dialog2	60	26	57
Dialog3	59	23	61
Dialog4	45	15	67
Total	227	93	59

5.6.4 Effects on labor-saving

The labor-saving benefit of grammatical error simulation was measured by calculating the TPGR. In all dialogs, the elapsed time was reduced considerably when the users exploited grammatical error simulation (Table V.8), which means that the construction of erroneous sentences with the help of grammatical error simulation is much more efficient than manual production. None of the evaluators opted to create errors manually when given the results of grammatical error simulation. They revised only several surface forms, which were incorrectly synthesized by the morphological synthesizer.

Despite the popularity of grammar training in foreign language learning, manual construction of grammar exercises remains a time-consuming and labor-intensive task. The experimental results show that the elapsed time required to generate grammatical errors was reduced considerably when the users employed grammatical error simulation, which means that the construction of erroneous sentences with the help of grammatical error simulation is much more efficient than manual production. We believe that this is a crucial advantage of grammatical error simulation. This has direct financial advantages because the time and cost of developing erroneous sentences would be dramatically reduced.

5.7 Conclusion

This chapter has introduced an automatic method for simulating realistic grammatical errors. Whereas previous studies did not consider sophisticated contextual information due to the lack of a flexible language, we could import the characteristics of intralingual errors of English learners and the interlingual errors of native Japanese speakers into statistical models by using Markov logic. The proposed method outperformed the baseline method, increasing the precision by 6% and the recall by 8.33% averaged across all proficiency levels. The proposed method also generated an error distribution that is more similar than that generated by the baseline method to the error distribution produced by real language learners. The proposed method produced a relative improvement of 37.5% in the average KL divergence. In accordance with acoustic scores, the enhanced quality of the simulated error distribution might increase an automatic speech recognizer's ability to detect grammatical errors by weighting errors that are generated more frequently more heavily. In addition, as the simulated errors become more similar to real learners' errors, the tutoring strategies of DB-CALL systems learned using language learner simulators would become more realistic and effective for real learners. Human evaluators judged that the proposed model is more likely to produce realistic errors than the baseline model. The proposed mode increased the average scores in two different evaluation tasks by 7 and by 0.411. In addition, the time performance gain rate showed that the construction of erroneous sentences with the help of grammatical error simulation is much more efficient than purely manual production. Using the proposed method reduced the grammatical error generation time by 59% in average.

During the experiments, we identified the need for a post-editing tool to ease the task of revising improper surface forms. The post-editing tool could also allow tutors to semi-automatically generate grammar quizzes from simulated errors if tutors want to focus on some specific grammar structures. Therefore, the development of the post-editing tool can be our future work. We also plan to add more expert knowledge through further error analysis to incrementally improve the performance. Because acquiring grammar is very important in learning a foreign language, we believe that the proposed method can be useful for developing intelligent language tutoring systems and generating materials for grammar training.

Chapter VI

FIELD STUDY

6.1 Introduction

There have been few serious attempts to verify educational effectiveness of DB-CALL systems that embody most of the aforementioned attributes. Therefore, we have provided an opportunity to learn English in an immersive environment in which learners experience free conversations about everyday life in real situations with intelligent robots. They can perceive the utterances of learners, especially Korean learners of English, and can provide corrective feedback to erroneous utterances. Recent development of robot-related technologies has drawn attention to the utilization of robots in real life, and increased interest in robots can give students integrative motivation to have a successful conversation with a robot. A major purpose of this investigation is to estimate the magnitude of the contributions that DB-CALL makes to the achievement of oral skills in the foreign language.

The remainder of this chapter is structured as follows: Section 6.2 describes related studies; Section 6.3 introduces the technologies for Human Robot Interaction (HRI); Section 6.4 presents a detailed description of the experimental design; Section 6.5 includes the results and discussion, and finally, Section 6.6 gives our conclusion.

6.2 Related Work

In recent years, there has been a shift in CALL research towards conversational interaction. This trend has been motivated by rapid globalization and great emphasis on communicative competence in the target language in a variety of situations. Recent development of spoken dialog systems has enabled CALL systems to bear a closer resemblance to oral conversation than the earlier CALL applications. There have been several systems that allow the user to engage in some form of meaningful dialog with embodied or disembodied agents in virtual words. DEAL [25] is a spoken dialog system for providing a multidisciplinary research platform, particularly in the areas of human-like utterance generation, game dialogue, and language learning. The domain is the trade domain, specifically a flea market situation. DEAL provides hints about things the user might try to say if he or she is having difficulties remembering the names of things, or if the conversation has stalled for other reasons. SPELL [27] provides opportunities for learning languages in functional situations such as going to a restaurant, expressing (dis-)likes, etc. Recast feedback is provided if the learner's response is semantically correct but has some grammatical errors. SCILL [26] covers the topics of weather information and hotel booking. Researchers also implemented the simulated user to produce example dialogs to expose language learners to language use and to expand the training corpus for the system. Let's Go [28] is a spoken dialog system that provides a bus schedule for the area around Pittsburgh, PA, USA. The researchers modified an extant system for the native speaker to adapt non-native speakers' data for the use of language learning. Modifications include the addition of new words, new constructs and the relaxation of some syntactic constraints to accept ungrammatical sentences.

Within the DB-CALL literature, however, there has been a dearth of empirical research on the developmental benefits engendered by the task environments in this line of research. Most discussions of the publications are largely system descriptions. Generally, it has been difficult to reconcile CALL research with SLA research due to contextual differences between computer-centered and human-centered tasks. Thus, making task conditions comparable to tasks performed under more traditional language learning conditions is one of the important challenges for CALL research. Unlike DB-CALL systems based on virtual worlds, robots as conversational agents bear a closer resemblance to human-centered tasks than DB-CALL applications because the only difference is the replacement of humans with robots in real life situations.

In Japan, the educational use of robots has been studied, mostly with Robovie [29] in elementary schools, focusing on English language learning. Robovie has one hundred behaviors. Seventy of them are interactive behaviors such as hugging, shaking hands, playing paper–scissors–rock, exercising, greeting, kissing, singing, briefly conversing. For the purpose of English education in this study, the robot could only speak and recognize English. In total, the robot could utter more than 300 sentences and recognize about 50 words. To identify the effects of a robot in English language learning, the researchers placed a robot in the first grade and sixth grade classrooms of an elementary school for two weeks, and compared the frequency of students' interaction with their English test score. While the interaction between the children and the robots generally diminished in the second week, a few children sustained a relationship with the robot. The results showed that the amount of time children spent with the robot during the first week had no effect on their improvement in English by the second week, but the amount of time that children interacted with the robots during the second week did have a significant and positive impact on improvement in English in the second week. This implies that robots which can maintain long-term relationships with students can be effective for language learning. Yet, Robovie has tended to be extremely restrictive in the number of words it can recognize so that the conversations have been confined to a chain of short-time interactions.

IROBI [30] was recently introduced by Yujin Robotics in Korea. IROBI was specifically designed and trialled for tutoring and educational services. IROBI, which has a sitting child-like appearance, is designed with an LCD panel on its chest to support easy communication with children, allowing voice and touch screen input without face and gesture recognition. IROBI was used to compare the effects of non-computer-based media and web-based instruction with the effects of robot-assisted learning for children. Robot-assisted learning is thought to improve children's concentration, interest, and academic achievement. It is also thought to be more user-friendly than other types of instructional media. But the discourse context of IROBI is slightly different from DB-CALL applications in that language learners interact with the robot largely though the virtual agents displayed in the LCD panel. The physical actions of the robot are merely employed to magnify the expressive power of content displayed in the LCD panel.

To the best of our knowledge, there have not been approaches combining authentic situations in the real world and real robots, which can provide a more realistic and active context than other approaches. Specifically, Engkey, the robot we developed, acts as a sales clerk in a fruit and vegetable store, and in a stationery store, so that it can interact in real life situations with language learners who play the part of customers. Given that studies on DB-CALL are still relatively new and most are in the early stages, this study aims to find general and approximate effects of DB-CALL which can motivate subsequent in-depth research. There is a need for much more research into the use of robots for educational purposes, and the effects of their use in this field.

The following section gives an account of the Human Robot Interaction (HRI) technologies used in the project.

6.3 Human Robot Interaction (HRI) Technology

We developed our robots as educational assistants called Mero and Engkey. They were designed with expressive faces, and have typical face recognition and speech functions allowing them to communicate. Mero is a head-only robot. The penguin-like robot Engkey is 80 cm tall and weighs 90 kg, and is equipped with stereo vision. In recent robotics research, several pioneering studies have suggested that humans can also establish relationships with pet robots. Many people actively interact with animal-like pet robots [82, 83, 84]

6.3.1 Speech and language processing

This section describes the speech and language processing component of the robots. At the high level, the speech and language processing component consists of a series of sub-components connected in a classical, pipeline architecture (see Fig. VI.1). The audio signal for the user utterance is captured and passed through a speech recognition module that produces a recognition hypothesis (e.g. "apple"). The recognition hypothesis is then forwarded to a language understanding component that creates a corresponding semantic representation (e.g. [item-apple]). Next, the dialog manager integrates this semantic input into the current discourse context, and produces the next system action in the form of a semantic output (e.g. {request quantity}). A language generation module produces the corresponding surface form, which is subsequently passed to a speech synthesis module and rendered as audio output.

Automatic speech recognition

The goal of ASR is to map from an acoustic signal to a string of words. Modern general purpose speech recognition systems are based on Hidden Markov Models (HMM). They take an acoustic model (AM), a dictionary of word pronunciations, and a language model (LM) to output the most likely sequence of words. The AM computes the likelihood of the observed acoustic signal given linguistic units such as phones or subparts of words for each time frame. The dictionary is a list of word pronunciations, each pronunciation represented by a string of phones. The language model (generally an n-gram[1] grammar) expresses the probability that a given string

[1]An n-gram is an n-token sequence of words: a bigram is a two-word sequence of words like 'Here is', 'is twenty', 'twenty five', or 'five dollars' and a trigram is a three-word sequence of words like 'Here is twenty', 'is twenty five', or 'twenty five dollars'.

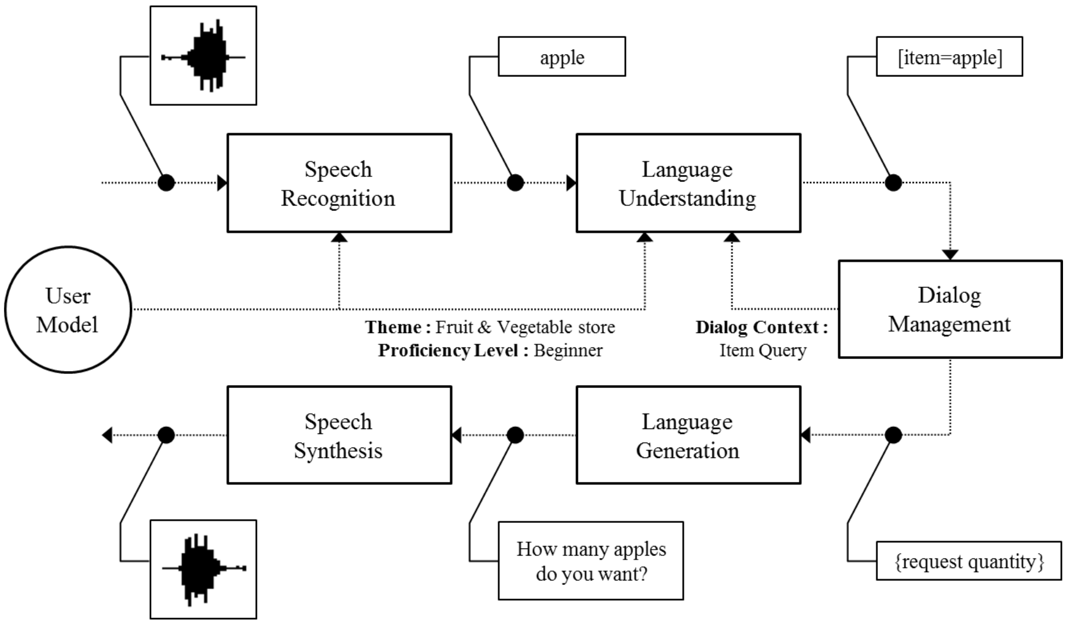

Figure VI.1: The architecture of the speech and language processing component

of words is a sentence in English.

In this study, speech recognition is performed by the DARE recognizer [85], a HMM-based speaker independent continuous speech recognizer. The target speech to recognize is the conversational speech of Korean elementary school students for shopping situations in a fruit and vegetable store, and a stationery store. Generally, speech is easier to recognize if the speaker is speaking a standard dialect, thus recognition is harder on foreign accented speech. Besides, it is rarely practical to collect enough training data to build an AM for a particular user group because it requires ASR experts to design a significant number of phonetically rich texts (hundreds of thousands of sentences), to record hundreds of hours of audio files with equal numbers of male and female speakers carefully chosen for diversity of voice quality and dialect. Therefore most previous studies on non-native speech recognition employed acoustic model adaptation techniques which improve the recognition performance with a small amount of non-native data [86]. We used a small amount of Korean children's transcribed speech (17 hours) to adapt acoustic models that were originally trained on the Wall Street Journal corpus [87] using standard adaptation techniques, both of maximum likelihood linear regression (MLLR) [88] and maximum a posteriori (MAP) adaptation [89]. The Korean children's speech was collected by a hundred Korean elementary school students (equal numbers of female and male students) using educational materials for the shopping domain (Section 6.4.2) which include small talk, purchases, exchanges and refunds.

ASR systems usually expect words to be pronounced in a certain way. If they are pronounced differently, which happens frequently in non-native speech, the automatic system is incapable of relating the 'wrong' pronunciation to the right word. Solely applying speaker adaptation techniques is therefore not sufficient to achieve a satisfactory performance for non-native speakers, so an additional modification of the pronunciation dictionary is necessary. We detected the occurrence of pronunciation variants with a speech recognizer in forced-alignment using a lexicon expanded according to all the possible substitutions between confusable phonemes. Korean speakers tend to replace the following consonants with the correspondingly similar consonants; the eight pronunciation variants of vowels shown in Table VI.1 are common to Korean speakers.

While large-vocabulary ASR systems focus on transcribing any sentence on any topic, for domain-dependent dialog systems it is of little use to be able to transcribe such a wide variety of sentences. The sentences that the speech recognizer needs to be able to transcribe are just those that are related to an ongoing dialog context.

Table VI.1: List of possible substitutions

Consonant	Vowel
CH → T	IH → IY
DH → D	OY → IY
TH → T	ER → R
TH → S	UH → OW
ZH → JH	EH → AE
F → P	AA → AO
R → L	AO → OW
V→ B	AH → AA

We call such a dialog-state dependent LM a restrictive LM. When we require the system to improve recognition accuracy, we can use a restrictive LM, thus achieving better accuracy at the cost of input diversity. We made different LMs around combinations of the study theme (small talk, fruit and vegetable store, and stationery store) and the student's English proficiency level (beginner and intermediate). The speech recognizer loads and unloads LMs dynamically according to the student's English proficiency level and the study theme. The student's level is indicated by the radio frequency ID (RFID) person identification process when every student starts a learning session by scanning their RFID card (see Section 6.3.3). The study theme is updated by the dialog manager which tracks dialog states during a conversation (see Section 6.3.1).

The standard evaluation metric for speech recognition systems is word error rate (WER). When given a pair of the reference sentence (supposedly the correct one) and the recognized one, WER can be computed as: $WER = (S + D + I)/N$, where S is the number of substitutions, D deletions, I insertions, and N is the number of words in the reference. By virtue of adaptation of acoustic model and pronunciation dictionary, and use of restrictive grammars, the average WER was about 22.8% at the vocabulary size of 1250.

Spoken language understanding

The Spoken Language Understanding (SLU) component of dialog systems must produce a semantic representation that is appropriate for the dialog task. Many speech-based dialog systems, since as far back as the GUS system [90], are based on the frame-and-slot semantics. A shopping task would have a frame with slots for information about items and price, thus a sentence like "Here is twenty five dollars" might correspond to the following filled-out frame:

```
Intention2:
    Speech Act: declare
    Main Goal: payment
Additional Information:
    Num: twenty five
    Unit: dollar
```

To generate this semantic representation, some dialog systems use general-purpose unification grammars with semantic attachments [91]. Other dialog systems rely on simpler domain specific semantic analyzers, such as semantic grammars [92]. Since language learners commit numerous and diverse errors, CALL systems should be able to understand language learners' utterances in spite of these obstacles. To accomplish this purpose, rule-based systems (i.e., general-purpose unification grammars and se-

[2] In the robots that we developed, we represent an intention in the form of 'Speech act(Main goal)'.

mantic grammars) usually anticipate error types and hand-craft a large number of error rules, but this approach makes these methods weak in dealing with ambiguity and insensitive to unexpected errors and diverse error combinations [27, 28, 33]. An alternative to rule-based systems that is probabilistic and also avoids hand-coding of grammars is machine learning-based techniques such as classification models and sequence labeling models. The task of classification is to take an utterance, extract some useful features describing the observation (e.g., bag of words, bag of n-grams), and then, based on these features, to classify the observation to one of a set of discrete classes (e.g., one of user's intentions).

There are often many ambiguities in interpreting a user's intention. For example, the following utterance looks like a yes-no question.

```
Can you give me a list of healthy foods?
```

In fact, however, this person was not interested in whether the system was capable of giving a list; this utterance was a polite form of a request. To resolve these ambiguities we need not only features from utterance itself but also features based on conversational context. In addition, the learners' numerous and diverse errors can make the classification of user's intention even harder, so systems should rely more on conversational context, as human tutors do.

Therefore we statistically infer the actual learner's intention by taking into consideration not only the utterance itself but also the dialog context. We can achieve this goal by employing features from dialog context and utterances together to make a classification model for intention recognition. For CALL, however, such approaches

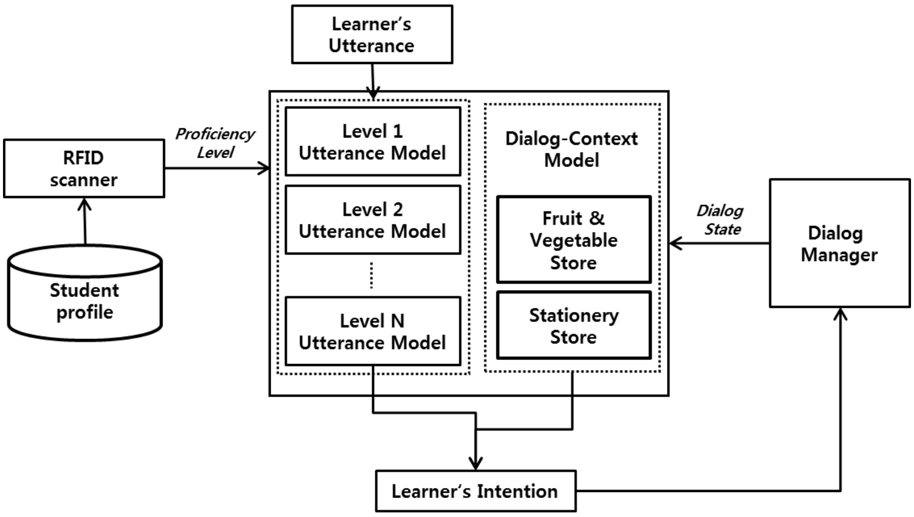

Figure VI.2: Hybrid model of language understanding

can be problematic, because separate handling for each of the proficiency levels is important in a language learning setting. Given a dialog scenario, the dialog-context model is relatively invariant; thus we prefer a hybrid model that combines the utterance model and the dialog-context model in a factored form, as shown in Fig. VI.2. This approach allows us to adjust the hybrid model to a required proficiency level by replacing only the utterance model [93].

The hybrid model merges n-best hypotheses[3] from the utterance model with n-best hypotheses from the dialog-context model to find the best user's intention. In the language production process, user intentions are first derived from the dialog context; subsequently the user intentions determine utterances [37]. By using this dependency and the chain rule, the most likely expected user's intention $I(U, D)$

[3]Instead of just producing the single best hypothesis, we produce a ranked list of hypotheses together with their probabilities. We call this ranked list of N hypotheses the n-best hypotheses.

given the utterance U and the dialog context D can be stated as follows:

$$I(U,D) = argmax_I P(I|U,D) \tag{VI.1}$$

$$I(U,D) = argmax_I \frac{P(I,U,D)}{P(U,D)} \tag{VI.2}$$

$$I(U,D) = argmax_I \frac{P(U|I)P(I|D)P(D)}{P(U,D)} \tag{VI.3}$$

By using Bayes' rule, Eq. VI.3 can be reformulated as:

$$I(U,D) = argmax_I \frac{P(U)P(U|I)P(I|D)P(D)}{P(U,D)P(I)} \tag{VI.4}$$

$P(U)$, $P(D)$, and $P(U,D)$ can be ignored, because they are constant for all I (Eq. VI.5):

$$I(U,D) = argmax_I \frac{P(I|U)P(I|D)}{P(I)} \tag{VI.5}$$

In this formula, $P(I|U)$ represents the utterance model and $P(I|D)$ represents the dialog-context model.

In order to distinguish the user's intention from the utterance itself, we use maximum entropy model [38] trained on linguistically motivated features. The objective of this modeling is to find the I that maximizes the conditional probability, $P(I|U)$ in Eq. VI.5, which is estimated using Eq. VI.6:

$$P(I|U) = \frac{1}{Z} exp\left(\sum_{k=1}^{K} \lambda_k f_k(I,U)\right), \tag{VI.6}$$

where K is the number of features, f_k denotes the features, λ_k the weighted parameters for features, and Z is a normalization factor. This model offers a clean way to combine diverse pieces of linguistic information. We used the following linguistic features for the utterance model:

- **Lexical word features:** Lexical word features consist of lexical trigrams using current, previous, and next lexical words. They are important features, but the lexical words appearing in training data are limited, so data sparseness problems can arise.

- **Part-of-speech (POS) tag features:** POS tag features also include POS tag trigrams matching the lexical features. POS tag features provide generalization power over the lexical features.

Determining the user's intention from the dialog state can be solved by finding similar dialog states within a dialog-state space (see Fig. VI.3), which was inspired by example-based dialog modeling [40]. Each dialog segment is represented as one dialog state (Table VI.2). A dialog-state space is built by first collecting a dialog corpus. Semantic tags (e.g., speech act, main goal, and additional information) are then manually annotated to utterances. A hand-crafted automatic system is also used to extract discourse contextual features (e.g., previous intentions and exchanged information status) by keeping track of the dialog states for each point in the dialog. Then the possible user intentions can be selected from dialog states similar to the current dialog state. The best user's intention is obtained from the dialog state that maximizes the similarity.

This idea can be formulated as the k-nearest neighbors (KNN) problem [39](Dasarathy, 1990) which provides high controllability for incrementally tuning the model dur-

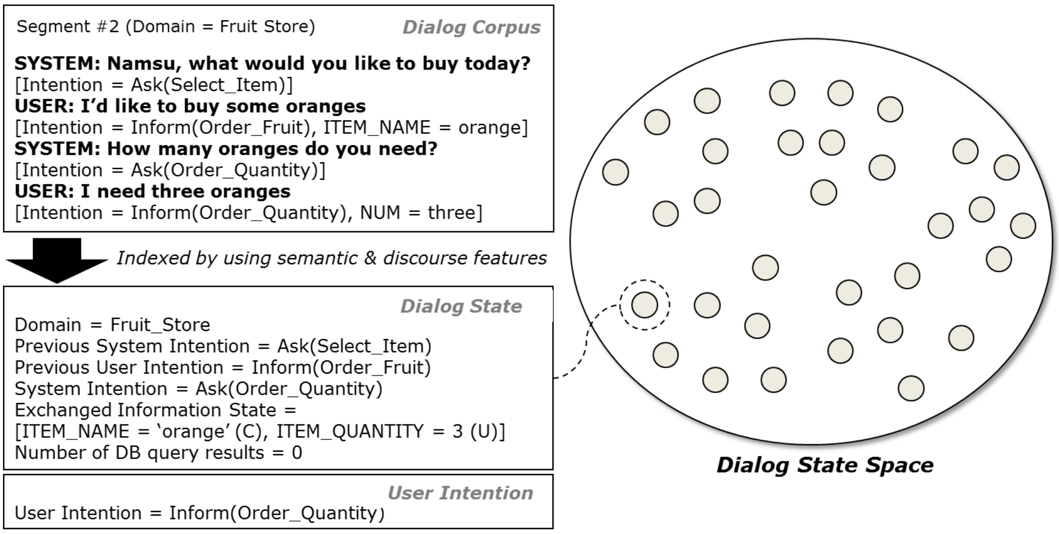

Figure VI.3: Indexing scheme for building a dialog-state space for the shopping domain

Table VI.2: Representation of dialog context and an example for the shopping domain

Attributes	Detail descriptions
PREV_SYS_INT	Intention of the previous system's intention
PREV_USR_INT	Intention of the previous user's intention
SYS_INT	Intention of the current system's intention
INFO_EX_STAT	A list of exchanged information states which is essential to successful task completion; (c) denotes confirmed, (u) unconfirmed
DB_RES_NUM	Number of database query results

ing operation, which is in practical terms a very desirable property. The similarity function is defined as the following equation:

$$Similarity(D, D') = \sum_{k=1}^{K} \lambda_k f_k(D, D'), \tag{VI.7}$$

where D and D' are dialog states, K is the number of features, f_k denotes the feature functions, λ_k the weighted parameters for features. Our feature functions first include the simplest tests, whether a feature is shared or not, for each feature of a dialog context (Table VI.2). In addition, we include a number of feature functions based on general discourse and world knowledge. For example, if the system's intention is "inform(list_items)", the number of database query results becomes an important feature. If the number of results is greater than one, the most likely expected user's intention would be "declare(select_item)". If the number of results equals one, "delcare(buy_item)" would be the most probable intention. To let the dialog-context model be a probability distribution, the score function is divided by the normalization factor:

$$P(I|D) = \frac{\sum_{D_I} Similarity(D_I, D)}{\sum_{I'} \sum_{D'_I} Similarity(D'_I, D)} \tag{VI.8}$$

The task of sequence labeling is to assign a label to each element in some sequence, for which the assigned tags capture both the boundary and the type of any detected entities (e.g., values of additional information). This approach makes use of IOB encoding [94]; 'I' is used to label tokens inside an entity, 'B' is used to mark the beginning of an entity, and 'O' labels tokens outside any entity of interest. Consider the following sentence:

```
Here/O is/O twenty/B-NUM five/I-NUM dollars/B-UNIT
```

From the IOB tagging result, 'twenty five' is identified as a numerical expression and 'dollars' detected as a unit of money. To extract additional information, we use a linear-chain conditional random field (CRF) model [95]. A linear-chain CRF is de-

Table VI.3: A portion of the inventory of semantic labels for the shopping domain

Intention	**Speech Act**	greet, bye, ask, apologize, declare, inform, ack, request, suggest, order, thank, confirm, reject, feedback
	Main Goal	welcome, person_info, school_info, transportation, preference, compliment, homework_check, feeling_check, weather_check, bring_items, advertise_items, list_items, select_item, buy_item, scan_item, mistake_happen, total_price, change, payment, item_shortage, item_position, cancel, refund, exchange, recommend, given_tip, closing
Additional Information		student_name, student_age, student_grade, school_name, time, location, difficulty, weather, season, feeling, treatment_num, num, unit, item_name, item_type, currency, tagged_question

fined as follows. The objective of this modeling is to find the $\mathbf{S}$ that maximizes the conditional probability, $P(\mathbf{S}|X)$ in which $\mathbf{S} = \{S_t\}$ and $\mathbf{X} = \{X_t\}$ for $t = 1, \ldots, T$, such that $\mathbf{S}$ is a semantic class labeling of an observed word sequence $\mathbf{X}$. The conditional probability is estimated using Eq. VI.9:

$$P(\mathbf{S}|D) = \frac{1}{Z} exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\mu_k g_k(S_{t-1}, S_t, \mathbf{X}, t)\right), \qquad \text{(VI.9)}$$

where K is the number of features, g_k denotes the features, μ_k the weighted parameters for features, and Z is a normalization factor. This model offers a clean way to combine diverse pieces of linguistic information. As in the utterance model, we use lexical word features and POS tag features for the sequence labeling model.

The parameters of the hybrid model for intention and the sequence labeling model for additional information were trained on the labeled training corpus, for which we annotated the educational materials (Section 6.4.2) with speech act, main goal, additional information, and discourse features aforementioned. Table VI.3 shows the inventory of speech act, main goal, and additional information for the shopping domain.

Dialog management

The dialog manager plays a key controlling role in any conversational spoken language interface: given the semantic input corresponding to the current user utterance and the current discourse context, it determines the next action of the system. In essence, the dialog manager is responsible for planning and maintaining the coherence of the conversation. To accomplish this goal successfully, the dialog manager must maintain a history of the discourse and use it to interpret the perceived semantic inputs and a representation of the system task is typically required.

The simplest dialog manager is a finite-state manager. This system completely controls the conversation with the user. It asks the user a series of questions, ignoring anything that is not a direct answer to the question and then going on to the next question. Systems that control the conversation in this way are called system-initiative systems. System-initiative dialog managers may be sufficient for simple tasks such as entering a credit card number, but pure system-initiative dialog managers are probably too restrictive for a relatively complicated task like shopping. The problem is that pure system-initiative systems require that the user answer exactly the question that the system asked. But this can make a dialog awkward and annoying. In addition, it is theoretically possible to create a finite-state system that has a separate state for each possible subset of questions that the user's statement could be answering, but this would require a vast explosion in the number of states. Therefore we avoid the pure system-initiative approach and use an architecture that allows mixed initiative, in which the conversational initiative can shift between system and user at various points in the dialog.

In this study, dialog management is performed by RavenClaw [96], a plan-based, task-independent dialog management framework. RavenClaw isolates the domain-specific aspects of the dialog control logic from domain-independent conversational skills, and facilitates rapid development of mixed-initiative systems operating in complex, task-oriented domains. System developers can focus exclusively on describing the dialog task control logic, while a large number of domain-independent conversational skills such as error handling, timing and turn-taking are transparently supported and enforced by the RavenClaw dialog engine. Consider for instance error handling. System developers construct a dialog task specification under the assumption that inputs to the system will always be perfect, therefore ignoring the underlying uncertainties in the speech recognition channel. The responsibility for ensuring that the system maintains accurate information through confirmation actions (e.g., explicit/implicit confirmation) and that the dialog advances normally towards its goals is delegated to the dialog engine. Apart from the error handling strategies, the RavenClaw dialog management framework provides automatic support for a number of additional domain-independent conversational strategies. Examples include the ability to handle timeouts, requests for help, for repeating the last utterance, suspending and resuming the conversation, or starting again.

The dialog task specification describes a hierarchical plan for the interaction. More specifically, a dialog task specification consists of a tree of dialog agents, where each agent is responsible for handling a subpart of the interaction. For instance, Fig. VI.4 depicts a portion of the dialog task specification for the shopping domain.

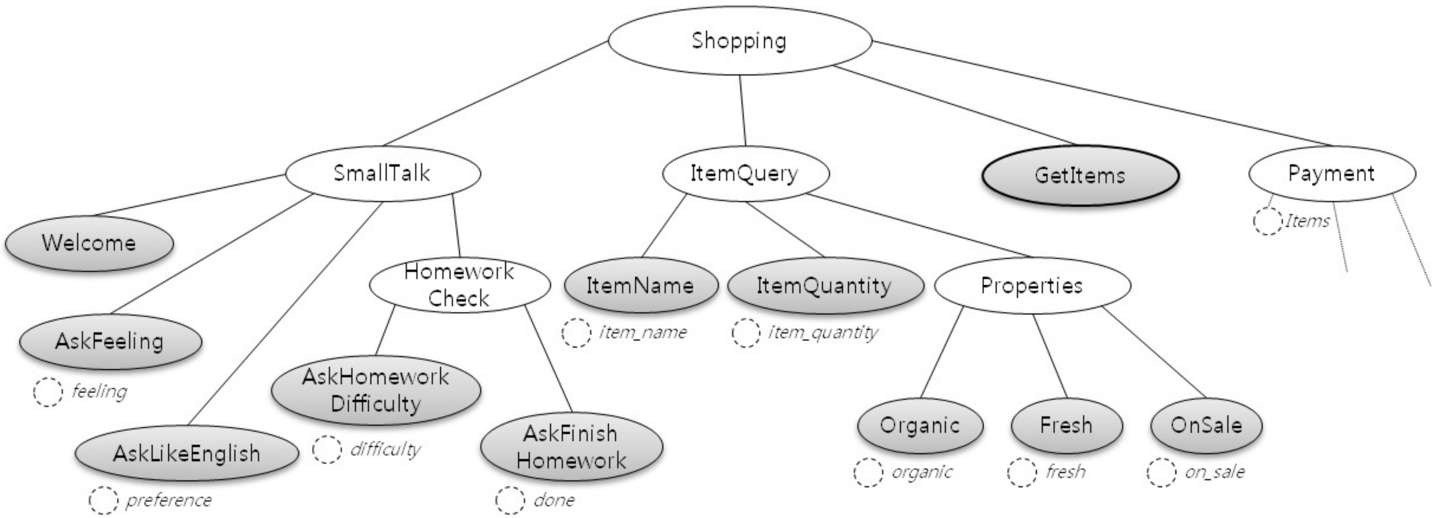

Figure VI.4: A portion of the dialog task tree for the shopping domain; clean circles – dialog agency, filled circles – dialog agent, dotted circles – concepts

The root node subsumes several ‘children’: SmallTalk, which engages the user in a daily conversation; ItemQuery, which obtains the fruit and vegetable properties from the user; GetItems, which executes the query against the backend; Payment, which presents the obtained results and handles the forthcoming negotiation for total price and performs payment. Moving one level deeper in the tree, the SmallTalk agent decomposes into Welcome, which provides a short welcome prompt and calls the user to come toward the robot; AskFeeling, which has a chat with users about their feelings, and finally AskLikeEnglish, which asks users whether they like English or not. The dialog agents in a dialog task specification fall into two categories: fundamental dialog agents, shown grayed in Fig. VI.4, and dialog agencies, shown in clear in Fig. VI.4. The fundamental dialog agents are located at the terminal positions in the tree (e.g., Welcome, AskFeeling) and implement atomic dialog actions, or dialog moves. There are four types of fundamental dialog agents: Inform – produces an output (e.g., Welcome); Request – requests information from the user (e.g., AskFeeling); Expect – expects information from the user, but without explicitly requesting it (e.g., Organic), and Execute – performs a domain-specific operation, such as database access (e.g., GetItems). The dialog agencies occupy non-terminal positions in the tree (e.g., SmallTalk, ItemQuery); their purpose is to control the execution of their subsumed agents, and encapsulate the higher level temporal and logical structure of the dialog task. Each dialog agent implements an Execute routine, which is invoked at runtime by the dialog engine. The execute routine is specific to the agent type. For example, inform agents generate an output when executed, while request agents generate a request but also collect the user’s response. For dialog agencies, the Execute routine is in charge of planning the execution of their subagents. In addition to the Execute routine, each dialog agent can define preconditions, triggers, as well as success and failure criteria. These are taken into account by the dialog engine and parent dialog agencies while planning the execution of the various agents in the tree. The tree representation captures the nested structure of dialog and thus implicitly represents context (via the parent relationship), as well as a default chronological ordering of the actions (i.e., left-to-right traversal). However, this developer-specified plan does not completely prescribe a fixed order for the execution of the various dialog agents. When the dialog engine executes a given dialog task specification, a particular trace through this hierarchical plan is followed, based on the user inputs, the encoded domain constraints and task logic, as well as the various execution policies in the dialog engine. If the dialog agents are the fundamental execution units in the RavenClaw dialog management framework, the data that the system manipulates throughout the conversation is encapsulated in concepts. Concepts can be associated with various agents in the dialog task tree, for instance feeling and preference in Fig. VI.4, and can be accessed and manipulated by any agent in the tree. Several basic concept types are predefined in the RavenClaw dialog management framework: Boolean, string, integer and float. Additionally, the framework provides support for more complex, developer–defined concept types such as (nested) structures and arrays. Internally, the ‘‘value’’ for each concept is represented by a set of value/confidence pairs, for instance item_name{apple/0.35; pineapple/0.27}. The dialog engine can therefore track multiple alternate hypotheses for each concept, and can capture the level of uncertainty in each hypothesis [97, 98]. Additionally, each concept also maintains the history of previous values, as well as information about the grounding state, when the concept was last updated,

Figure VI.5: Facial expressions for various emotions

etc.

When it is desirable to offer corrective feedback, the robot provides implicit and explicit feedback by using the methods described in Chapter III and IV.

6.3.2 Emotional expression

The human perception of a robot's emotional expressions plays a crucial role in human robot interaction. Mero and Engkey were designed with expressive faces that can represent different emotions: pleasure, dislike, neutrality, hope, fear, joy, distress, surprise, embarrassment, pride, shame and sadness (see Fig. VI.5). By virtue of its movable body, Engkey can also make diverse gestures by conducting a series of facial and body motions such as winking, yawning, cheering, sulking, etc, in accordance with the meaning of a verbal response:

6.3.3 Person identification

As previous research on interpersonal communication indicates, it is vital that two parties recognize each other for their relationship to develop [29]. We can develop a unique relationship with individuals because we can identify each of them [99, 100]. Although person identification is an essential requirement for an educational robot, current visual and auditory sensing technologies cannot reliably support it. Lighting conditions may vary, and the shapes and colors of the objects in the environment may be too complex for current computer vision technologies to function. In addition, the method of person identification must be robust because misidentification can ruin a relationship. Here, the robots identify individuals using a RFID system. Recent RFID technologies enabled using contactless ID cards in practical situations. Consequently, the robots can show some human-like behavior in which the robot can call a child's name if that child is at a certain distance. This behavior is useful for encouraging the child to come and interact with the robot.

6.4 Experimental design

To find general cognitive and affective effects of DB-CALL approaches which can motivate subsequent in-depth research, we designed and performed a field study at a Korean elementary school. The following subsections describe the method of the study in more detail.

6.4.1 Setting and participants

A total of 24 elementary students (12 male and 12 female) were enrolled in English lessons two days a week for a total of about two hours per day and had chant and dance time on Wednesdays for eight weeks during the winter vacation. However, three students left the study, resulting in a total of 21 students. Because the program was administered during the vacation, there was no other English class. The students ranged from third to fifth grade (nine students for third grade, seven for fourth, and eight for fifth); in general, there are six grades in a Korean elementary school and students start learning English from third grade. All of them were South Korean, spoke Korean as their first language and were learners of English as a foreign language. The participants were recruited by the teachers at the school from volunteers, through interviews, according to motivation and English proficiency. Then they were divided into beginner-level and intermediate-level groups, according to the pre-test scores. The evaluation rubric in Table VI.6 shows that the pre-test scores reflect the students' initial proficiency: students' pronunciation was understandable with some confirmation and misunderstanding; students' responses showed heavy reliance on beginner-level expressions with some communication breakdowns; students' responses contained grammar errors that are sometimes distracting to listeners and cause confusion about meaning; students replied with relatively short answers, requiring encouragement. The design of the field study, however, makes the precise role of DB-CALL approaches in facilitating L2 development less than clear. This is due to the lack of a control group which was necessitated by financial and scheduling constraints. Fig. VI.6 shows the layout of the classroom: (1) PC room

Figure VI.6: Students interacting with Mero and Engkey

where students took lessons by watching digital content; (2) Pronunciation training room where the Mero robot performed automatic scoring of pronunciation quality for students' speech and provided feedback; (3) Fruit and vegetable store, and (4) Stationery store where the Engkey robots acted as sales clerks and the students as customers.

6.4.2 Material and treatment

The researcher produced training materials including a total of 68 lessons, with 17 lessons for each combination of the level (beginner and intermediate) and the theme (fruit and vegetable store and stationery store). Among other things, the course involves small talk, homework checking, purchases, exchanges and refunds. When dealing with task assignment, the instructors proceeded in subtle gradations, moving from the simple to the complex. Throughout the course of the study, each student was asked to enter the four rooms in the order of PC room, Pronunciation training room, Fruit and vegetable store, and Stationery store so that students were gradually exposed to more active oral linguistic activities. Students were expected to spend about ten minutes in each training room. Although there were assistants, their roles were confined to fixing any technical problems with the robots. There was no English instruction in addition to the interaction with the robots during the period of this study.

6.4.3 Data collection and analysis

Cognitive effects

In order to measure the cognitive effects of the DB-CALL approach, i.e., improvement of listening and speaking skills, all students took a pre-test at the beginning of the study and a post-test at the end. For the listening skill test, 15 multiple-choice questions were used, which were developed by experts in evaluation of educational programs (see Fig. VI.7). The items in the test were mainly selected from the content taught during the course, as shown in Table VI.4.

1. You hear someone giving a talk. In this talk you have to listen for certain facts, and then decide what you have heard (   )

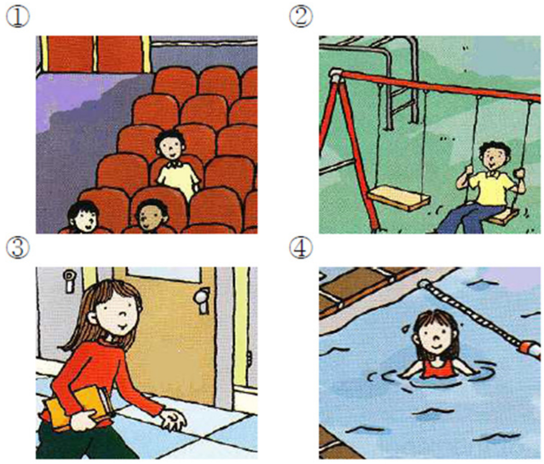

Figure VI.7: A multiple-choice question for the listening skill test

Table VI.4: Assessment items for listening tests

Question Number	Assessment Items
1	Words with similar sounds
2	Expressions for asking about items
3	Expressions about transportation
4	Expressions about weather
5	Expressions about location of building
6	Expressions for asking and answering about time
7	Expressions about item features
8	Expressions about price and number
9	Expressions about emotion and body condition
10	Expressions about time
11	Expressions about quantity of items
12	Expressions for purchasing items
13	Expressions about what has been done
14	Expressions for purchasing items
15	Expressions about type of currency

Table VI.5: Assessment items for speaking tests

Question Number	Assessment Items
1	Greeting, introducing oneself, and asking about present states
2	School name, transportation, amount of time required to go to one's school
3	Expressions related to learning English
4	Expressions about item names, price, and refund
5	Expressions related to weather and recommendation
6	Expressions about item names and ordinal numbers
7	Asking for items and understanding confirmation
8	Comparative expressions
9	Expressions about getting back change
10	Expressions about item features

The test was used as the assessment tool in both the pre-test and the post-test phases of the study. The internal consistency estimates of reliability, Cronbach's alpha [101], were computed for the listening test. The value of Cronbach's alpha for the pre-test was .87 and the value for the post-test was .66, each indicating satisfactory reliability. The speaking skill test consisted of 10 one-on-one interview items. All speaking assessment tasks were carried out by a teacher from the participating school with an advanced degree in Education. The topics of the interviews were selected from the content taught (see Table VI.5).

The evaluation rubric measured speaking proficiency on a five-point scale in four categories: pronunciation, vocabulary, grammar, and communicative ability, as shown in Table VI.6.

The value of Cronbach's alpha for the pre-test was .93 and the value for the post-test was .99, each indicating satisfactory reliability. A paired t-test was performed using the meanscoresandstandarddeviationstodetermine if any significant differences

Table VI.6: Assessment items for speaking tests

Category	Criteria	Score
Pronunciation	Student's pronunciation was relatively accurate.	5
	Student's pronunciation showed some problems with individual sounds, but did not cause problems in intelligibility.	4
	Student's pronunciation was understandable with some confirmation and misunderstanding.	3
	Student's pronunciation made understanding difficult due to numerous errors.	2
	Student's pronunciation was incomprehensible.	1
Vocabulary	Student's response showed appropriate words and idioms.	5
	Although one or more words may not be precise, the response was informationally appropriate.	4
	Student's response showed heavy reliance on beginner-level expressions with some communication breakdowns.	3
	Student can speak at the phrase level, but showed plenty of repeats and repairs.	2
	Student had difficulty in speaking even one or two words.	1
Grammar	Student's response was well structured.	5
	Student's response had at most minor lapses and did not cause confusion about meaning.	4
	Student's response contained errors that are sometimes distracting to listeners and cause confusion about meaning.	3
	Student's response contained many errors leading to communication breakdowns.	2
	Student's response was unintelligible.	1
Communicative ability	Student actively engaged in conversation with high confidence and the response was clear and intelligible.	5
	Student showed a lack of confidence in gestures and facial expressions, but sustained coherent discourse.	4
	Student replied with relatively short answers, requiring encouragement.	3
	Student replied with very short answers with a lack of confidence.	2
	Student often refused to speak.	1

occurred.

Affective effects

In order to investigate the effects of DB-CALL on affective factors such as satisfaction in using robots, interest in learning English, confidence with English, and

Table VI.7: Internal consistency estimates of reliability

Affective Factor	N[a]	R[b]
Satisfaction in using robots	10	0.73
Interest in learning English	16	0.93 (0.96)
Confidence with English	12	0.91 (0.90)
Motivation for learning English	14	0.91 (0.83)

[a] Number of questions
[b] Cronbach's alpha in the form of pre-test (post-test)

motivation for learning English, a questionnaire was designed by ten teachers and experts in the evaluation of educational programs. It consisted of some personal information and 52 statements in accordance with a four-point Likert scale, which had a sliding answer scale of 1–4, ranging from “strongly disagree” to “strongly agree”, without a neutral option. Mean and standard deviation were used to evaluate the effect on students' satisfaction, whereas a pre-test/post-test method was used for other factors. The internal consistency estimates of reliability, Cronbach's alpha, was computed to indicate satisfactory reliability (see Table VI.7).

6.5 Results and discussion

6.5.1 Cognitive effects

The achievement of the students in the beginner group on pre- and post-test is presented in Table VI.8. According to the findings in this table there were large improvements in the participants' speaking skills achievement in the post-test. The score in the post-test is significantly better than that of the pre-test. The effect sizes, which were calculated following the formula proposed in [102], range over 0.82–0.90, showing large effects. We conducted the Bonferroni test [103] for a simultaneous

Table VI.8: Cognitive effects on oral skills for the beginner group

Category	N	Pre-test Mean	Pre-test SD[a]	Post-test Mean	Post-test SD[a]	Mean difference	t	df	Effect size
Listening	10	8.60	2.84	9.50	1.90	0.90	1.03	9	0.32
Speaking									
Pronunciation	10	26.90	8.39	41.80	1.81	14.90	6.10*	9	0.90
Vocabulary	10	27.50	8.02	38.10	3.51	10.60	4.85*	9	0.85
Grammar	10	27.20	7.45	37.30	3.37	10.10	5.74*	9	0.89
Communicative ability	10	30.60	12.00	45.60	2.91	15.00	4.37*	9	0.82
Total	10	112.20	35.23	162.80	10.81	50.60	5.34*	9	0.87

* $p<.01$
[a] Standard Deviation

inference to test whether or not the four categories under speaking skills have significant differences simultaneously. The result showed that there was a significant difference in a simultaneous inference at the significance level of 0.01. The listening skill, however, showed no significant difference.

Significant differences in speaking skills were also found in the result of the intermediate group and the effect sizes are also large, whereas the listening skill showed a significantly negative effect (see Table VI.9). In addition, the result of the Bonferroni test showed that the four categories under speaking skills have a significant difference in a simultaneous inference at the significance level of 0.01.

The combined results of both groups showed no significant differences in listening skills (see Table VI.10). This finding can be explained by a number of factors such as the unsatisfactory quality of the text-to-speech component and the robots' various sound effects (e.g., alarms, musical instruments) which can distract learners' attention from the robots' speech. However, significant differences in speaking skills were found in the overall results and the result of the Bonferroni test showed that the

Table VI.9: Cognitive effects on oral skills for the intermediate group

		Pre-test		Post-test		Mean			Effect
Category	**N**	**Mean**	**SD[a]**	**Mean**	**SD[a]**	**difference**	***t***	**df**	**size**
Listening	11	13.09	1.64	11.73	1.19	21.36	23.32*	10	0.72
Speaking									
Pronunciation	11	36.91	6.43	49.09	2.43	12.18	7.72*	10	0.93
Vocabulary	11	36.00	6.24	46.27	3.23	10.27	6.41*	10	0.90
Grammar	11	35.64	6.28	43.64	2.84	8.00	4.94*	10	0.84
Communicative ability	11	36.27	6.83	49.18	2.09	12.91	7.53*	10	0.92
Total	11	144.82	25.60	188.18	10.19	43.36	6.82*	10	0.91

* p<.01
[a] Standard Deviation

Table VI.10: Cognitive effects on oral skills for overall students

		Pre-test		Post-test		Mean			Effect
Category	**N**	**Mean**	**SD[a]**	**Mean**	**SD[a]**	**difference**	***t***	**df**	**size**
Listening	21	10.95	3.2	10.67	1.91	20.29	20.55	20	0.12
Speaking									
Pronunciation	21	32.14	8.86	45.62	4.28	13.48	9.48*	20	0.90
Vocabulary	21	32.95	8.21	42.38	5.31	10.43	8.00*	20	0.87
Grammar	21	31.62	7.96	40.62	4.43	9.00	7.59*	20	0.86
Communicative ability	21	33.57	9.83	47.48	3.06	13.91	7.60*	20	0.86
Total	21	123.13	34.13	176.10	16.53	46.81	8.48*	20	0.88

* p<.01
[a] Standard Deviation

four categories under speaking skills have a significant difference in a simultaneous inference at the significance level of 0.01. The large improvement in speaking skills in the overall results agrees with the findings of previous studies in general. Specifically, based on the evaluation rubric, the gain in the vocabulary area reveals that before the treatments students were limited to heavy reliance on very simple expressions with some communication breakdowns, but after the treatments their responses became informationally appropriate with only one or more imprecise words. This may indicate that the authentic context facilitated form-meaning mapping and the vocabulary acquisition process. The improved accuracy of pronunciation shows that the treatments made students' pronunciation more intelligible. Before the treatments, their pronunciation was understandable only with some confirmation and misunderstanding. The improvement in the grammar area shows that after the treatments there were at most minor lapses that did not cause confusion about meaning, compared to serious errors before the treatment that sometimes distracted listeners and caused confusion about meaning. This may support the output hypothesis and the effects of corrective feedback. The fact that learners had feedback at any related point made them reflect on their erroneous utterances. The increase in communicative ability means that before the treatments students required encouragement even for replying with short answers, yet they could sustain coherent discourse by themselves after the treatments. This may show that learners were getting accustomed to speaking English. It can also be attributed to the fact that when using robot-assisted learning the student gained confidence in a relaxed atmosphere. A lack of confidence and a feeling of discomfort were more related to students' participation in face-to-face traditional discussions, and less to participation in computer-based learning. Although the absence of a control group makes the result less than clear, given the results of a previous study [104] showing decreased scores in control groups in which learners do not participate in any treatment sessions and the positive affective effects (Section 6.5.2) related to oral skills, the likely interpretation is that the treatments contributed to the improvement in oral skills.

6.5.2 Affective effects

As shown in Table VI.11, the students were highly satisfied about using robots for language learning. It is worth noting that a large portion of the satisfaction was associated with students' recognition of robots as intellectual beings capable of human-like social interactions such as watching, listening and moving toward students. This result supports the benefit of the robots' capacity to create interpersonal relationships with the students [99, 100]. In comparison to the other questions, the questions about the robot's outer appearance (e.g., ''The robot's body looks comfortable for moving around in a classroom'' and ''The robot's facial expression looks comfortable to you) and voice (e.g., ''You like the robot's voice'') showed the lowest level of satisfaction, showing the need to develop a more anthropomorphic appearance and a natural voice. The low level of satisfaction regarding the robot's voice can explain, in part, the lack of improvement in students' listening skills. The robot's speech synthesizer has only addressed comprehensibility whereas CALL places demands on naturalness, accuracy, and expressiveness as well. In order to fully meet the requirements of CALL, further attention needs to be paid to accuracy and naturalness, in particular at the prosodic level, and to expressiveness [105].

The students' responses to the questions about their interest in learning English on pre- and post-test are presented in Table VI.12, showing a large improvement of interest with a significance level of 0.01. The response to the question ''Singing a song, chanting, and other games are interesting'' shows that the robot's physical body, one of its unique features, had a great influence on students' interest by enabling the robots to dance, make gestures and use facial expressions. In addition, the

Table VI.11: Cognitive effects on oral skills for overall students

Question	Strongly disagree	Disagree	Agree	Strongly agree	N	Mean	SD[a]
The robot looks smart	0	3	10	8	21	3.24	0.70
The robot can watch you	0	2	10	9	21	3.33	0.66
The robot can listen to your song and speech	0	7	7	6	20	2.95	0.83
The robot can come to you	1	7	6	7	21	2.90	0.94
The robot's appearance looks comfortable for learning	2	5	7	7	21	2.90	1.00
The robot's body looks comfortable for moving around in a classroom	2	6	11	2	21	2.62	0.80
The robot's facial expression	3	3	13	2	21	2.67	0.86
looks comfortable to you	1	0	10	10	21	3.38	0.74
The robot's compliment is pleasing to you							
You like the robot's voice	3	5	9	4	21	2.67	0.97
The robot seems secure	0	3	12	6	21	3.14	0.65
Total						2.98	0.44

[a] Standard Deviation

large improvement shown by the response to the question "You think English is easier and more familiar than before" may support the affective filter hypothesis [106] which states that the blockage can be reduced by sparking interest and providing low anxiety environments. The question "You want to use the expressions learned" also showed a large improvement, which may be attributable to the immediate application of the learned expressions through conversations with robots. The only question that had a score of less than three is "You want to talk with other people in English". This result agrees with the findings of previous studies that a feeling of discomfort was more related to students' participation in face-to-face traditional conversations and less to participation in robot-assisted learning.

Table VI.12: Students' interest in learning English

Question	Stage	Strongly disagree	Disagree	Agree	Strongly agree	N	Mean	SD[a]	MD[b]	*t*
English class is interesting	Pre	2	5	5	6	18	2.83	1.04	0.500	1.58
	Post	2	1	4	11	18	3.33	1.03		
Listening and speaking English is interesting	Pre	4	2	7	4	17	2.65	1.11	0.294	0.96
	Post	2	3	6	7	18	3.00	1.03		
Singing a song, chanting, and other games are interesting	Pre	3	3	6	6	18	2.83	1.10	0.824	3.00**
	Post	1	0	4	12	17	3.59	0.80		
You want to talk with other people in English	Pre	7	4	4	3	18	2.17	1.15	0.500	1.84
	Post	3	6	3	6	18	2.67	1.14		
You enjoy learning English	Pre	2	4	9	3	18	2.72	0.89	0.444	2.20*
	Post	2	2	5	9	18	3.17	1.04		
You want to use the expressions learned	Pre	3	1	10	4	18	2.83	0.99	0.667	3.37**
	Post	1	1	4	12	18	3.50	0.86		
You inquire about an unknown word to a dictionary or others	Pre	2	3	7	6	18	2.94	1.00	0.222	1.46
	Post	2	1	7	8	18	3.17	0.99		
You want to participate in English class with passion	Pre	1	2	10	5	18	3.06	0.80	0.056	0.33
	Post	1	2	9	6	18	3.11	0.83		
You want more English classes in school	Pre	1	6	5	6	18	2.89	0.96	0.176	0.82
	Post	1	2	8	6	17	3.12	0.86		
You are looking forward to English class	Pre	2	4	6	6	18	2.89	1.02	0.333	2.92**
	Post	1	2	7	8	18	3.22	0.88		
You want to study English more in the future	Pre	3	4	4	7	18	2.83	1.15	0.278	1.76
	Post	2	2	6	8	18	3.11	1.02		
You pay attention to what you are going to learn in English class	Pre	2	6	7	3	18	2.61	0.92	0.556	3.83**
	Post	2	2	5	9	18	3.17	1.04		

You try to remember what you have heard	Pre	2	3	8	5	18	2.89	0.96	0.444	2.68*
	Post	2	1	4	11	18	3.33	1.03		
You think English is easier and more familiar than before	Pre	3	4	7	4	18	2.67	1.03	0.611	2.83*
	Post	2	0	7	9	18	3.89	0.96		
You are curious about what you are going to learn in your next English class	Pre	2	3	8	5	18	2.89	0.96	0.333	1.46
	Post	1	3	5	9	18	3.22	0.94		
You are interested in English	Pre	4	3	4	7	18	2.78	1.22	0.611	3.05**
	Post	1	1	6	10	18	3.39	0.85		
Total	Pre					18	2.78	0.71	0.430	3.21**
	Post					18	3.21	0.74		

* $p<.05$
** $p<.01$
[a] Standard Deviation
[b] Mean Difference

A significantly large increase in confidence was found in the responses to the questions about confidence in English on the pre- and post-test with a significance level of 0.01 (see Table VI.13). This can also be attributed to the fact that robot-assisted learning allows the students to achieve academically and gain confidence through repeated exercises in a relaxed atmosphere. However, relatively low scores were given to the questions relating to individual levels of fear or anxiety associated with either real or anticipated communication with another person or persons (e.g., ‘‘You are not afraid of speaking English,’’ ‘‘You are not afraid of being questioned by the English teacher,’’ and ‘‘You feel no shame about your English mistakes’’). Therefore robots should help students to feel that they can learn the foreign language well by using more encouragement and praise. Classroom atmosphere is very important; it should be happy, lively, friendly and harmonious to help students overcome their psychological barriers, and lower their anxiety. Robots should also tolerate a few small mistakes made by students provided those mistakes do not affect the communication process, because this can release pressure and strengthen their confidence. Given the large improvement shown by the response to the question ‘‘You can greet foreigners with confidence’’, we can infer that the confidence gained through the repeated greetings with robots was transferred, to some extent, to greetings with foreigners. We can also find some grounds for assuming an improvement in the students’ cognitive abilities from the large gain in respect of the question ‘‘You can answer the questions about what you have learned with confidence’’. Finally, the great improvement regarding the question ‘‘You think that you can speak English better if you study harder’’ is very impressive, given that many Korean people regard English as a very difficult language to learn, however hard they study.

Table VI.13: Students' confidence with English

Question	**Stage**	**Strongly disagree**	**Disagree**	**Agree**	**Strongly agree**	**N**	**Mean**	**SD**[a]	**MD**[b]	*t*
You understand what you have learned	Pre	4	3	8	5	20	2.70	1.08	0.400	1.71
in English class well	Post	3	2	5	10	20	3.10	1.12		
You can answer the questions about what you	Pre	3	6	6	4	19	2.58	1.02	0.579	2.63*
have learned with confidence	Post	2	2	6	9	19	3.16	1.01		
You are not afraid of speaking English	Pre	3	8	6	3	20	2.45	0.94	0.300	1.10
	Post	2	6	7	5	20	2.75	0.97		
You sing songs and chant with confidence	Pre	2	7	4	7	20	2.80	1.06	0.400	1.80
	Post	2	2	6	10	20	3.20	1.01		
You are not afraid of being questioned	Pre	3	6	9	2	20	2.50	0.89	0.250	1.10
by the English teacher	Post	1	7	8	4	20	2.75	0.85		
You will participate in English learning activities	Pre	3	5	6	6	20	2.75	1.07	0.200	1.29
(role-play, game) actively	Post	2	4	7	7	20	2.95	1.00		
You are not afraid of English homework	Pre	3	6	8	3	20	2.55	0.94	0.600	3.04**
	Post	1	3	8	8	20	3.15	0.88		
You think you can make a good presentation	Pre	3	5	6	6	20	2.75	1.07	0.450	3.33**
in English classes	Post	1	1	11	7	20	3.20	0.77		
You fully understand what you have learned	Pre	2	2	11	5	20	2.95	0.89	0.400	2.99**
in English classes	Post	1	0	10	9	20	3.35	0.75		
You feel no shame about your English mistakes	Pre	4	11	4	1	20	2.10	0.79	0.550	2.77*
	Post	1	7	10	2	20	2.65	0.75		

You can greet foreigners with confidence	Pre	4	4	7	5	20	2.65	1.09	0750	3.94**
	Post	0	2	8	10	20	3.40	0.68		
You think that you can speak English better	Pre	3	0	8	9	20	3.15	1.04	0.700	3.20**
if you study harder	Post	0	0	3	17	20	3.85	0.37		
Total	Pre					20	2.66	0.70	0.460	3.53**
	Post					20	3.12	0.70		

* p<.05
** p<.01
[a] Standard Deviation
[b] Mean Difference

The responses to the questions about motivation for learning English are presented in Table VI.14. There has been a large enhancement of motivation, with a significance level of 0.01. Interestingly, there was a relatively large increase in the score for the question ‘‘You are aware of the necessity of studying English’’ compared to the small difference for the question ‘‘You recognize the importance of English to your present and future life.’’ It may mean that the experience of having a conversation with a robot enabled the learners to focus on their practical needs in English, thus overcoming some of their communicative difficulties. The low scores for the questions related to preparing to study English (e.g., ‘‘You want to buy English books and materials,’’ ‘‘You make plans to study English,’’ and ‘‘You enjoy preparing for English classes’’) may illustrate that traditional education does not work for the new generation of children. The popularity of e-Learning in Korea is promoting an increasing disengagement of the ‘‘Net Generation’’ or ‘‘Digital Natives’’ from traditional instruction.

Table VI.14: Students' confidence with English

Question	Stage	Strongly disagree	Disagree	Agree	Strongly agree	N	Mean	SD[a]	MD[b]	*t*
You are aware of the necessity of studying English	Pre	3	1	9	7	20	3.00	1.03	0.600	2.85**
	Post	0	2	4	14	20	3.60	0.68		
You want to read signboards and lyrics written in English	Pre	3	4	5	6	18	2.78	1.11	0.500	3.43**
	Post	2	3	3	11	19	3.21	1.08		
You recognize the importance of English to your present and future life	Pre	0	0	5	15	20	3.75	0.44	0.053	1.00
	Post	0	0	3	16	19	3.84	0.37		
You want to learn English more	Pre	2	4	4	10	20	3.10	1.07	0.500	3.25**
	Post	0	1	6	13	20	3.60	0.60		
You want to study hard in English classes	Pre	1	1	7	11	20	3.40	0.82	0.350	1.93
	Post	0	0	5	15	20	3.75	0.44		
You have a desired level of English ability	Pre	1	2	10	7	20	3.15	0.81	0.400	2.99**
	Post	1	1	4	14	20	3.55	0.83		
You want to buy English books and materials	Pre	2	10	6	2	20	2.40	0.82	0.550	3.58**
	Post	2	2	11	5	20	2.95	0.89		
You make plans to study English	Pre	4	7	6	3	20	2.40	0.99	0.300	2.04
	Post	3	4	9	4	20	2.70	0.98		
You spend more time on studying English by yourself	Pre	4	5	7	4	20	2.55	1.05	0.600	3.27**
	Post	2	2	7	9	20	3.15	0.99		
You enjoy preparing for English classes	Pre	7	5	4	4	20	2.25	1.16	0.550	2.78*
	Post	2	5	8	5	20	2.80	0.95		
You want to get praised for your English	Pre	0	1	7	12	20	3.55	0.60	0.350	2.67*
	Post	0	0	2	18	20	3.90	0.31		

You want to write down your ideas in English	Pre	5	5	4	6	20	2.55	1.19	0.500	2.13*
	Post	2	4	5	9	20	3.05	1.05		
You want to use what you have learned in English classes in everyday life	Pre	4	3	7	6	20	2.75	1.12	0.450	2.65*
	Post	1	3	7	9	20	3.20	0.89		
You want to converse with foreigners actively	Pre	6	5	5	4	20	2.35	1.14	0.600	3.27**
	Post	3	2	8	7	20	2.95	1.05		
Total	Pre					20	2.86	0.64	0.440	4.99**
	Post					20	3.30	0.45		

* p<.05
** p<.01
[a] Standard Deviation
[b] Mean Difference

6.6 Conclusion

In this study, we introduced the HRI technologies we used to implement the educational assistant robots. To investigate the cognitive and affective effects of robot-assisted learning, a course was designed in which intelligent robots act as sales clerks in a fruit and vegetable store, and in a stationery store so that they can interact in real life situations with language learners who play the part of customers. A pretest/post-test design was used to investigate the cognitive effects of the DB-CALL approach on the students' oral skills. The results showed no significant difference in the listening skill, but the speaking skills improved with a large effect size at the significance level of 0.01. This may mean that DB-CALL approaches can provide valuable leads in helping students to enhance their speaking ability, particularly in the case of Korean students where teaching is generally focused predominantly on vocabulary and grammar. Descriptive statistics and pre-test/post-test design were used to investigate the affective effects of the DB-CALL approach. The results showed that DB-CALL promotes and improves students' satisfaction, interest, confidence, and motivation at the significance level of 0.01. In addition, the result of the Bonferroni test showed that the three categories under affective factors (i.e., interest, confidence, motivation) have a significant difference in a simultaneous inference at the significance level of 0.01. Throughout all affective areas, unique features of robots made a great influence on the students' responses showing that DB-CALL can be an enjoyable and fruitful activity for students. Although the results of this study bring us a step closer to understanding DB-CALL approaches, subsequent in-depth research should be conducted to ascertain the detailed effects of each possible factor involved in DB-CALL methods. Also, the results are only valid for Korean elementary students. More studies are needed to consolidate/refute the findings of this study over longer periods of time using different activities with samples of learners of different ages, nationalities, and linguistic abilities. Given that studies on DB-CALL are still relatively new and most are in the early stages, further research is needed into the use of DB-CALL systems for educational purposes and the effects of their use in this field.

Chapter VII

CONCLUSION

In this thesis, we considered intelligent corrective feedback for language learning that handles learners' errors and helps learners to use more appropriate words and expressions. On the basis of the SLA theory and practice, we have distinguished global error from local error and designed different methods for each type of error. When a learner speaks, we judge whether the utterance has a global error or local errors. In case of global error, a DB-CALL system should provide matching fluent utterances. We proposed a novel hybrid model that allows natural decomposition between the utterance model and the dialog context model in terms of the psychological language production process. It led to an efficient way for adjusting to diverse fluency levels with minimal efforts. In addition, our elaborate dialog context model using enhanced k-nearest neighbors algorithm gave rise to more accurate inference of the language learners' intention. Also, it proved to be effective to provide appropriate context-aware feedback so that the learners can obtain positive rewards by successfully completing dialogs.

When language learners make local errors, it would be wiser to pinpoint and correct the erroneous part with keeping other parts untouched. It is not a trivial task to detect grammatical errors in oral conversations because of the unavoidable errors of automatic speech recognition systems. To provide corrective feedback, a novel method to detect grammatical errors in speaking performance was proposed. The proposed method consists of two sub-models: the grammaticality-checking model and the error-type classification model. We automatically generate grammatical errors that learners are likely to commit and construct error patterns based on the articulated errors. When a particular speech pattern is recognized, the grammaticality-checking model performs a binary classification based on the similarity between the error patterns and the recognition result using the confidence score. The error-type classification model chooses the error type based on the most similar error pattern and the error frequency extracted from a learner corpus. The results showed that for the grammaticality-checking task, the proposed method largely outperformed the two comparative models respectively by 56.36% and 42.61% in F-score while keeping the false positive rate very low. For the error-type classification task, the proposed method exhibited very high performance with a 99.6% accuracy rate. Because high precision and a low false positive rate are important criteria for the language tutoring setting, the proposed method will be helpful for intelligent CALL systems.

An automatic method for simulating realistic grammatical errors is crucial for advanced technologies in CALL, including generating corrective feedback in DB-CALL systems, simulating a language learner to optimize tutoring strategies, and generating context-dependent grammar quizzes as educational materials. Whereas previous studies did not consider sophisticated contextual information due to the lack of a flexible language, we could import the characteristics of intralingual errors of English learners and the interlingual errors of non-native speakers into statistical models by using Markov logic. The proposed method outperformed the baseline method, increasing the precision by 6% and the recall by 8.33% averaged across all proficiency levels. The proposed method also generated an error distribution that is more similar than that generated by the baseline method to the error distribution produced by real language learners. The proposed method produced a relative improvement of 37.5% in the average KL divergence. In accordance with acoustic scores, the enhanced quality of the simulated error distribution might increase an automatic speech recognizer's ability to detect grammatical errors by weighting errors that are generated more frequently more heavily. In addition, as the simulated errors become more similar to real learners' errors, the tutoring strategies of DB-CALL systems learned using language learner simulators would become more realistic and effective for real learners. Human evaluators judged that the proposed model is more likely to produce realistic errors than the baseline model. The proposed mode increased the average scores in two different evaluation tasks by 7 and by 0.411. In addition, the time performance gain rate showed that the construction of erroneous sentences with the help of grammatical error simulation is much more efficient than purely manual production. Using the proposed method reduced the grammatical error generation time by 59% in average.

Finally, we introduced the educational assistant robots that we developed for foreign language learning and explores the educational effectiveness of DB-CALL which is in its early stages. To investigate the cognitive and affective effects, a course was designed in which intelligent robots act as sales clerks in a fruit and vegetable store, and in a stationery store so that they can interact in real life situations with language learners who play the part of customers. A pretest/post-test design was used to investigate the cognitive effects of the DB-CALL approach on the students' oral skills. The results showed no significant difference in the listening skill, but the speaking skills improved with a large effect size at the significance level of 0.01. This may mean that DB-CALL approaches can provide valuable leads in helping students to enhance their speaking ability, particularly in the case of Korean students where teaching is generally focused predominantly on vocabulary and grammar. Descriptive statistics and pre-test/post-test design were used to investigate the affective effects of the DB-CALL approach. The results showed that DB-CALL promotes and improves students' satisfaction, interest, confidence, and motivation at the significance level of 0.01. In addition, the result of the Bonferroni test showed that the three categories under affective factors (i.e., interest, confidence, motivation) have a significant difference in a simultaneous inference at the significance level of 0.01. Throughout all affective areas, unique features of robots made a great influence on the students' responses showing that DB-CALL can be an enjoyable and fruitful activity for students. Although the results of this study bring us a step closer to understanding DB-CALL approaches, subsequent in-depth research should be conducted to ascertain the detailed effects of each possible factor involved in DB-CALL methods. Also, the results are only valid for Korean elementary students. More studies are needed to consolidate/refute the findings of this study over longer periods of time using different activities with samples of learners of different ages, nationalities, and linguistic abilities. Given that studies on DB-CALL are still relatively new and most are in the early stages, further research is needed into the use of DB-CALL systems for educational purposes and the effects of their use in this field.

요약문

외국어 회화 교육을 위한 지능적인 교정적 피드백 방법

국제화 시대를 맞아 영어 교육에 전 세계적인 투자가 이루어졌지만 아직도 영어 교육 방법에는 별 다른 차이가 생기지 않고 있다. 현 영어 교육 방식의 문제점들을 고려하여, 이 논문은 대화 기반 컴퓨터 언어 교육 시스템 (Dialog-Based Computer-Assisted Language Learning)에 관한 연구를 다룬다. 이 논문은 제 2 언어 습득 이론, 관련 기술, 시스템 및 현장 실험을 포함한 일련의 연구를 담고 있다. 일반적으로 많은 발음 및 문법 오류를 가지는 외국어 학습자들이 최근 활발히 연구되고 있는 음성 대화 시스템을 이용하여 외국어 회화 연습을 할 수 있도록 본 연구는 먼저 기존 음성 대화 시스템의 비 모국어 화자 사용을 위한 다양한 적응화 방법들을 적용하였다. 하지만 자유로운 회화 연습이 외국어 습득에 매우 중요한 것은 사실이나, 그것만으로는 외국어 학습자들이 높은 외국어 능력을 습득하는데는 한계가 있다. 외국어 학습자들의 정확한 언어 구사를 위해서는 학습자들의 문법 오류에 대한 교정적 피드백 (Corrective Feedback) 제공이 필요하다. 따라서 본 연구의 중심적인 부분은 학습자들이 더 적합한 단어와 표현을 사용할 수 있도록 학습자들의 오류를 검출하고 적절한 피드백을 제공하는 기술에 있다. 지금까지 많은 연구들이 쓰기에 관한 문법 오류 검출과 교정적 피드백 제공에 관한 것인 반면 음성 발화에 대한 연구는 거의 이루어지지 않았다. 이 논문은 제 2 언어 습득 이론에 기반한 음성 발화에 대한 효과적인 교정적 피드백 전략 및 그것을 가능하게 하는 기술에 대해 서술한다. 그리고 논문에 제시된 모든 기술들을 통합하여 개발한 영어 교육 로봇과 3D 가상 환경에서의 외국어 교육 게임을 소개한다. 마지막으로, 외국어 회화 교육에 대한 본 연구의 실제적인 효과를 검증하기 위해서 영어 교육 로봇을 한국 초등학교에 배치하여 현장 실험을 수행하였고 그 결과 본 연구가 외국어 학습자들에게 즐겁고

효과적인 영어 교육 방법론임을 보였다.

REFERENCES

[1] S. D. Krashen, *The input hypothesis: Issues and implications.* New York: Longman, 1985.

[2] M. Swain, S. Gass, and C. Madden, "Communicative competence: Some roles of comprehensible input and output in its development," 1985.

[3] M. H. Long, "Focus on form in task-based language teaching," in *Language policy and pedagogy: Essays in honor of A. Ronald Walton*, p. 179–192, Amsterdam: John Benjamins Publishing Company, 2000.

[4] A. M. Masgoret and R. C. Gardner, "Attitudes, motivation, and second language learning: A Meta-Analysis of studies conducted by gardner and associates," *Language Learning*, vol. 53, no. S1, p. 167–210, 2003.

[5] G. Brown and G. Yule, *Discourse analysis.* Cambridge: Cambridge University Press, 1983.

[6] D. Byrne, *Teaching Oral English.* Harlow: Longman, 1986.

[7] S. Garrod, "Language comprehension in context: A psychological perspective," *Applied Linguistics*, vol. 7, no. 3, p. 226, 1986.

[8] G. Brown, "Investigating listening comprehension in context," *Applied Linguistics*, vol. 7, no. 3, p. 284, 1986.

[9] W. S. R., "Attention," in *Cognition and second language instruction*, pp. 3–32, Cambridge: Cambridge University Press, 2001.

[10] H. Johnson, "Defossilizing," *ELT Journal*, vol. 46, no. 2, p. 180, 1992.

[11] A. Liang and R. J. McQueen, "Computer assisted adult interactive learning in a Multi-Cultural environment.," *Adult Learning*, vol. 11, no. 1, pp. 26–29, 1999.

[12] J. Roed, "Language learner behaviour in a virtual environment," *Computer Assisted Language Learning*, vol. 16, no. 2, p. 155–172, 2003.

[13] H. Yi and J. Majima, "The teacher-learner relationship and classroom interaction in distance learning: A case study of the japanese language classes at an american high school," *Foreign Language Annals*, vol. 26, no. 1, p. 21–30, 1993.

[14] S. Bryan, "The relationship between negotiated interaction, learner uptake, and lexical acquisition in task-based computer-mediated communication," *Tesol Quarterly*, vol. 39, no. 1, p. 33–58, 2005.

[15] C. Lai, F. Fei, and R. Roots, "The contingency of recasts and noticing.," *CALICO Journal*, vol. 26, no. 1, p. 21, 2008.

[16] C. Lai and Y. Zhao, "Noticing and text-based chat," *Language Learning & Technology*, vol. 10, no. 3, p. 102–120, 2006.

[17] S. Loewen and R. Erlam, "Corrective feedback in the chatroom: An experimental study," *Computer Assisted Language Learning*, vol. 19, no. 1, p. 1–14, 2006.

[18] R. Sachs and B. Suh, "Textually enhanced recasts, learner awareness, and l2 outcomes in synchronous computer-mediated interaction," in *Conversational interaction in second language acquisition: A collection of empirical studies*, p. 197–227, Oxford: Oxford University Press, 2007.

[19] B. Smith, "Computer-mediated negotiated interaction and lexical acquisition," *Studies in Second Language Acquisition*, vol. 26, no. 03, p. 365–398, 2004.

[20] N. Nagata, "BANZAI: an application of natural language processing to web-based language learning," *CALICO JOURNAL*, vol. 19, no. 3, p. 583–600, 2002.

[21] T. Heift and D. Nicholson, "Web delivery of adaptive and interactive language tutoring," *International Journal of Artificial Intelligence in Education*, vol. 12, no. 4, p. 310–324, 2001.

[22] T. Heift and M. Schulze, *Errors and Intelligence in CALL. Parsers and Pedagogues.* New York: Routledge, 2007.

[23] J. Dalby and D. Kewley-Port, "Explicit pronunciation training using automatic speech recognition technology," in *Research in technology and second language education: developments and directions*, pp. 379–398, Connecticut: Information Age Publishing, 2005.

[24] A. Neri, C. Cucchiarini, and H. Strik, "Effective feedback on l2 pronunciation in ASR-based CALL," in *the workshop on Computer Assisted Language Learning, Artificial Intelligence in Education Conference, San Antonio, Texas*, p. 40–48, 2001.

[25] J. Brusk, P. Wik, and A. Hjalmarsson, "DEAL: a serious game for CALL practicing conversational skills in the trade domain," in *The Proceedings of SlaTE-Workshop on Speech and Language Technology in Education, Pennsylvania, USA*, 2007.

[26] S. Seneff, C. Wang, and J. Zhang, "Spoken conversational interaction for language learning," in *InSTIL/ICALL Symposium, Venice, Italy*, 2004.

[27] H. Morton and M. A. Jack, "Scenario-based spoken interaction with virtual agents," *Computer Assisted Language Learning*, vol. 18, no. 3, p. 171–191, 2005.

[28] A. Raux and M. Eskenazi, "Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges," in *InSTIL/ICALL Symposium, Venice, Italy*, 2004.

[29] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Human-Computer Interaction*, vol. 19, no. 1, p. 61–84, 2004.

[30] J. Han, M. Jo, S. Park, and S. Kim, "The educational use of home robots for children," in *IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005*, p. 378–383, 2005.

[31] M. H. Long, "The role of the linguistic environment in second language acquisition," *Handbook of second language acquisition*, vol. 2, p. 413–468, 1996.

[32] M. Swain, "The output hypothesis: Theory and research," *Handbook of research in second language teaching and learning*, p. 471–483, 2005.

[33] D. Schneider and K. F. McCoy, "Recognizing syntactic errors in the writing of second language learners," in *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, p. 1198–1204, 1998.

[34] M. Poesio and A. Mikheev, "The predictive power of game structure in dialogue act recognition: Experimental results using maximum entropy estimation," in *Proceedings of ICSLP*, vol. 98, 1998.

[35] D. Bohus and A. I. Rudnicky, "RavenClaw: dialog management using hierarchical task decomposition and an expectation agenda," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[36] H. Ai, A. Roque, A. Leuski, and D. Traum, "Using information state to improve dialogue move identification in a spoken dialogue system," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[37] D. W. Carroll, *Psychology of Language.* Belmont, CA: Wadsworth Publishing, 4 ed., 2003.

[38] A. Ratnaparkhi, *Maximum entropy models for natural language ambiguity resolution.* Ph.d. dissertation, University of Pennsylvania, 1998.

[39] B. V. Dasarathy, *Nearest neighbor (NN) norms: NN pattern classification techniques.* Los Alamitos: IEEE Computer Society Press, 1990.

[40] C. Lee, S. Jung, S. Kim, and G. G. Lee, "Example-based dialog modeling for practical multi-domain dialog system," *Speech Communication*, vol. 51, no. 5, p. 466–484, 2009.

[41] S. Jung, C. Lee, K. Kim, M. Jeong, and G. G. Lee, "Data-driven user simulation for automated evaluation of spoken dialog systems," *Computer Speech & Language*, vol. 23, no. 4, p. 479–509, 2009.

[42] L. N. Michaud, I. Charge-McCoy, and F. Kathleen, *Modeling user interlanguage in a second language tutoring system for deaf users of american sign language.* University of Delaware, 2002.

[43] J. Foster, "Treebanks gone bad: Generating a treebank of ungrammatical english," in *Proc. IJCAI Workshop on Analytics for Noisy Unstructured Data*, 2007.

[44] J. Foster, "Good reasons for noting bad grammar: Empirical investigations into the parsing of ungrammatical written english," *Unpublished doctoral dissertation, Trinity College, University of Dublin*, 2005.

[45] S. Lee and G. G. Lee, "Realistic grammar error simulation using markov logic," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, p. 81–84, 2009.

[46] D. C. McClelland, *Achieving society.* Free Pr, 1967.

[47] C. Cucchiarini, J. van Doremalen, and H. Strik, "DISCO: development and integration of speech technology into courseware for language learning," *Proceedings of Interspeech 2008, Brisbane, Australia, 22-26 September 2008*, p. 2791, 2008.

[48] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks* 1," *Computer Speech & Language*, vol. 14, no. 4, p. 373–400, 2000.

[49] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines, 2001," *Software available at http://www. csie. ntu. edu. tw/cjlin/libsvm*, 2001.

[50] M. Richardson and P. Domingos, "Markov logic networks," *Machine Learning*, vol. 62, no. 1, p. 107–136, 2006.

[51] E. Izumi, K. Uchimoto, and H. Isahara, "Error annotation for corpus of japanese learner english," in *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005)*, p. 71–80, 2005.

[52] S. C. Rhee, S. H. Lee, Y. J. Lee, and S. K. Kang, "Design and construction of Korean-Spoken english corpus," in *Eighth International Conference on Spoken Language Processing*, 2004.

[53] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, *et al.*, "The HTK book (for HTK version 3.4)," *Cambridge University Engineering Department*, vol. 2, no. 2, p. 2–3, 2006.

[54] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Proceedings of the international conference on spoken language processing*, vol. 2, p. 901–904, 2002.

[55] J. Jia, "CSIEC: a computer assisted english learning chatbot based on textual knowledge and reasoning," *Knowledge-Based Systems*, vol. 22, no. 4, p. 249–255, 2009.

[56] P. Seedhouse, "The case of the missing" no": The relationship between pedagogy and interaction," *Language Learning*, vol. 51, no. s1, p. 347–385, 2001.

[57] A. Ferreira and J. Atkinson, "Designing a feedback component of an intelligent tutoring system for foreign language," *Knowledge-Based Systems*, vol. 22, no. 7, p. 496–501, 2009.

[58] Y. Afacan and H. Demirkan, "An ontology-based universal design knowledge support system," *Knowledge-Based Systems*, vol. 24, pp. 530–541, May 2011.

[59] K. Scheffler and S. Young, "Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning," in *Proceedings of the second international conference on Human Language Technology Research*, p. 12–19, 2002.

[60] J. Schatzmann, K. Georgila, and S. Young, "Quantitative evaluation of user simulation techniques for spoken dialogue systems," in *6th SIGdial Workshop on DISCOURSE and DIALOGUE*, p. 2–3, 2005.

[61] J. Wagner, J. Foster, and J. van Genabith, "Judging grammaticality: Experiments in sentence classification.," *CALICO Journal*, vol. 26, no. 3, p. 17, 2009.

[62] J. Wagner, J. Foster, and J. V. Genabith, "A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors," *Proceedings of the joint EMNLP/CoNLL, Prague*, 2007.

[63] J. Foster and \. E. Andersen, "GenERRate: generating errors for use in grammatical error detection," in *Proceedings of the fourth workshop on innovative use of nlp for building educational applications*, p. 82–90, 2009.

[64] P. V. Hentenryck, "The OPL optimization programming language," 1999.

[65] P. V. Roy, P. Brand, D. Duchier, S. Haridi, C. Schulte, and M. Henz, "Logic programming in the context of multiparadigm programming: the oz experience," *Theory and Practice of Logic Programming*, vol. 3, no. 06, p. 717–763, 2003.

[66] K. R. Apt and M. Wallace, *Constraint logic programming using Eclipse*. Cambridge Univ Pr, 2007.

[67] K. Marriott, N. Nethercote, R. Rafeh, P. J. Stuckey, M. G. D. L. Banda, and M. Wallace, "The design of the zinc modelling language," *Constraints*, vol. 13, no. 3, p. 229–267, 2008.

[68] N. Nethercote, P. Stuckey, R. Becket, S. Brand, G. Duck, and G. Tack, "Minizinc: Towards a standard CP modelling language," *Principles and Practice of Constraint Programming–CP 2007*, p. 529–543, 2007.

[69] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics," *Pearson Education*, vol. 56, p. 64–65, 2008.

[70] S. Riedel, “Markov the beast.” `<http://code.google.com/p/thebeast/>`, 2008.

[71] S. Kok, P. Singla, M. Richardson, P. Domingos, M. Sumner, H. Poon, and D. Lowd, “The alchemy system for statistical relational AI,” *Department of Computer Science and Engineering, University of Washington, Technical Report. http://www. cs. washington. edu/ai/alchemy*, vol. 2, no. 6, 2005.

[72] D. Jain and M. Beetz, “Probcog.” `<http://ias.cs.tum.edu/research-areas/knowledge-processing/probcog>`, 2007.

[73] H. Poon and P. Domingos, “Sound and efficient inference with probabilistic and deterministic dependencies,” in *Proceedings of the 21st national conference on Artificial intelligence - Volume 1*, p. 458–463, AAAI Press, 2006. ACM ID: 1597612.

[74] E. Izumi, K. Uchimoto, and H. Isahara, “The overview of the sst speech corpus of japanese learner english and evaluation through the experiment on automatic detection of learners’ errors,” in *Proceedings of Language Resource and Evaluation Conference (LREC)*, p. 1435–1438, 2004.

[75] E. Dagneaux, S. Denness, S. Granger, and F. Meunier, “Error tagging manual version 1.1,” *Centre for English Corpus Linguistics, Universite Catholique de Louvain*, 1996.

[76] Y. Tono, “The role of learner corpora in SLA research and foreign language teaching: The multiple comparison approach,” *Unpublished Ph. D. Thesis. Lancaster University, UK*, 2002.

[77] M. Sumner and P. Domingos, “The alchemy tutorial,” 2010.

[78] D. Klein and C. D. Manning, “Accurate unlexicalized parsing,” in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, p. 423–430, 2003.

[79] I. Zukerman and D. W. Albrecht, “Predictive statistical models for user modeling,” *User Modeling and User-Adapted Interaction*, vol. 11, no. 1, p. 5–18, 2001.

[80] C. Cucchiarini, A. Neri, and H. Strik, “Oral proficiency training in dutch l2: The contribution of ASR-based corrective feedback,” *Speech Communication*, vol. 51, p. 853–863, Oct. 2009. ACM ID: 1576974.

[81] H. Strik, J. van de Loo, J. van Doremalen, and C. Cucchiarini, "Practicing syntax in spoken interaction: Automatic detection of syntactical errors in non-native utterances," in *Proceedings of the SLaTE 2010 Workshop*, (Tokyo, Japan), Sept. 2010.

[82] B. Friedman, P. H. K. Jr, and J. Hagman, "Hardware companions?: What online AIBO discussion forums reveal about the human-robotic relationship," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 273–280, 2003.

[83] M. Fujita, "AIBO: toward the era of digital creatures," *The International Journal of Robotics Research*, vol. 20, no. 10, p. 781, 2001.

[84] K. Wada, T. Shibata, T. Saito, and K. Tanie, "Analysis of factors that bring mental effects to elderly people in robot assisted activity," in *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, vol. 2, p. 1152–1157, 2002.

[85] D. H. Ahn and M. Chung, "One-Pass Semi-Dynamic network decoding using a subnetwork caching model for large vocabulary continuous speech recongnition," *IEICE TRANSACTIONS on Information and Systems*, vol. 87, no. 5, p. 1164–1174, 2004.

[86] S. Goronzy, *Robust adaptation to non-native accents in automatic speech recognition*. Springer-Verlag New York Inc, 2002.

[87] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *the workshop on Speech and Natural Language, Harriman, NY*, p. 357–362, 1992.

[88] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer speech and language*, vol. 9, no. 2, p. 171, 1995.

[89] G. Zavaliagkos, R. Schwartz, and J. McDonough, "Maximum a posteriori adaptation for large scale HMM recognizers," in *IEEE International Conference on Acoustics, Speech and Signal Processing, Atlanta, Georgia*, vol. 2, p. 725–728, 1996.

[90] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd, "GUS, a frame-driven dialog system," *Artificial intelligence*, vol. 8, no. 2, p. 155–173, 1977.

[91] S. M. Shieber, *An introduction to unification-based approaches to grammar.* PhD thesis, CSLI. Lecture Notes No.4, CSLI, Stanford, 1986.

[92] R. R. Burton, *Semantic grammar: a technique for efficient language understanding in limited domains.* Ph.D. dissertation, University of California, Irvine, 1976.

[93] S. Lee, C. Lee, J. Lee, H. Noh, and G. G. Lee, "Intention-based corrective feedback generation using context-aware model," in *Proceedings of International Conference on Computer Supported Education, Valencia, Spain*, 2010.

[94] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *the Third ACL Workshop on Very Large Corpora, MIT, Cambridge, Massachusetts, USA*, p. 82–94, 1995.

[95] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning, Williams College, Williamstown, MA, USA*, p. 282–289, 2001.

[96] D. Bohus and A. I. Rudnicky, "The RavenClaw dialog management framework: Architecture and systems," *Computer Speech and Language*, vol. 23, no. 3, p. 332–361, 2009.

[97] D. Bohus and A. I. Rudnicky, "Constructing accurate beliefs in spoken dialog systems," in *IEEE Automatic Speech Recognition and Understanding Workshop, San Juan, Puerto Rico*, 2005.

[98] D. Bohus and A. I. Rudnicky, "A k hypotheses + other belief updating model," in *AAAI Workshop on Stochastic Methods in Spoken Dialog Systems, Boston, MA*, 2006.

[99] S. J. Cowley and K. MacDorman, "Simulating convesations: The communion game," *AI & Society*, vol. 9, no. 2, p. 116–137, 1995.

[100] R. A. Hinde, *Individuals, relationships & culture: Links between ethology and the social sciences.* New York: Cambridge Univ Press, 1987.

[101] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, p. 297–334, 1951.

[102] R. L. Rosnow and R. Rosenthal, *Beginning Behavioral Research: A Conceptual Primer.* Upper Saddle River, NJ: Prentice Hall, 6 ed., May 2007.

[103] M. Kutner, C. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*. New York: McGraw-Hill/Irwin, 5th ed., Aug. 2004.

[104] K. A. Petersen, *Implicit Corrective Feedback in Computer-Guided Interaction: Does Mode Matter?* Ph.D. dissertation, Georgetown University, 2010.

[105] Z. Handley, "Is text-to-speech synthesis ready for use in computer-assisted language learning?," *Speech Communication*, vol. 51, no. 10, p. 906–919, 2009.

[106] S. D. Krashen, "Explorations in language acquisition and use: The taipei lectures," *TESL-EJ*, vol. 7, no. 2, p. 96, 2003.

Acknowledgements
감사의 글

2007년의 시작과 함께 6년간의 회사 생활을 뒤로하고 다시 학생이 되어 포항으로 돌아왔습니다. 당장 세상에서 감당해야할 많은 짐들이 여전히 있었지만 회사 생활 동안 내내 끊임없이 일어나던 많은 질문을 던지기 위해 다시 돌아올 수 밖에 없었습니다. 이 질문들은 비단 컴퓨터 공학에만 한정된 것은 아니었습니다. 우리의 사회를 더 깊이 알고 싶었고 또한 우리의 삶을 더 알고 싶었습니다. 인간을 좀 더 이해해 보고자 언어를 공부하기로 선택하였습니다. 4년 반이 지난 지금 언어 처리로 박사 학위를 받게 되었지만 그 당시 품었던 질문들에 대한 답은 아직도 얻지 못했습니다. 아니 그런 답들에 우리가 얼마나 멀리 있는 지를 알게 되었습니다. 그런데 이 깨달음이 은혜였습니다. 계속해서 찾아 헤매지만 길을 찾지 못하는 인간의 본질을 보게 되었습니다.

“예수께서 그들을 보시며 이르시되 사람으로는 할 수 없으나 하나님으로서는 다 하실 수 있느니라” (마태복음 19:26). 꼬리에 꼬리를 무는 질문들 과학, 경제, 사회, 종교, 또는 인간에 관한 것이든... 그 어떤 질문에도 제대로 답할 수 없었습니다. 제가 인간이기 때문입니다. 제가 답하려고 해서는 답이 없습니다. 오직 그분에게 답을 구해야 한다는 것을 깨달은 것이 바로 제가 체험한 은혜입니다. 이것이 저의 학위입니다. 저의 지난 시간은 다 이것을 위해 쓴 셈입니다. 주님 이 학위는 오로지 당신의 것입니다.

불꽃같이 해가 내리 쬐던 2007년 7월 7일에 그녀의 표현대로면 털끝만큼의 망설임도 없이 저의 아내가 되어 준 정은이의 농도 100%의 사랑과 믿음 그리고 도움에 가슴 깊이 감사를 드립니다. “내가 그를 위하여 돕는 배필을 지으리라 하시니라” (창세기 2:18). 우리는 지난 4년 반 동안 많은 일들을 함께 헤쳐왔습니다. 같이 북한 경제 연구를 하여 아내가 석사 학위를 받았고 세상의 그 어떤 보석보다 밝은 우리의 두 아이 예인과 혜성을 얻었습니다. 제가 힘들 때, 중요한 일이 있을 때, 멀리 출장을 다녀 올 때마다 받아 든 아내의 편지는 세상에서 제일 큰 위로와 격려였습니다. 저는 답장조차 그 만큼 하지 못해서 미안합니다. 이 글이 조금이라도 대신할 수 있기를 바랍니다. 우리는 분명 하나님이 예비하신 돕는 배필임을 감사합니다. 앞으로 가야할 길도 당신과 함께여서 용감한 마음으로 기대할 수 있습니다. 사랑합니다.

아버지 어머니 무슨 말씀을 드려야 할 지 모르겠습니다. 다른 어떤 집이 우리와 같이 깊이 삶을 나누었을까요. 함께 길을 걸었을까요. 그 분의 손 아래서 부모님 같은 스승을 평생 둘 수 있는 것이 저의 복이지요. 지나온 수 많은 삶의 굴곡 가운데 기도와 말씀으로 함께 이겨나온 것이 저의 자랑이며 든든한 버팀목입니다. 날마다

한결같은 사랑으로 기도해주신 그 은혜 감사합니다. 그 덕으로 제가 이 과정을 마칠 수 있었습니다.

장인어른 장모님 감사드립니다. 아무 것 없는 학생 사위를 반갑게 맞아주시고 바쁘다고 잘 찾아뵙지 못한 점도 너그러이 이해해주셨습니다. 늘 저희의 부족함을 먼저 채워주시고 따뜻한 말씀을 해주시고 안아주신 것을 잊을 수 없을 것입니다.

ISoft 연구실 식구들에게 감사합니다. 먼저 저의 연구 주제를 정해 주시고 많은 기회를 만들어 주신 이근배 교수님께 감사드립니다. 첫 룸메이트이자 성실한 연구자세를 보여준 민우, 좋은 연구 안목을 가진 발표의 달인 상근, 늘 밝고 명랑하게 랩을 만들어준 청재와 성희, 나를 랩에서 수면 위로 부상시킨 경덕, 엄청난 말솜씨를 자랑하는 석환, 화이트닝의 달인 진식, JH21983 진리교를 접하게 해준 종훈, 새침함 속의 순수함을 간직한 형종, 귀여운 야구광 동현, 박학 다식 파코신 지민, 열정적인 스마일맨 규송, 자유인 준휘, 순진무구 어리광 총, 색깔있는 능력자 홍석, 듬직한 인재, 믿음과 정이 넘치는 영선씨까지... 참 즐거운 시간이었습니다. 잊을 수 없는 수많은 기억들을 선물해주셔서 감사합니다. 제 대학원 생활의 소중한 액자에 당신들을 그려넣겠습니다.

그리고 마지막으로 화평교회 식구들 보환, 경란, 충효, 은정, 민규, 옥진, 철준, 수현, 성녕, 유라 너희들과 함께한 공동체가 내게 큰 힘이 되었어. 주님이 가르치신 삶은 공동체가 없다면 이루어질 수 없는 것이기에 너희들은 나의 삶에서 그 무엇보다도 소중한 존재란다. 사랑한다.

Curriculum Vitae

Research Interest

Computer-Assisted Language Learning, Spoken (and Multimodal) Dialog Systems, Error Handling in Spoken Dialog Systems, Machine Learning, Natural Language Processing, Spoken Language Understanding, Automated English Assessment, Educational Data Mining, User Modeling, Emotion Recognition.

Education

- **Ph.D. / M.S. Computer Science and Engineering** **POSTECH**
  September, 2007 – Present
  Thesis: Intelligent Corrective Feedback for Communicative Computer-Assisted Language Learning

- **B.S. Computer Science and Engineering** **POSTECH**
  March, 1997 – February, 2006

Work Experience

- **PIOLINK** **Jun 2004 – Jan 2007**

  PIOLINK is a leading application networking company in Korea. It manufactures Application Switches, Network Load Balancers, and Web Security Switches. I worked as a senior team member for the development and design of multi processors and multitasking systems. I implemented and managed firmware and software for embedded Linux systems mixing PowerPC and MIPS. I wrote numerous device drivers including communications processor drivers (BroadCom BCM1250, BCM1480), SSL chip drivers (Britestream BN1010; Cavium CN1000), network chip drivers (BroadCom BCM5690, BCM5464SR), and peripheral device drivers (PCI, HT, System Monitoring chip, RTC & NVRAM, Flash memory, UART).

● **CORECESS** **Jan 2001 – May 2004**

CORECESS manufactures telecommunications equipment for broadband access networks such as Optical Link Technologies (GEPON and WDM PON), Intelligent Mutilayer Switches, VDSL, and DSLAM. I worked on the development and design of quality of service (Corecess QoS architecture design and implementation), layer 2 network protocols (RSTP, LACP), and a Logging File System. I implemented and managed system software for embedded Linux and pSOS. I wrote numerous device drivers including programmable network processor drivers (Agere APP550), network chip drivers (Switchcore CXE 1000, CXE16; Galileo Galnet II, Galnet II+, Galnet III), and peripheral device drivers (PCI, IIC, System Monitoring chip, RTC & NVRAM, Flash memory, UART).

***R**esearch **E**xperience*

● Selected Project Experience (Project title / Term / Funding Institute)

1. Research on English Speaking Assessment and Provision of Feedback for Korean, May 2010 – Present, SK Telecom Corporation:
   I participated in making the project proposal and drawing the system architecture. In particular, I am developing a component for grammaticality judgment and provision of feedback for oral output.

2. Research Laboratory for Natural Language-based Immersive English Tutoring System, Mar 2010 – Present, MEST (Ministry of Education Science and Technology):
   I am responsible for project management and participated in designing the system architecture consisting of various technologies such as Speech, Vision, and Haptic. I am developing dialog strategies in consideration of students' proficiency level, emotion, and gameplay.

3. Development of Mobile Platforms for Dialog-based Speech Interfaces, Mar 2010 – Present, MKE (Ministry of Knowledge and Economy):
   I am responsible for investigating dialog management for English tutoring. I am working on data collection, robust language understanding, dialog management, and corrective feedback generation. I am researching a ASR combination to generate feedback on both global and local errors.

4. Development of Intelligent Robots for English Conversation Tutoring, Aug 2009 – Present, MKE (Ministry of Knowledge and Economy):
   I am responsible for speech and language processing of intelligent robots. I collaborated with English teachers to make educational material and developed communicative robots capable of providing recast feedback in response to students' errors. I participated in a pilot project for elementary students.

5. Conversational Agent for English Conversation Tutoring July 2008 – July 2009, MSRA (Microsoft Research Asia Funded Project):
   I was responsible for project management. I designed and implemented a hybrid language understanding component for robust language understanding and corrective feedback

generation.

6. Research on Dialog-based Computer Aided Language Learning for Spoken Dialog System in English, Mar 2007 – Nov 2007, KT Corporation:
   I was responsible for project management. I designed and implemented a spoken dialog system for English conversation practice in the immigration domain. I developed a method to provide recast feedback and suggest expressions in the case of timeout.

Management Experience

- **Laboratory Director**
  - Intelligent Software Laboratory, POSTECH 2010
- **Team Leader**
  - Intelligent Computer-Assisted Language Learning (ICALL) Team 2007 - 2010

Honors and Awards

- Award of Excellence (EIFIT-08-19) from Brain Korea 21 of the Korea Research Foundation 2009
- Award of Excellence (EIFIT-08-18) from Brain Korea 21 of the Korea Research Foundation 2009
- Award of Excellence from Knowledge Information Processing and Ontology Challenge 2008
- Best Paper Award, Joint Conference of the Korean Society of Phonetic Sciences and Speech Technology and the Korean Association of Speech Sciences 2008
- Best Paper Award, Conference on Human & Cognitive Language Technology 2007

Patents

● International Patents

1. Sungjin Lee, Gary Geunbae Lee. *Grammar Error Simulation Apparatus and Method*. PCT Application, 2010

2. Sungjin Lee, Gary Geunbae Lee. *Apparatus for Foreign Language Learning and Method for Providing Foreign Language Learning Service*. PCT Application 2010

3. Hyungjong Noh, Sungjin Lee, Jonghoon Lee, Gary Geunbae Lee. *Document Processing Apparatus and Method for Extraction of Description*. PCT Application 2010

● Domestic Patents

1. Sungjin Lee, Hyungjong Noh, Kyusong Lee, Gary Geunbae Lee, *Method and Apparatus for Grammatical Error Detection for Conversational Agents for Foreign Language Conversation Tutoring*, Korea Application, 2011

2. Hyungjong Noh, Sungjin Lee, Kyusong Lee, Gary Geunbae Lee, *Method of Managing Communication and System for the Same*, Korea Application, 2011

3. Kyusong Lee, Sungjin Lee, Jonghoon Lee, Hyeongjong Noh, Gary Geunbae Lee. *Foreign Language Learning Game System and Method based on Natural Language Dialogue Technology.* Korea Application, Apr 2010

4. Sungjin Lee, Sangkeun Jung, Gary Geunbae Lee. *Grammar Error Simulation Apparatus and Method*. Korea Application, Dec 2009

5. Sungjin Lee, Cheongjae Lee, Gary Geunbae Lee. *Apparatus for Foreign Language Learning and Method for Providing Foreign Language Learning Service*. Korea Application, Dec 2009

6. Hyungjong Noh, Sungjin Lee, Jonghoon Lee, Gary Geunbae Lee. *Document Processing Apparatus and Method for Extraction of Description*. Korea Application, Oct 2009

7. Jonghoon Lee, Sungjin Lee, Hyungjong Noh, Gary Geunbae Lee. *Document Processing Apparatus and Method*. Korea Application, Sep 2009

8. Sungjin Lee, Sangkeun Jung, Kyungduk Kim, Cheongjae Lee, Gary Geunbae Lee. *Foreign Language Conversation Training Method and Apparatus and Trainee Simulation Method and Apparatus for Qucikly Developing and Verifying The Same*. Korea Application, Jun 2009

9. Cheongjae Lee, Sungjin Lee, Sangkeun Jung, Kyungduk Kim, Jinsik Lee, Gary Geunbae Lee. *Speech-interfaced Dialog Systems*. Korea Application, Apr 2009

10. Sungjin Lee, Minwoo Jeong, Cheongjae Lee, Gary Geunbae Lee. *Example-based Communicating System for Foreign Conversation Education and Method Therefor*, Korea Application, Dec 2007

Publications

● Journal Publications

1. Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee, Seongdae Sagong, Moonsang Kim. (2011) *On the Effectiveness of Robot-Assisted Language Learning*, ReCALL Journal, Vol.23(1), SSCI.

2. Sungjin Lee, Jonghoon Lee, Hyungjong Noh, Kyusong Lee, Gary Geunbae Lee. (2011) *Grammatical Error Simulation for Computer-Assisted Language Learning*, Knowledge-Based Systems, SCI

3. Jonghoon Lee, Sungjin Lee, Hyungjong Noh, and Gary Geunbae Lee. (2011) *Iteratively Constrained Selection of Word Alignment Links from Knowledge and Statistics*, Knowledge-Based Systems, SCI

4. Sungjin Lee, Hyungjong Noh, Kyusong Lee, Gary Geunbae Lee. (2011) *Foreign Language Tutoring in Oral Conversations Using Spoken Dialog Systems*, IEICE Trans. on Information & Systems, SCIE (**invited**)

5. Sungjin Lee, Cheongjae Lee, Gary Geunbae Lee. (2010) *Example-based Dialog System for English Conversation Tutorin*g. Journal of KIISE: Software and Applications, Vol. 37(2) .129-136.

6. Jinsik Lee, Sungjin Lee, Jonghoon Lee, Byungchang Kim, Gary Geunbae Lee. *Incorporating Linguistically Motivated Features into Two-step Prediction Model for Korean Prosodic Phrasing*, ACM Transactions on Asian Language Information Processing (**under review**)

● Conference/Workshop/Symposium Publications

1. Hyungjong Noh, Sungjin Lee, Kyusong Lee, Gary Geunbae Lee, *Ranking Dialog Acts using Discourse Coherence Indicator for Language Tutoring Dialog Systems*, Proceedings of the international workshop on spoken dialog systems (IWSDS 2011), Granada, Sep 2011

2. Sungjin Lee, Hyungjong Noh, Kyusong Lee, Gary Geunbae Lee, *Grammatical Error Detection for Corrective Feedback Provision in Oral Conversations*, Proceedings of the AAAI 2011: Twenty-Fifth Conference on Artificial Intelligence, San Francisco, Aug 2011

3. Hyungjong Noh, Kyusong Lee, Sungjin Lee, Gary Geunbae Lee, *POMY: A Conversational Virtual Environment for Language Learning in POSTECH*, Proceedings of The 12th annual

SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2011), Portland, June 2011

4. Kyusong Lee, Soo-Ok Kweon, Sungjin Lee, Hyungjong Noh, Gary Geunbase Lee, *Effects of Language Learning Game on Korean Elementary School Student*, Proceedings of the ISCA SIG on Speech and Language Technology in Education (SLaTE 2011), Venice, Aug 2011

5. Sungjin Lee, Changgu Kim, Jonghoon Lee, Hyungjong Noh, Kyusong Lee, Gary Geunbae Lee. *Affective Effects of Speech-enabled Robots for Language Learning*. Proceedings of the 2010 IEEE Workshop on Spoken Language Technology (SLT 2010), Berkeley, Dec 2010

6. Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee. *POSTECH Approaches for Dialog-based English Conversation Tutoring*. Proceedings of the 2010 APSIPA annual summit and conference, Singapore, Dec 2010 (**nominated for student best paper**)

7. Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee. *Cognitive Effects of Robot-Assisted Language Learning on Oral Skills*. Proceedings of Interspeech Second Language Studies Workshop, Tokyo, Sep 2010.

8. Kyusong Lee, Sungjin Lee, Jonghoon Lee, Hyeongjong Noh, Gary Geunbae Lee. *Natural Language-based Immersive English Tutoring System.* Proceedings of Conference on Human & Cognitive Language Technology, Gwangju, Oct 2010.

9. Kyusong Lee, Sungjin Lee, Jonghoon Lee, Hyeongjong Noh, Gary Geunbae Lee. *Natural Language Dialog-based Language Learning Game Platform*. Proceedings of the Korean Society of Speech Sciences. Daejeon, May 2010.

10. Sungjin Lee, Cheongjae Lee, Jonghoon Lee, Hyungjong Noh, Gary Geunbae Lee. *Intention-based Corrective Feedback Generation using Context-aware Model*. Proceedings of the International Conference on Computer Supported Education, Valencia, Spain, Apr 2010.

11. Hyungjong Noh, Minwoo Jeong, Sungjin Lee, Jonghoon Lee, Gary Geunbae Lee. *Script-Description Pair Extraction from Text Documents of English as Second Language Podcast.* Proceedings of the International Conference on Computer Supported Education, Valencia, Spain, Apr 2010.

12. Cheongjae Lee, Sungjin Lee, Sangkeun Jung, Kyungduk Kim, Donghyeon Lee, Gary Geunbae Lee. *Correlation-based Query Relaxation for Example-based Dialog Modeling*. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (IEEE-ASRU), pages 474-478, Merano, Italy, Dec 2009.

13. Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Gary Geunbae Lee. *Importing Human Tutor's Conversation Strategy into Dialog Systems for Language Learning using Fuzzy Logic*. Proceedings of the international workshop on spoken dialog systems (IWSDS2009), Germany, Dec 2009

14. Sungjin Lee, Gary Geunbae Lee. *Realistic Grammar Error Simulation using Markov Logic*. Proceedings of the ACL 2009, Singapore, August 2009

15. Cheongjae Lee, Sangkeun Jung, Kyungduk Kim, Sungjin Lee, Gary Geunbae Lee. *Automatic agenda graph construction from human-human dialogs using clustering method*. Proceedings of the Korean Society of Speech Sciences. Jeonju, May 2009.

16. Sungjin Lee, Gary Geunbae Lee. *Recognizing Language Learners' Intention using Discourse Information*. Joint Conference of the Korean Society of Phonetic Sciences and Speech Technology and the Korean Association of Speech Sciences, Seoul, Nov 2008

17. Sungjin Lee, Cheongjae Lee, Gary Geunbae Lee. *Example-based dialog modeling for English conversation tutoring*. Proceedings of the 2nd International Conference on Next Generation Computing (NGC), Seoul, Nov 2007.

18. Sungjin Lee, Cheongjae Lee, Gary Geunbae Lee. *Example-based dialog modeling for Computer-Assisted Language Learning*. Proceedings of Conference on Human & Cognitive Language Technology, Daegu, Oct 2007.

***P**rofessional **A**ctivities*

• Journal Reviewing

- Knowledge-Based Systems, SCI — 2010 - 2011
- Journal of Computing Science and Engineering — 2011

• Conference Reviewing

- International Joint Conference on Natural Language Processing — 2011

• Organizing Committee

- Young Researchers' Roundtable for Spoken Dialog Systems — 2011

• Talks

- POSTECH approaches for dialog-based English conversation tutoring (Invited Talk)
  World Voice Congress, Joint Workshop of Phonetics and Engineering, Seoul, Korea — 2010

***T**eaching **E**xperience*

•Teaching Assistant — POSTECH

Fall, 2010

CSED442: Artificial Intelligence (Undergraduate Level)

●**Teaching Assistant** **POSTECH**
Spring, 2007
CSED101: Introduction to Computing (Undergraduate Level)