Computer Assisted Language Learning
Vol. 18, No. 3, July 2005, pp. 171–191

Routledge
Taylor & Francis Group

Scenario-Based Spoken Interaction with Virtual Agents

Hazel Morton* and Mervyn A. Jack
The University of Edinburgh, Scotland

This paper describes a CALL approach which integrates software for speaker independent continuous speech recognition with embodied virtual agents and virtual worlds to create an immersive environment in which learners can converse in the target language in contextualised scenarios. The result is a self-access learning package: SPELL (Spoken Electronic Language Learning). The SPELL system has been implemented to run in real time on a standard desktop PC. A prototype of SPELL is in user testing with learners of Italian and learners of Japanese in five high schools in Scotland. The speech recogniser is programmed to recognise grammatical and some ungrammatical utterances so that the learner can receive feedback on their language production. The dialogues can be modified in cases of communication difficulties, and are reactive to the learner's spoken input so that the learner receives relevant and immediate feedback to their utterance. Feedback takes two key forms: *reformulations*, where the system modifies the initial input, and *recasts*, where the system repeats the learner's utterance, implicitly correcting any errors. The SPELL system will offer a test bed for the effectiveness of the feedback within a self-access virtual learning environment.

Introduction

Chapelle (1997) suggests that designers of computer-assisted language learning (CALL) programmes should look more towards the theories which drive instructed second language acquisition (SLA) research: there is a need for innovations in the CALL field to be driven by pedagogical principles rather than technical ones. The present paper discusses the use of a theory of SLA, namely the *interaction hypothesis* (Long, 1996), as a foundation for the design of a new multimedia CALL approach employing automatic speech recognition, embodied virtual agents and virtual worlds for the creation of scenarios in which learners can develop oral skills in their target language (L2).

*Corresponding author. Centre for Communication Interface Research, The King's Buildings, The University of Edinburgh, Edinburgh EH9 3JL, UK. Email: hmorton@ccir.ed.ac.uk

ISSN 0958-8221 (print)/ISSN 1744-3210 (online)/05/030171–21

DOI: 10.1080/09588220500173344

According to the interaction hypothesis, learners need to interact in the L2 in such a way that the interactions can be modified in instances of communicative difficulty. Interaction provides learners with opportunities to receive comprehensible input and feedback (Gass, 1997; Long, 1996; Pica, 1994). It also allows learners to make changes to their own linguistic output (Swain, 1985, 1995). Interaction between learners and their interlocutors forces learners to test their hypotheses about the structure of the language. Furthermore, the input provided to the learner in the interaction may be above their current level, which may prompt the learner to ''notice the gap'' (Schmidt & Frota, 1986, p. 311), and in some way seek modification of the input.

These modifications can provide the learner with the additional input they need for their L2 development, and push them to make more target-like utterances. Long (1996) claims that negotiation of meaning ''facilitates acquisition because it connects input, internal learner capacities, particularly selective attention, and output in productive ways'' (p. 452). Negotiation of meaning can be made through input modifications such as repetitions, confirmation checks, clarification requests, reformulations, comprehension checks and recasts. This is further exemplified by the fact that empirical research relating to classroom interactions has demonstrated the importance of conversational interaction in the development of the L2. Mackey (1999) found that learners who actively participated in an interaction produced more advanced structures than those learners whose participation in the interaction was less active. Further, conversational interaction between interlocutors provides opportunities in which learners can receive feedback on their utterances directly within the ongoing discourse, with implicit reactive feedback being offered to the learner.

Long (1991) makes a distinction between what he calls focus-on-forms, discrete point, step-by-step grammar instruction, and focus-on-form, corrective feedback which is fully integrated within ongoing communicative activities. A number of studies have compared learners' language development in communicative language teaching (CLT) without focus-on-form to that which is achieved in CLT with focus-on-form. The results of these studies have provided strong support for the inclusion of focus-on-form in the CLT classroom.

Recasting is a focus-on-form strategy used in the language classroom. For oral language development, recasting provides implicit corrective feedback to the learner in an unobtrusive way, so that the correction does not interfere with the ongoing discourse. Recasts are a reactive source of information which are specific to the individual learner, whose original meaning is taken and restated by the interlocutor with some aspect (phonological, syntactic, lexical, etc) being modified. In cases of one-to-one interactions, it has been shown that focused recasts are beneficial to the learner's development (Long, Inagaki, & Ortega, 1998; Mackey & Philip, 1998).

Although much of this research has been conducted in the realm of human–human interaction, either between a native speaker and a non-native speaker or between two non-native speakers, the work described in this present paper advocates applying this theory as a central principle in learner-computer interaction. Previous CALL studies have analysed learner interactions using the interaction hypothesis where learners interact with other learners or native speakers in network-based

communication (Blake, 2000; Kitade, 2000). In contrast, the approach described here (SPELL) draws on the interaction hypothesis to create situations in which a learner can engage in meaningful spoken interactions with the computer, and whose interactions can be negotiated in order that the learner can develop their oral language in the L2. Further, the system is reactive to the learner's spoken input so that the learner receives relevant and immediate feedback to their utterance.

The SPELL approach combines virtual worlds and virtual agents with automatic speech recognition technology to create a language learning application in which learners can converse in the target language and receive feedback from the agents. The following sections give an account of the technologies used within the project design, and a detailed description of the SPELL system.

The Technologies

Virtual Worlds

Virtual reality has been defined as ''an event or entity that is real in effect but not in fact'' (Heim, 1994, p. 29). The virtual worlds presented to the learner in SPELL offer a highly contextualised environment in which the learner can first observe the interactions between the animated agents and then can enter the environment as an active dialogue participant. In a manner that is similar to role-playing exercises often used in the language classroom, the learner can, for example, become a customer in a café. SPELL therefore uses virtual worlds in creating situations in which the learner can engage with other interlocutors within a context that pre-determines the kinds of spoken interaction that will occur.

In their use of virtual environments, users may experience *presence*, that is, the subjective sense of ''being there'' in the virtual world (Slater, Usoh, & Steed, 1994). The underlying assumption is that if users experience such a sense of presence in a virtual environment they will come to behave in the virtual environment in a way that is similar to the way they would behave in a similar environment in the real world. Thus, if learners have the opportunity to communicate in the virtual environment, the skills learned there are likely to carry over to similar situations in the real world. Virtual environments offer features which are superior to video presentations because of the sense of presence in the environment created through the manipulation of certain aspects of that environment.

In the SPELL approach described here, the user's presence is suggested through the use of predefined viewpoints in the scene and is re-enforced by the gaze of the agents. SPELL is able to mimic the learner moving around in the scene by the movement of the camera. The viewpoint of the scene is always from the learner's perspective (as if from the learner's eyes). The scene does not depict an embodied representation of the learner; however, at certain times the learner is able to see their virtual hand, for example whenever another agent in the scene hands something to them. Figure 1 shows the learner receiving a menu from the virtual waiter in the ''at the café'' scenario.

Within this multi-agent environment, it is apparent who is being addressed at any one time by the gaze of the virtual agents. The user is directly addressed by the agents, and the agents are able to hand items within the scene to the implied body of the learner. Further, the learner's interaction with the virtual agents in the environment has consequences with respect to subsequent actions by the agents, just as the interactions between humans in the real world can affect their subsequent actions.

Animated Agents

Animated agents as graphical representations of characters are being increasingly used in computer interfaces to offer a more personalised interaction between human and computer. Animated agents have been created for a variety of applications such as a virtual presenter (Noma & Badler, 1997), a virtual real estate agent, (Cassell, Bickmore, Billinghurst, et al., 1999), as training agents for the training peace corps (Marsella, Gratch, & Rickel, 2003) and as retail agents (McBreen, 2002). Research has shown (André, Rist, & Müller, 1999) that adult users of an application overwhelmingly prefer a system which employed an animated agent to a similar system without the agent.

Animated agents have also been used in pedagogical applications. Johnson, Shaw, and Ganeshan (1998) define animated pedagogical agents as ''lifelike characters that

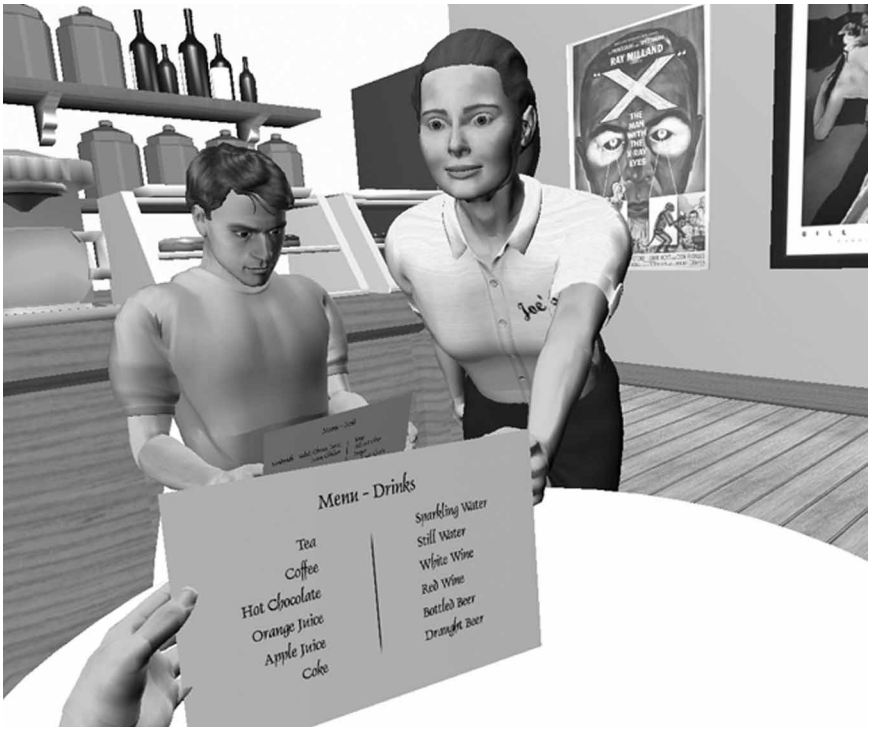

Figure 1. The learner receives a menu from the virtual waiter agent (interactive scenario)

facilitate the learning process'' (p. 2). Johnson and Shaw (1997) used a 2-D animated agent to support students as they work through medical problem solving activities in a Web-based learning environment. Johnson, Rickel, and Lester (2000) also used a 3-D animated agent immersed in a simulated virtual world as a teaching aid for engineering students. Early research in the use of animated agents in pedagogical applications has shown such agents to be effective in tutoring systems in which they can improve the learning experience by engaging students in effective conversations with the agent (Lester, Converse, Kahler, et al., 1997a). Massaro (1998) found that using an animated face in an application for the hearing impaired could improve the learning experience.

In related research by Lester, Converse, and Kahler, et al. (1997b), school students using an application with an animated ''bug'' generally showed improvement in post-experience tests. Interestingly the study also revealed a *persona* effect of having the animated agent in which the agent positively affected the learner's perception of their own learning experience by its ''presence''.

Johnson et al. (2000) believe that a learning environment can be enhanced with the addition of animated agents. They detail a number of types of interaction that agents can display which benefit the learning context. Each agent would not need to exhibit all of these types of interaction, nor would each of these be applicable to pedagogical agents for language learning. However, this summary serves as a general account of the benefits that animated pedagogical agents can bring to human–computer interaction:

- Animated agents employed in simulated worlds provide new opportunities to demonstrate to students how to perform a particular task in that world. They can physically show how to complete a task.
- Animated agents can act as navigational guides to students in complex environments.
- Animated agents can steer a student's attention to something in the virtual world (for example, features which give hints or extra information to the student) by way of gaze behaviours and deictic gestures.
- Animated agents can provide non-verbal feedback to a student's input or actions, as well as verbal feedback.
- Animated agents are able to express conversational signals which people are accustomed to in human–human interaction, for turn taking, expressing personal opinions or acknowledging the user's utterance.
- Animated agents can convey emotion to the user which in turn may elicit emotion from the user and serve to increase learning motivation.
- Animated agents can serve the role of a virtual teammate, where working in a team is an element of the task design. Here, the agent can either act as an instructor, helping the student to accomplish a task, or substitute as a missing team member, allowing the student to practice working in a team.

The use of animated agents within the contextualised virtual world used in SPELL offers the learner an opportunity for one-to-one conversation, designed to contribute

to an enhanced learning experience. Animated pedagogical agents have been shown to ''increase the computer's ability to engage and motivate students'' (Johnson et al., 2000, p. 48). In the context of CALL, Chapelle (1998) suggests that it may be important for learners to have an audience for their linguistic output so that the learners can ''attempt to use the language to construct meanings for communication rather than solely for practice'' (p. 24). In this way, animated pedagogical agents could serve as the cyber audience for language learners' output.

Virtual agents which are able to show affective behaviours, have expressive behaviours, exhibit a personality, speak to the user and listen to the user can offer a further dimension to the experience of human-computer interaction. Indeed, it has been found that users of new media treat the machines they interact with in a social way (Reeves & Nass, 1996). The construction of animated agents makes it possible to create another dimension within the communicative setting—delivering a capability where the spoken output can be enhanced by expressive agents. The point of departure for the approach described in this paper is the proposal that this new technology can be used in the development of effective pedagogical applications for foreign language learning.

Speech Recognition

As a simple description, speech recognition uses statistical approaches in order to recognise speech. The recogniser first detects a speech signal, which is then processed, by making reference to internal phoneme models, into a string of words. These word hypotheses are then evaluated by means of the language models which hold the possible occurrences of strings of words. For a more detailed account of speech recognition, see Ehansi and Knodt (1998).

The value and relevance of speech recognition technology in CALL has been the subject of debate. Derwing, Munro, and Carbonaro (2000) found that a commercial speech recognition package for dictation which performed acceptably for native speakers (90% accuracy) did not perform to acceptable levels with non-native speakers. However, as Neri, Cucchiarini, and Strik (2003) point out, standard speech recognition packages are not designed for non-native speakers. Pronunciation CALL programs which make use of speech recognition sometimes use the information to show learners their speech waveform in order to allow comparison with a model. However, ''good'' pronunciations can look quite different from the ''model'' especially if they are not produced at the same speed. In reviewing a CALL product which took this approach, Miura (2002) commented: ''the use of sound waves and pitch curves is especially problematic. In order to make the comparison, words and phrases need to be pronounced at exactly the same speed as the models''. The use of comparing waveforms as a means of improving oral skills in a second language is not a beneficial learning method. Further studies have found problems with the pronunciation evaluations made by the speech recognisers. Wildner (2002) found that the input from native speakers was sometimes scored lower than that of non-native speakers. Miura (2002) found that the pronunciation evaluation given by the

system differed from those given by a native speaker. In view of these limitations, the waveform matching approach has not been used as a means of feedback in SPELL.

Hincks (2002) conducted research on a commercial language learning product, *Talk to Me*, for the development of the pronunciation for immigrant learners of English in Sweden. These results showed that although the learners reported high satisfaction with using the software overall there was little improvement in pronunciation for the sample group as a whole. However, those students who initially had strongly accented speech improved considerably in their pronunciation post-test.

Some studies have stated the benefits of using automatic speech recognition in CALL applications. Eskenazi (1999) found that learners' difficulties with specific areas of pronunciation (namely phones and prosody) could be detected in the speech waveform used in the system, which could then be used in giving correction to learners. Further, Holland, Kaplan, and Sabol (1999) found that despite the limitations of the speech recogniser in their study, and the misrecognitions it generated, the end users in the study still enjoyed the interactions with the system and would prefer a speech interactive component to be included in the CALL application.

Creating a stress free, low-anxiety environment in which a learner can practice their speaking skills in the L2 with a tireless individual tutor who is able to offer feedback on the learner's spoken output seems an enticing proposition in the field of second language learning. The use of automatic speech recognition technology is the tool by which this proposition can become a reality. It is in the adaptation of the technology that it will find its usefulness in second language learning.

Semantic Interpretation

In a dialogue context, the main goal of speech recognition is to achieve a correct *semantic interpretation* of the user's utterance, so that the system can respond to the utterance appropriately. This can be achieved by including task-relevant semantic information in the grammar used by the speech recogniser, so that the outcome of the recognition process is not just a literal transcription of the utterance but also an interpretation of its meaning. The semantic information is typically expressed within the grammar in the form of slot-value assignments, so that the recognition of a particular word or phrase leads to the filling of a semantic slot with the information contained in that word or phrase. The goal is thus primarily understanding rather than transcription of the utterance. The information which the system obtains from the utterance can then be used to effect in the subsequent dialogue with the user. In order that a learner can interact in a simulated dialogue with the system, the SPELL program described in this paper has been implemented to run in real time using the semantic interpretation approach.

Derwing et al. (2000) state that for speech recognition to be usable for language learners the recognition accuracy must be at an acceptable level and the system must be able to detect errors in the learner's speech in the same way as a native speaker can. One facet in the user testing of SPELL is to determine if the recognition accuracy within the constrained environments available to the learner is of an acceptable level,

and whether the system is able to determine errors in the learner's speech. The focus in SPELL however is not on the learner's pronunciation. Even with accented speech, the conversation can proceed without problem. Only if the accented speech is incomprehensible, when it deviates so far from the pronunciation models that the system cannot recognise it, will the communication need to be negotiated.

In a review of commercial products which use speech recognition, Wachowicz and Scott (1999) found the following to be desirable characteristics in speech interactive CALL:

1. Task-based instruction with an emphasis on communicative authenticity.
2. Implicit, as opposed to corrective feedback in those tasks.
3. Multimodal formats (video, drawings, photos) to enhance authenticity.
4. Focus on schematised, relatively predictable conversations.
5. Verification procedures and repair strategies to counter speech recogniser errors.

The SPELL approach described here adopts these characteristics in order to create *speech interactive* CALL. The following section gives a detailed account of the SPELL approach which utilises the technologies described here in implementing CALL which has at its basis the interaction hypothesis. The approach at the heart of the design is an implementation of the interaction hypothesis within the constraints of the technology.

The Design of Spell

Chapelle (1998) urges designers of multimedia CALL to consider the application as a participant in the L2 interaction: ''It is useful to view multimedia design from the perspective of the input it can provide to learners, the output it allows them to produce, the interactions they are able to engage in, and the L2 tasks it supports'' (p. 26).

The aim in the SPELL approach is to create environments in which both the learner and the computer are co-participants in the interaction. The learner becomes an active participant in the dialogue, where their participation encourages their language development, and they are not simply passive recipients of the language situation. The key feature to the interactive dialogues in SPELL is that the learner's relevant participation in the dialogue is *necessary* for the dialogue to continue.

Garcia-Carbonella, Rising, Montero, and Watts (2001) suggest the positive outcome of using games and simulations in language education since they can:

1. Address student–teacher asymmetry in conventional language classrooms.
2. Increase the amount of exposure to Linput.
3. Lower the affective filter (Krashen, 1985).
4. Enhance interactions through negotiated meaning.

SPELL uses speech recognition so the learner can interact with the system through speech in a simulated, constrained environment. By creating situations in which it is

possible to predict the type of responses, the learner can engage in a dialogue with the computer.

SPELL Scenarios

There are three main components of SPELL: observational, one-to-one and interactive. Each component uses animated virtual agents who reside in a contextualised virtual world. Learners can interact with the animated agents using speech recognition in the one-to-one and interactive components. SPELL is written in Java, and the prototype runs in real time on a standard PC with dedicated resources for speech recognition and advanced graphics.

The observational scenario. Within each lesson, the learner is first offered an opportunity to observe a spoken dialogue between multiple agents within the given scene (see Figure 2). This is termed the observational scenario. Animated agents display three types of action: speech, gesture and facial animation and manipulation of objects in the environment. Animation files hold the gestures and facial expressions that the agents will display. These are also written into the dialogue flow. Some animations will express the agent's internal state (for example frowning during miscommunication), some will have the agent make deictic reference to the virtual world (for example, pointing to a list of options), some will have the agent manipulate a virtual object in the virtual world (for example, holding a glass). The agents 'speak' by means of pre-recorded audio files. It is judged that since the audio files from the

Figure 2. Virtual customer agents in observational scenario

system constitute an important form of input to the learner, the superior quality of pre-recorded audio files outweighs the gain in flexibility of using synthetic (text-to-speech synthesis) speech for the agents. In order for the agents to display speaking animations, scripts are run over each audio file after recording which generates an approximation of the lip movements for each word in each individual audio file.

The language level is often higher than the learner will be expected to produce. However, as it is highly contextualised, the essence of the scene should be apparent to the learner. There are various options within the observational scenario available to the learner: subtitling can be switched on or off, learners have control over the flow of the scenario in that they can pause, stop and restart the dialogue, and learners can access various other files relevant to the lesson such as a transcription of the dialogue, vocabulary, grammar information and cultural information. The observational stage encourages the learner to listen to the contextualised dialogue between the agents, in a low anxiety environment. This also gives the learner the opportunity to become accustomed to the virtual world in which they will become an active participant in the interactive scenarios. In the example, the observational scenario takes place in a virtual café. There are two animated agents in the scene initially, talking about the café. A third agent, the waiter, comes in to take their order. The observational scenario highlights the functional language use of ordering in a café while also showing the agents expressing their likes and preferences.

The one-to-one scenario. After watching the observational scenario, the learner can then engage in a one-to-one dialogue with the tutor agent (see Figure 3). After presentation of the contextualised learning environment, the virtual tutor agent asks the learner some questions relating to the scenario they have just watched, as well as some questions about the learner themselves, thus preparing them for the interactive scenario. Using a headset with the PC, the learner can interact through speech with the agent. The agents use pre-recorded audio files which are played depending on the flow of the dialogue between user and agent. The one-to-one scenarios incorporate various levels of help for the learner, both through spoken audio from the virtual agents and also from text help menus for cases where the learner is experiencing some difficulties. Additionally, the virtual tutor agent offers implicit feedback to the learner when the learner's utterance has been ungrammatical.

Following the observational scenario, the virtual tutor then asks the learner questions about the other virtual agent participants in the scene: what foods they like, etc, then the virtual tutor asks the learner about their own likes and preferences with regards to food and drink. As the scenario in which the interaction takes place is constrained and the topic of the interaction is highly contextualised, predictions can be made on what the learner might say at any given stage of the interaction.

The interactive scenario. The interactive scenario creates an environment in which the learner acts as an active dialogue participant. In this example, the learner ''enters'' the virtual café and sits at the table with the virtual tutor, now acting as a participant in the scenario. The learner's presence is implied through camera

Figure 3. Interaction with the virtual tutor agent in one-to-one scenario

viewpoint movement. The learner interacts with the tutor agent, and when the waiter comes in the scene, the learner receives a menu and then gives their order to the waiter (see Figure 4).

The interactive scenario allows the consequences of what the learner says to be explicitly and immediately demonstrated. For example, in the café environment, following the learner's order of a glass of water from the waiter agent, the agent walks to the counter, chooses the appropriate drink from the available virtual objects, and delivers it to the learner ''seated'' at the table.

A key element in the design of SPELL is that learner input in the dialogue is necessary for the dialogue to continue. In the case of such transactional dialogues, whatever the learner says has consequence for the rest of the dialogue. Errors from the learner will either result in the system giving implicit feedback in the form of a recast, or will prompt the system to reformulate the initial proposition so that the learner can respond again. In cases where the learner has made a semantically appropriate response, confirmation strategies are used so that the learner can confirm that the system has heard correctly. In this way, the learner can engage in an on-going dialogue where their goal is to effectively communicate within any given scenario. Speech recognition makes this a possible activity for learners.

Figure 4. Interacting with virtual agents in interactive scenario

SPELL—Enabling Speech Recognition

As Goodwin-Jones (2000) points out, ''the needs of language learners in respect to speech recognition software are quite different from those of regular consumers'' (p. 7). In order that the learner can receive relevant feedback on their spoken language, the system must be prepared for their unique kind of input. The speech recogniser needs to understand what the learner is trying to achieve, that is, what meaning they are intending. In addition, the recogniser must be able to understand the particular form of the utterance the learner has used.

Within the constrained environment defined for SPELL, it is readily possible to predict to a reasonable degree what a learner might say at any given stage; similarly, it is then possible to predict certain grammatical errors that they might make for any given stage. The aim is to develop recognition grammars specifically for non-native speakers which take into account predicted responses for any given stage in a dialogue, both grammatical and ungrammatical. This is the key element to the development of the language models in SPELL. At each stage in a dialogue, recognition files are accessed which contain a wide range of possible utterances that might be expected from the user at that point. These files are known as *recognition grammars*. In SPELL, at each dialogue stage, the recogniser is required to recognise the possible utterances that the learner might say in order to respond appropriately.

However, to deliver relevant and immediate feedback to the learner's utterance, SPELL also has to recognise those utterances which are non-target like. Recognition grammars specific for the target user group have been developed and are referred to here as ReGaLL: *Recognition Grammar for Language Learners.*

As each new scenario is developed, it is essential to optimise all the possible responses that a learner might make after the prompt wording is finalised. In this way, the wording of the prompt is taken into consideration along with alternative responses in the creation of the ReGaLL. The ReGaLL is written by hand in a shorthand format specific to the recognition engine, and is refined after early testing of each scenario.

In SPELL, to process the incoming speech signal with the relevant language model, the recogniser loads the specific ReGaLL for each individual stage in the dialogue. This constrains the potential number of utterances the recogniser has to evaluate the speech signal against. At any given stage in the dialogue, it is assumed that there are four types of response that the learner can make:

1. Response is semantically appropriate, with no grammatical errors.
2. Response is semantically appropriate, but the response contains grammatical errors.
3. No response has been made.
4. Response is semantically inappropriate for the defined context of the dialogue stage.

Of these types of response, the first two, as far as possible within the constrained environment, are predicted and covered within the ReGaLL (see Figure 5). If the speech signal, i.e. the learner's utterance, matches any of the given utterances coded as acceptable for the learning scenario, then the dialogue will proceed to the next stage. However, if the speech signal is judged to correspond with any of the given ''error'' utterances, feedback is presented to the learner on their utterance.

Within the main file (.Like) two sub files are called. LikeOK contains the expected ''correct'' response to the question. (FoodList is also a sub-file which might contain any number of food items which are relevant to this stage.) Thus a learner might respond to the question with ''Katie likes pizza'', or ''she likes pizza'', both of which might be acceptable answers, and fill the response field *food* with the item ''pizza''. Additionally, the sub-file LikeError is referenced, so that the learner might say, ''she like pizza'' or ''Katie is liking pizza''. In these cases, the *food* field is filled with the response, but a command is triggered in the dialogue which says ''recast''. This is a command for the virtual agent to react in a particular way. A sample dialogue between the virtual agent and learner might be:

- Virtual Agent: What food does Katie like?
- Learner: Umm Katie like pizza.
- Virtual Agent: That's right. Katie likes pizza. What food do you like?

```
.Like
[       LikeOK
        LikeError       ]


LikeOK
[       ([katie she] likes FoodList:f)          ]
        {<food $f>}

LikeError
[       (?it's the FoodList:f)
        ([katie she] like FoodList:f)
        ([katie she] is liking FoodList:f)      ]
{<food $f><command recast>}
```

Figure 5. Sample SPELL recognition grammar for language learners

Thus, the system detects that the learner has responded appropriately for the context, but the response contains a grammatical error, so the system offers some implicit feedback to the learner on their utterance and then moves the dialogue on to the next stage. Implicit feedback is useful in the scenarios as it does not impede the flow of the dialogue between system and learner. Additionally, as speech recognition systems invariably are prone to error themselves, that is, they sometimes misrecognise user responses, the implicit approach allows the system to offer feedback without having to explicitly state that a learner's response was right or wrong. Some hesitations from learners are to be anticipated when they are speaking in the target language. The prototype of SPELL accounts for sentence initial hesitations in the ReGaLL as seen in the above example.

Furthermore, the design of SPELL has taken into account the potential benefits of interactional modifications in its design. The animated agents are able to instigate some forms of interactional modifications in the dialogue where there is a communicative breakdown between virtual agent and learner. In this way, SPELL is able to deduce from the learner's input that for any stage in the dialogue, the learner is having some difficulty. So in the cases of learner response (3) and (4) above, the system assumes that either:

1. the learner has had some difficulty comprehending the initial proposition from the virtual agent; or
2. the learner has difficulty in formulating a response to the proposition.

In either case, SPELL handles this perceived learner difficulty by implementing the *reformulation component* of the dialogue.

The reformulation component in SPELL takes the initial proposition from the system and with increasing levels of help, aims to guide the learner to an understanding of what has been said and how to respond. In the first instance, the virtual agent repeats the initial utterance to the learner, giving more time for a response. At the next level, the virtual agent offers a contextualised example followed by the initial proposition. Further help may then be offered by visual aids. The following example exemplifies how SPELL offers additional input to the learner.

- Virtual Agent: What drink does Katie like?
- Learner: [Silent]
- Virtual Agent: What drink does Katie like? [Slower]
- Learner: Umm—drink . . .
- Virtual Agent: John likes red wine. What drink does Katie like?

Here the virtual agent repeats the initial question, and in further help, the question is contextualized with an example from the observational scenario. Further help still would be in the form of a menu displaying the possible options (see Figure 6).

This general reformulation structure is used throughout SPELL for all cases where the learner has remained silent, asked for clarification (''what?'', ''sorry?'', ''I don't understand'', etc) or said something that is not appropriate for the given stage in the dialogue.

Schmidt and Frota (1986) found that for learners who are attending, being able to hear a corrected version helped them understand what they are doing wrong. Two features are key to making correction effective:

- It must draw students' attention to their own errors.
- It must do so in meaningful, communicative contexts.

One feature of the implicit feedback that has been noted as important is for the learner to be aware that they are being corrected and be aware of the correct form. Mackey, Gass, and McDonough (2000) showed that in cases where feedback is given to the learner from a range of features (morphosyntactic, phonological, lexical, and semantic), the learner does not always perceive the reasons for the feedback, particularly for morphosyntactic errors. In the majority of cases, the learners thought the feedback was either for semantic errors or not corrective feedback at all. It seems then that for morphosyntactic errors, the corrective feedback was not able to meet Schmidt and Frota's first feature to bring attention to the learner's errors, at least not on a conscious level. However, it is impossible to tell whether subconsciously they were noticed or even if the learner was ready for the structure. A future ambition for SPELL is to assess the effects of feedback for morphosyntactic errors, whether they are perceived as corrective feedback and whether the feedback has an effect on uptake of the corrected form. It should be noted that SPELL is designed to give implicit feedback by means of the recast for morphosyntactic errors. One ambition for SPELL

Figure 6. Menu pop-up in one-to-one scenario (Italian version)

is that more detailed L1 – L2 phoneme sets are created, so that it may also be possible to give feedback on phonological errors.

Error Handling in SPELL

In order that relevant feedback can be made, potential errors must be accounted for in the ReGaLL. Whilst it is impossible to predict all possible errors, certain morphosyntactic errors, which are known to cause difficulties for language learners, can be predicted. In order that a profile of any individual learner can be created, SPELL first categorises the types of errors that may occur within the dialogue. The following is an example error classification:

- ErrorType1 = article insertion for uncountable nouns
- ErrorType2 = omission of present tense third singular (−s)
- ErrorType3 = present progressive(−ing) used for present tense

Each error is given an error type which can then be used in keeping a record of the types and frequency of errors made by each learner. Figure 7 shows the error categorisation as applied to the previous ReGaLL example, LikeError.

So, as in the above example, if a learner answers ''she like pizza'', SPELL will record that ErrorType2 has been made. The animated agent then recasts the sentence, and moves on to the next stage in the dialogue. The scenarios depicted in the lessons are created within a constrained language domain such that a reasonable prediction of responses by the learners can be made.

Whilst automatic speech recognition technology cannot replace a human tutor, or human-human interaction for language learning, and will not be able to achieve as accurate a level of listening as can a human interlocutor, one area is which SPELL is superior to the human interlocutor is in the precise logging of particular aspects of the language that any individual learner may have difficulty with. Thus, each time the learner makes an error, SPELL can ''flag'' the error depending on its classification. A frequency count can then be made for each error type, which after extended use then creates a learner profile which can automatically create additional input and exercises targeting this specific area.

Further Research and Development

A prototype of the SPELL system has been implemented for Italian and Japanese and is in user testing with learners of these two languages in five high schools in Scotland. User attitudes to the system are being investigated along with accuracy results from the speech recogniser and accuracy of error handling in SPELL.

Informal user tests of the SPELL prototype have noted that the animated agents could express more affective behaviours in their interaction with the learners. Johnson

```
LikeError
[       (?it's the FoodList:f)
        {<ErrorType1flag SET>}
        ([katie she] like FoodList:f)
        {<ErrorType2flag SET>}
        ([katie she] is liking FoodList:f)
        {<ErrorType3flag SET>}
]
{<food $f><command recast>}
```

Figure 7. Sample error flags in SPELL recognition grammar

et al. (2000) note that animated agents "could exploit the visual channel to advise, encourage and empathise with learners" (p. 71). Elliott, Rickel, and Lester (1999) claim that animated agents, which display emotions, can enhance the student's learning experience: if an agent demonstrates that they care about a student's progress, then the student may also learn to care. They also claim that an agent which exhibits an interesting personality may serve to make the learning experience more fun for the learner. Any application created for pedagogical purposes must strive to make the learning experience as enjoyable and comfortable as possible.

It has been suggested that agents in role-playing situations should be able to display autonomous emotive responses and be able to react to users' emotive input (Marsella, Gratch, and Rickel, 2000). Agents who display autonomous emotive characteristics are possibly more interesting and effective learning companions than hardwired agents. Indeed, Trappl and Petta (1997) state that the agent who displays personality and emotion can serve to promote interest and empathy in the user.

One of the criticisms of the focus-on-form technique of *recasting* has been that in the classroom it is often not used consistently (Lyster, 1998). Therefore, learners may not be aware of the recast, or if they are, they are not always sure what part of their utterance has warranted a recast. Assuming Schmidt's noticing hypothesis (Schmidt, 1990), the learner must at some level notice the form in order that the input can become intake. If inconsistent recasting is used, so that in some instances of an error the form is recast but other instances of the same error the form is not recast, the learner may not notice the form. One of the advantages of SPELL is that consistency of feedback can be met since it can be programmed to give feedback at all occurrences of a particular error.

The unique use of both speech recognition technology and animated virtual agent technology make SPELL a viable resource for the assessment of feedback in one-to-one simulated dialogues. Further research is planned on various areas relating to the feedback the learner receives, such as frequency of recast, type of recast (audio *v.* text), and verbal *v.* non-verbal feedback. Additionally, plans are in place to assess learners' motivation when using SPELL to develop their oral language skills.

Summary

Creating a low-anxiety environment in which a learner can practice their speaking skills in the L2 with a tireless individual tutor who is able to offer feedback on the learner's spoken output offers an enticing proposition in the field of language learning. The use of automatic speech recognition technology in the context of virtual worlds is the tool by which this proposition can become a reality. SPELL develops automatic speech recognition technology specifically for non-native speakers as a tool for effective oral language development. SPELL constrains the available recognition to expected learner output at any given stage, and allows the learner to interact through a number of turns with the animated agents present in the virtual environments.

SPELL offers the learner simulations of every day situations in which they can interact through speech with virtual agents. Topics which readily lend themselves to this format are functional situations such as going to a restaurant (ordering food,

expressing likes and dislikes), at the railway station (buying tickets, expressions of time), and in the town centre (asking for directions). Some personal topics can also be addressed in this format, for example asking (the virtual agent) about their family or talking about sports and hobbies. A limitation of this approach is that the situations do not immediately lend themselves well to more complex types of language. In a truly communicative context, various topics can be introduced from the learners and the language needed to express complex thoughts or situations be raised accordingly. SPELL requires that the situations are predictable to some degree in order that the recogniser can be programmed with possible learner output. This also enables recorded prompts to be used as agent responses to the learner, avoiding speech quality problems associated with text-to-speech synthesis.

Although SPELL cannot replace one-to-one interaction with a native speaker, it can offer a realistic and beneficial simulation in a way that is absent from traditional materials and other CALL approaches. The interaction hypothesis states that conversational interaction between a learner and, for example, a native speaker can facilitate the learner's development as the learner can be involved in negotiated interaction which then gives them relevant and necessary input. In SPELL, learners can engage in negotiated interaction with the animated agents. Furthermore, interactions with the agents can provide learners with immediate and relevant feedback to their spoken output, through reformulations and recasts.

SPELL creates highly contextualised graphical representations of context. The use of speech recognition and the design of the dialogues in SPELL allows the learner to have a simulation of a real world conversation, in which their speech is evaluated in real time, they receive relevant and immediate feedback and their participation within the dialogue is necessary for all dialogue participants to proceed.

Acknowledgements

The authors wish to acknowledge project funding from the Scottish Enterprise Proof of Concept Fund which has made this development possible; and the generous support of Nuance Communications Inc in this work.

References

André, E., Rist, T., & Müller, J. (1999). Employing AI methods to control the behavior of animated interface agents. *Applied Artificial Intelligence*, *13*, 415–448.

Blake, R. (2000). Computer mediated communication: a window on L2 Spanish interlanguage. *Language Learning & Technology*, *4*(1), 120–136. Retrieved from http://llt.msu.edu/vol4num1/blake/. (accessed 24 June 2005).

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., & Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of CHI '99*, 520–527. Pittsburg, PA.

Chapelle, C. A. (1997). CALL in the year 2000: Still in search of research paradigms? *Language Learning & Technology*, *1*(1), 19–43. Retrieved from http://llt.msu.edu/vol1num1/chapelle/. (accessed 24 June 2005).

Chapelle, C. A. (1998). Multimedia CALL: Lessons to be learned from research on instructed SLA. *Language Learning & Technology*, *2*(1), 22–34. Retrieved from http://llt.msu.edu/vol2num1/article1. (accessed 24 June 2005).

Derwing, T. M., Munro, M. J. & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly*, *34*, 592–603.

Ehsani, F., & Knodt, E. (1998) Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning & Technology*, *2*(1), 45–60. Retrieved from http://llt.msu.edu/vol2num1/article3. (accessed 24 June 2005).

Elliot, C., Rickel, J., & Lester, J. (1999). Lifelike pedagogical agents and affective computing: an exploratory synthesis. In M. Wooldridge, & M. Veloso (Eds.), *Artificial intelligence today*, volume 1600 of *Lecture Notes in Computer Science*. Springer–Verlag (pp. 195–212).

Eskenazi, M. (1999). Using a computer in foreign language pronunciation training: What advantages? *CALICO Journal*, *16*(3), 447–469.

Garcia-Carbonella, A., Rising, B., Montero, B., & Watts, F. (2001). Simulation/gaming and the acquisition of communicative competence in another language. *Simulation and Gaming*, *32*(4), 481–491.

Gass, S. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Lawrence Erlbaum associates.

Goodwin-Jones, R. (2000). Emerging technologies: Speech technologies for language learning. *Language Learning & Technology*, *3*(2), 6–9. Retrieved from http://llt.msu.edu/vol3num2/emerging. (accessed 24 June 2005).

Heim, M. (1994). *The metaphysics of virtual reality*. Oxford: Oxford University Press.

Hincks, R. (2002). Speech recognition for language teaching and evaluating: A study of existing commercial products. In *Proceedings of the Seventh international Conference on Spoken Language Processing (ICSLP)*, *Denver Colorado*, 733–736 .

Holland, V. M., Kaplan, J. D. & Sabol, M. A. (1999). Preliminary tests of language learning in a speech-interactive graphics microworld. *CALICO Journal*, *16*(3), 339–359.

Johnson, W. L., & Shaw, E. (1997). *Using agents to overcome difficulties in web-based courseware*. Paper presented at the AI–ED'97 Workshop on Intelligent Educational Systems on the World Wide Web. Kobe: Japan.

Johnson, W. L., Shaw, E., & Ganeshan, R. (1998). Pedagogical agents on the web. In *Working Notes of the ITS '98 Workshop on Pedagogical Agents*, 2–7. San Antonio, Texas.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, *11*, 47–78.

Kitade, K. (2000). L2 learners' discourse and SLA theories in CMC: Collaborative interaction in Internet chat. *Computer Assisted Language Learning*, *13*(2), 143–166.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. New York: Longman.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997a). The persona effect: Affective impact of animated pedagogical agents. In *Proceedings of CHI '97*, 359–366. Atlanta, Georgia.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997b). Animated pedagogical agents and problem-solving effectiveness: A large-scale empirical evaluation. In *Proceedings of the Eighth World Conference on Artificial intelligence in Education*, 23–30. Kobe, Japan: IOS Press.

Long, M. H. (1991). Focus on form: a design feature in language teaching methodology. In K. de Bot, R. B. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39–52). Amsterdam: John Benjamins.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie, & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). New York: Academic Press.

Long, M., Inagaki, S., & Ortega, L. (1998). The role of implicit negative feedback in SLA: Models, and recasts in Japanese and Spanish. *The Modern Language Journal*, *82*, 357 – 371.

Lyster, R. (1998). Recasts, repetition, and ambiguity in L2 classroom discourse. *Studies in Second Language Acquisition*, *20*, 51 – 81.

Mackey, A. (1999). Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, *21*, 557 – 587.

Mackey, A. & Philip, J. (1998). Conversational interaction and second language development: recasts, responses and red herrings? *The Modern Language Journal*, *82*, 338 – 356.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*, 471 – 497.

Marsella, S., Gratch, J., & Rickel, J. (2000). The effect of affect: Modelling the impact of emotional state on the behavior of interactive virtual humans. *Proceedings of the Workshop on Multi – modal Communications and Context in Embodied Agents*, *5th international Conference on Autonomous Agents*, 47 – 52

Marsella, S., Gratch, J., & Rickel, J. (2003). Expressive behaviors for virtual worlds. In H. Prendinger, & M. Ishizuka (Eds.), *Life-like characters tools, affective functions and applications*. Springer Cognitive Technologies Series.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

McBreen, H. (2002). Embodied conversational agents in ecommerce applications. In K. Dautenhahn, A. Bond, D. Canamero, & B. Edmonds (Eds.), *Socially intelligent agents—creating relationships with computers and robots* (pp. 267 – 275). Dordrecht: Kluwer Publications.

Miura, K. (2002) Calico software review. Retrieved from www.calico.org/CALICO_Review/review/tmm – japan00.htm. (accessed 24 June 2005).

Neri, A., Cucchiarini, C., & Strik W. (2003). Automatic speech recognition for second language learning: How and why is actually works. *Proceedings of the 15th international Conference on Phonetic Sciences*, *Barcelona*, 1157 – 1160

Noma, T., & Badler, N. (1997). A virtual human presenter. In *Proceedings of the IJCAI '97 Workshop on Animated Interface Agents: Making them intelligent*, 45 – 51. Nagoya, Japan.

Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning*, *44*, 493 – 527.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, televisions and new media like real people and places*. Cambridge: Cambridge University Press.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 17 – 46.

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: a case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to Learn. Conversation in second language acquisition* (pp. 237 – 326). Cambridge, MA: Newbury House Press.

Slater, M., Usoh, M., & Steed, A. (1994). Depth of presence in virtual environments. *Presence, Teleoperators and Virtual Environments*, *3*, 130 – 144.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass, & C. Madden (Eds.), *Input in second language acquisition* (pp. 235 – 253). Rowley, MA: Newbury House Press.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125 – 144). Oxford: Oxford University Press.

Trappl, R., & Petta, P. (1997). *Creating personalities for synthetic actors*. Berlin: Springer-Verlag.

Wachowicz, K., & Scott, B. (1999). Software that listens: It's not a question of whether, it's a question of how. *CALICO Journal*, *16*(3), 253 – 276.

Wildner, S. (2002). Calico software review. Retrieved from www.calico.org/CALICO_Review/review/germanow00.htm.