Attitudes towards Subtitles and their Effect on User Responses in Speech Interactive Foreign Language Learning

Attitudes to Subtitle Duration and the Effect on User Responses in Speech Interactive Foreign Language Learning

Hazel Morton, Nancie Gunson and Mervyn Jack
University of Edinburgh, Centre for Communication Interface Research
Edinburgh, UK.
Email: {hmorton, nancie, maj}@ccir.ed.ac.uk

***Abstract*- The use of subtitles in a Computer Assisted Language Learning (CALL) program can assist learners in the comprehension of the language input they are exposed to. In an interactive CALL program, subtitles may also help learners in the formulation of their own responses.**
This paper describes the evaluation of a speech interactive CALL program that combines speech recognition technology, embodied animated agents and virtual worlds to create an environment in which learners can converse with virtual characters in the target language in real-time. In particular, the research focuses on the use of subtitles as a help strategy; and the effects of subtitle duration on user attitudes and user responses in the system. Two groups of users participated in the research: high school students learning Italian and high school students learning Japanese. Empirical results are presented from an experiment that evaluated user attitudes towards the speech interactive program and in particular subtitle duration.

***Index Terms*-Human Computer Interaction, Computer Assisted Language Learning, Speech Recognition, Embodied Conversational Agents, Subtitles**

I. INTRODUCTION

Speech interactive language learning applications provide the learner with opportunities to practice their spoken language skills. By using a multi-modal approach, help strategies can be built into the system to assist the learner in the comprehension and production of the language. One such help strategy is subtitling which provides a written transcript of the speech as the speech is played, whether the application depicts actors in a movie, video segments or animated characters.

Research has highlighted the benefits of subtitling in computer assisted language learning (CALL) applications. It has been found that lower ability learners rely on subtitling as part of the listening process whereas higher ability learners use subtitling as a 'back-up' to their listening [1]. It has also been found that language learners who have control over subtitles exhibit better comprehension and better production of the language [2]. Further, it has been found that subtitles are a preferred help option and are used more frequently and for longer durations than a transcript help option in CALL [3]. With respect to depicting subtitles in the target language or in the learner's native language, the target language is facilitative to the learning process whereas using the native language is detrimental to the learning [4].

The point of departure for this research is to investigate the use of subtitles in a speech interactive CALL environment, that is one in which the learner is required to interact with the system through speech rather than solely to listen to the target language. In this case, the subtitles may be used to assist comprehension while the learner listens to the target language. However, the use of subtitles may also assist learners in their own production of the target language.

The CALL program described in this research is predominantly focused on the practice of speaking skills, offering the learner opportunities to engage in the target language by speaking with virtual characters. Some textual information (for example, a dictionary) is available to the learner, accessed by pausing the interaction. The subtitles, however, are available during the interaction (switched on if chosen by the learner) as a means of helping the learner to comprehend the virtual characters' speech. Previous evaluations showed that the speech interactive CALL program described here was found to be enjoyable to use, and achieved high satisfaction scores [5] and boosted motivation [6]. In earlier assessments of the program it was noted that the subtitles were a heavily relied upon help feature. Previous evaluations using this software also showed that when formulating their responses in the program, the majority of users tended to produce one word or short phrase responses, with only a minority of users attempting full sentence responses in the system [5].

The purpose of the study described in this paper is to investigate user perceptions of and attitudes towards the CALL program with particular focus on the help strategy, subtitle functionality. The research investigates whether duration length of the subtitles used for comprehension assistance can also have an effect on assisting user responses.

II. CALL PROGRAM

The program described in this paper, SPELL (Spoken Electronic Language Learning), combines virtual worlds and virtual characters with automated speech recognition technology to create a speech interactive CALL application in which learners can interact in the target

doi:10.4304/jmm.6.5.436-446

language with virtual characters who 'listen' by means of a speech recogniser. The 3D virtual worlds, created in Virtual Reality Modeling Language, (VRML), depict the contextualized environment in which the interaction takes place.

The aim in the SPELL program is for learners to engage in a dialogue with the virtual characters within a defined context. Based on the interaction hypothesis [7], the virtual characters are designed to offer modifications of their input in cases where the user appears to be having difficulties. Interaction provides learners with opportunities to receive comprehensible input and feedback [7, 8, 9]. Further, interaction allows learners to make changes to their own linguistic output [10, 11]. In the SPELL program, the learners are not told in advance what to say, nor are they given a finite list from which to choose their utterances; the speech recognition grammars are programmed with predicted responses for each individual stage of the dialogue, accounting for grammatical and some ungrammatical responses.

It has been suggested [12] that implicit feedback is preferable to corrective feedback for speech interactive CALL systems, as implicit feedback is likely to minimise potential problems resulting from imperfect speech recognition. Feedback in the SPELL program is given implicitly in the form of recasts and reformulations. If the system detects that the learner has made an error in their utterance, the virtual character recasts the learner's utterance. If the learner does not respond, the virtual character repeats the question. If the system detects that the learner has given an answer that is not appropriate to the given stage, the system 'rejects' this and the virtual character reformulates the question, possibly offering a hint to the learner. These feedback strategies allow the dialogue with the learner to continue without explicit reference to a problem. This has the advantage of continuing the flow of the dialogue (and where necessary giving the learner another opportunity to respond, or implicitly correcting their response); and, by being implicit in the feedback, this minimises attention to any potential errors made by the speech recognition component.

The SPELL program offers the learner three scenario types within each 'lesson': observational, one-to-one and interactive. Supplementary materials are also available for the learners to access if they require: vocabulary, grammar files, a transcription of the observational dialogue and cultural information. The observational scenario gives the learner an opportunity to observe the virtual characters within the scene engage in a contextualized dialogue. Fig. 1 depicts the virtual characters in the observational scenario interacting with each other (this image also shows the subtitles switched on).

Figure 1: Virtual Characters in the Observational Scenario with Subtitles Selected

The learner is able to engage in a dialogue with the virtual characters in the one-to-one and interactive scenarios. Each one-to-one scenario raises a topic which is relevant to the lesson. The learner, using headphones and a microphone, is asked a number of questions relating to the given scene by one of the animated characters. For example, in the case of the 'At the station' lesson used in this study, the learner may be asked where in the country (relevant to the particular language lesson) they would like to go or at what time the train to a particular city leaves. These questions introduce topics and sentence structures which are relevant in the given context and which will be useful for participation in the interactive scenario. The one-to-one scenarios give the learner the opportunity for extended sentence practice prior to becoming 'immersed' in 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 railway station and orders tickets at the counter. In the 'At the station' scene, the goal is for the learner to purchase tickets to their preferred destination in the host country. The ticket agent asks the learner where they would like to go and subsequently takes them through a series of questions in order to sell the train ticket.

For a full description of the design of the SPELL program and for details on the speech recognition component, see Morton and Jack [13].

III. EVALUATION OF LESSON

User-centred design and usability studies are common in the field of Human Computer Interaction (HCI). It has been suggested that consideration of usability issues in the design and evaluation of CALL systems is important [14]. Usability engineering emphasises the importance of directly observing potential users [15]. As such a large component of any usability work involves observing user behaviour while interacting with a system. The evaluation

of the CALL program presented here has been designed with usability issues taken into consideration. Data are presented on user attitudes towards using the program and the subtitle types employed. In addition, user response data on the utterances made while interacting with the characters are presented.

The evaluation of the SPELL program took the form of a short, standardized procedure in which target users (language learners) interacted with a fully functional prototype of the system. Researchers were present throughout in order that any problems or issues arising in the use of the system could be observed and recorded. This approach has the advantage that the researcher may notice aspects of the interaction which the learner is unaware of; it also allows the researcher to guide the learner through the session in a pre-designed path so that each learner who takes part experiences, as much as possible, the same procedure.

As part of the full working version of 'At the station' lesson created for Italian and Japanese, two versions of the one-to-one dialogues were created which held different subtitle types. When the subtitle functionality was switched on, the timings/durations of the subtitles were different for the two types. In one subtitle type, the subtitles were visible for the duration of the virtual character's speech; that is, after the character's dialogue turn, during the learner's turn, the subtitle for the character's speech was not visible. As this was the default setting of the subtitle type, this is referred to here as subtitle_default. In the second subtitle type, the subtitle for the character's speech came on as the character's speech was played, and remained on while the turn had passed to the learner. This is referred to here as subtitle_extended. The subtitle settings were stored in the dialogue code.

All participants in the study were asked to try the same scenarios in the same order. However, in this study, investigation was also made of a particular design aspect: the subtitle type. All participants were asked to try the two versions of the one-to-one scenarios. Therefore, a within-subjects comparison could be made of the subtitle type. In order to avoid any order effect in the data, the order of presentation of the two subtitle types was systematically divided amongst the participants so that half of the cohort group experienced the subtitle_default version first, followed by the subtitle_extended version; the other half of the participant cohort experienced the two versions in the other order.

A. Participants

A total of 41 participants took part in the evaluation of the program. 24 students of Italian and 17 students of Japanese were recruited from two secondary schools in Scotland. The evaluations took part in the host schools. Each participant met with the researcher in a dedicated room and the evaluations were conducted on a one-to-one basis, each lasting the duration of one class period (between 45-50 minutes). Therefore, each participant interacted with the software only in front of the researcher and not in view of their classmates.

An equal distribution of male and female participants was not achieved in the evaluation, although it is representative of the gender distribution of students enrolled in these language classes at the schools. The participants were aged between 14 and 15 years. Table 1 details the participants in this evaluation:

TABLE 1
EXPERIMENT PARTICIPANTS

	Male	Female	Total
Italian	5	19	24
Japanese	11	6	17
Total	16	25	41

The participants differed in their prior exposure to the language. The Japanese group were studying Japanese as an extra-curricular activity and were not intending to take an exam in Japanese, whereas the Italian group were studying Italian as a school subject and were aiming to take an exam in Italian. The majority of participants (31) in the assessment had been studying the language for less than one year; 7 participants had been studying between 1 and 2 years and 2 participants had been studying for more than 2 years.

B. Experiment Procedure

Participants were first given a short tutorial on using the program. As stated above, each research session was conducted in private and therefore every participant was given the tutorial by the researcher. The tutorial comprised of using the navigation controls to navigate through the lesson, using the functionality controls to load, start and stop each scenario and access the supplementary materials, such as the dictionary and transcript, both in the lesson overview and during the interactive scenario (by 'pausing' the scenario). The tutorial included the researcher asking the participant to independently navigate the lesson and access various materials in order that the participant would be able to do so if required when using the software in the evaluations.

Following the tutorial, the participants were asked to attempt various aspects of the 'At the railway station' lesson. They were asked to watch the observational scenario, then try two of the one-to-one scenarios (O-O 1 and O-O 2). As part of the experiment design, participants were asked to try each one-to-one scenario twice, each time experiencing one of the two subtitle types. Finally, they were asked to try the interactive scenario. The participants were informed that they could access other features in the program, for example the dictionary, as they wished. All responses made by the participants while interacting with the virtual characters were automatically recorded and logged by the system. The researcher remained present during the program use. After each scenario, the participants were asked to complete an attitude questionnaire. Finally, the researcher engaged the participant in a verbal interview about the opinions of using the program.

C. One-to-One Scenarios

In the one-to-one scenarios, the virtual character asks the learner some questions relevant to the railway station scene, utilising the relevant constructions and vocabulary for the lesson. In this research, participants completed two one-to-one scenarios: *About Train Times* and *Journey Details*. In the About Train Times scenario, the virtual character asks the learner some questions about the departure and arrival times of trains in the host country. To the side of the character on the screen is a timetable depicting the times. In the Journey Details scenario, the virtual character first asks where in the host country the learner would like to go. This dialogue stage is accompanied by a pop-up of a map of Italy or Japan, with 6 cities in each detailed.

In the Italian version, the character asks:

"*Dove desideri andare in Italia?"*

In the Japanese version, the character asks:

"Nihon ni wa doko e ikitai desu ka?"

Following an appropriate response (that is the participant gives a city in the host country, either in a one word form, phrase, or full sentence response), the character then proceeds to ask about the departure and arrival times of the train to that city and the platform from which the train departs (again a timetable is displayed for these dialogue stages to the side of the virtual character). Following the completion of all four stages in the scenario, the character then summarises all the responses. In cases where the response was not appropriate for the given dialogue stage, or the learner has not given a response, the system initiates the reformulation strategy, which in this case would first give the learner another opportunity to respond to the same question and subsequently give a hint to the learner if necessary. If the learner makes a response that is appropriate for the particular dialogue stage, but makes a grammatical error or responds with a one-word reply, the system initiates the recast strategy where the virtual agent recasts the learner's response in a full sentence before moving on to the next dialogue stage.

D. User Attitude Questionnaires

User attitude questionnaires were used for each of the different scenarios that the participants experienced in the lesson. The usability questionnaire was created in order to gather attitude data to each of the scenarios that the participants experienced.

The user attitude questionnaire was based on the salient attributes of the perceived usability of speech interactive systems, identified by previous research [16,17]. The usability attributes were adapted for this speech interactive learning system and covered affective, engagement and interaction issues. The affective issues focused on the learners' levels of anxiety when using the program or interacting with the characters. It has been suggested [18] that anxiety is more of an issue in listening and speaking activities. Therefore, it is particularly important in the investigation of attitudes towards this program to investigate these attributes. Learners' feelings of engagement with the program were collected in order to ascertain the degree to which the learners enjoyed using the program and whether they felt they would be happy to use it again. The interaction issues related to attitudes towards the spoken interaction with the characters and the usefulness of such activities for learning.

A set of 14 statements were created for the evaluation of the interactive scenarios covering these issues. These questionnaires have been used in previous evaluations [1, 5]. As this evaluation also sought to investigate learners' attitudes to the subtitle functionality, an additional four statements were added to the questionnaires for the one-to-one scenarios specifically related to the subtitles. These additional statements were:

I felt the subtitles were distracting.
I felt that the subtitles interfered with my learning of Italian / Japanese.
I felt that the subtitles helped me understand the character.
I thought the subtitles helped me respond to the character.

The questionnaire consisted of a series of short, simple statements, each with a set of tick-boxes on a Likert [19] seven-point scale labelled from "strongly agree" through "neutral" to "strongly disagree", see Fig. 2. The polarity of the statements was balanced to avoid the response acquiescence effect, where respondents may have a natural tendency to agree with proposals.

Figure 2: Example User Attitude Questionnaire Statement

When analysing the results, responses to the questionnaire are first given a numerical value from 1 to 7; these values are then normalised for the polarity of the statements such that a 'strongly agree' response to a positive statement is given a value of 7, whereas a 'strongly agree' response to a negative statement is given a value of 1. After normalisation of the data, the overall attitude for each participant can be calculated as a mean of all of the scores on the items in the questionnaire. These values can then be used to calculate the overall attitude for all items in the questionnaire across all participants in the study. Additionally, mean scores for individual items in the questionnaire can be obtained for all participants.

IV. RESULTS

Due to the nature of the design, in which participants could control the scenarios, for example pausing to look up vocabulary or grammar information, the length of time spent on each scenario varied amongst participants. In consequence, completion rates of the above tasks in the procedure differed amongst learners. As this research focuses on the use of the subtitle duration, the results

section details the results from the one-to-one scenarios only.

37 participants out of 41 completed both one-to-one scenarios twice. The remaining four learners were unable to complete both one-to-one scenarios twice. It is noted that these four learners belong to the Japanese group and they relied heavily on the supplementary materials during the dialogues; therefore it took them longer to complete the scenarios. As some participants did not complete all aspects of the evaluation, the results section will have varying participant numbers. Therefore, for each results sub-section following, the participant numbers involved will be given.

A. User Attitude Results

Following their interactions in the scenarios, participants were asked to complete a short user attitude questionnaire to gather their opinions on such issues as using the system and interacting with the characters.

O-O 1(About train times)

A total of 39 participants experienced both versions of the first one-to-one scenario; attitude data were collected immediately following their experience. The overall mean score of the first usage of O-O 1 was 5.14; the mean score of the second usage was 5.23. Although the second usage was higher, this difference was not significant.

Investigation into differences between the two subtitle types was made. Table 2 details the mean scores for each individual usability statement for the first one-to-one scenario by subtitle type.

A repeated measures ANOVA was computed on participant responses to each of the individual statements in the O-O1 attitude questionnaires for the two subtitle types, with language of study, order of variable presentation and participant gender as between-group variables. Analysis was performed using a statistical software package (SPSS). As multiple comparisons were being made, the risk of obtaining statistical significance erroneously was taken into account by applying a Bonferroni correction. The significance value was set at $p < .05$, with a highly significant result being one where $p < .001$.

Overall, no significant differences were found between the two subtitle types for the first one-to-one scenario for the statements relating *specifically* to the subtitle type. However, it was found that there were significant main effects between the default and extended subtitle types in terms of *a preference for speaking the target language in class* [F=4.207, df=1, p=.049] and *being happy to talk to the characters again* [F=6.247, df=1, p=.018]. In both cases, the extended subtitle type scored higher.

Further, investigation was made on any between-subjects effects of language of study, order of presentation or participant gender across all individual attributes in the questionnaire.

TABLE 2
OVERALL MEANS FOR EACH QUESTIONNAIRE STATEMENT BY SUBTITLE TYPE O-O1
(standard deviation in parenthesis)

Questionnaire Statement	Subtitle_ default (S.D.)	Subtitle_ extended (S.D.)
I felt in control when interacting with the character.	4.45 (1.648)	4.33 (1.623)
I felt embarrassed when interacting with the character.	4.53 (1.739)	4.53 (1.664)
I felt relaxed when interacting with the character.	4.35 (1.460)	4.53 (1.396)
I felt stressed when interacting with the character.	4.68 (1.760)	4.95 (1.679)
I enjoyed interacting with the character.	5.20 (1.471)	5.28 (1.219)
I prefer speaking Italian / Japanese in class, rather than interacting with the character.	3.73 (1.536)	3.90 (1.598)
I would be happy to interact with the character again.	5.40 (1.277)	5.75 (1.032)
I felt that this interaction was useful for my learning of Italian / Japanese.	5.85 (.864)	6.03 (.832)
I felt I always understood what the character said.	5.55 (1.484)	5.63 (1.295)
I felt that the character did not understand what I said.	3.88 (2.267)	3.93 (1.817)
I felt I always knew how to respond to the character.	4.48 (1.585)	4.58 (1.299)
I felt that this dialogue was too easy for me.	5.08 (1.023)	5.15 (1.122)
I felt that the level of the language was too difficult for me to understand.	5.00 (1.569)	4.98 (1.544)
I felt the character was difficult to understand.	5.58 (1.318)	5.30 (1.436)
I felt the subtitles were distracting.	6.38 (.952)	6.30 (.791)
I felt that the subtitles interfered with my learning of Italian / Japanese.	6.15 (.975)	6.13 (1.017)
I felt that the subtitles helped me understand the character.	6.28 (.640)	6.28 (.751)
I thought the subtitles helped me respond to the character.	6.05 (1.011)	6.08 (.807)

Significant main effects of language were found on a number of attributes in the questionnaire, which indicates that irrespective of the version of the scenario tried there was a difference between the two between-subjects groups. Those attributes which showed a significant effect are summarised in Table 3.

TABLE 3
BETWEEN SUBJECTS EFFECTS OF LANGUAGE O-O1

Questionnaire Attribute	Type III Sum of Squares	df	F	Sig.
Prefer speaking <IT/JP> in class	25.860	1	6.030	.020
Understood what the characters said	35.958	1	17.381	.000
Always knew how to respond	11.938	1	4.185	.049
Dialogue was too easy	8.054	1	5.003	.033
Level of language too difficult to understand	54.633	1	23.291	.000
Character difficult to understand	23.034	1	9.268	.005

Between-subjects effects were found for a *preference for speaking in class* (p=.020), *understanding what the character said* (p=.000), *knowing how to respond to the characters* (p=.049), feeling that the *level of the language was too difficult to understand* (p=.000) and feeling that *the character was difficult to understand* (p=.005). From the estimated marginal means, it was found that in each case, the Italian group scored these attributes higher than the Japanese group. Additionally, the between-subjects effects showed one more significant main effect for language on feeling that *the dialogue was too easy* (p=.033). In this case, the estimated marginal mean for the Japanese group was higher than for the Italian group, indicating that regardless of the version used, the Italian group were significantly different from the Japanese group in finding the dialogue '*too easy*'.

With regards to order of presentation of the experimental variables, significant main effects of order were found on three attributes in the questionnaire. Interestingly, each of these attributes related specifically to the subtitles. Between-subjects effects were found for a *subtitles helped me understand the characters* [F=5.037, df=1, p=.032], *subtitles were distracting* [F=4.484, df=1, p=.042] and *subtitles interfered with my learning* [F=7.921, df=1, p=.008]. From the estimated marginal means, it was found that in each case the variables experienced in the order of extended subtitle version followed by default subtitle version scored significantly higher than the reverse order, suggesting that the extended subtitle type had a positive effect on participants attitudes towards the subtitles. However, given that these effects were found only for those questions relating specifically to subtitles, it is possible that the effects noticed here are due to participants noticing the subtitles more in the extended duration condition which they encountered first and therefore endured across their second (default) subtitle condition.

Finally, one significant effect of gender was found for feeling that the *subtitles interfered with learning* [F=5.210, df=1, p=.029]. In this case it was found that females were significantly less likely to feel that the subtitles interfered with their learning, across both conditions, than the male participants.

O-O 2(About journey details)

A total of 37 participants experienced both versions of this second one-to-one scenario; attitude data was collected immediately following their experience.

TABLE 4
OVERALL MEANS FOR EACH QUESTIONNAIRE STATEMENT BY SUBTITLE TYPE O-O2
(standard deviation in parenthesis)

Questionnaire Statement	Subtitle_ default (S.D.)	Subtitle_ extended (S.D.)
I felt in control when interacting with the character.	4.55 (1.519)	4.46 (1.536)
I felt embarrassed when interacting with the character.	4.79 (1.679)	4.90 (1.729)
I felt relaxed when interacting with the character.	4.92 (1.421)	4.79 (1.321)
I felt stressed when interacting with the character.	4.74 (1.735)	4.82 (1.604)
I enjoyed interacting with the character.	5.53 (1.133)	5.28 (1.572)
I prefer speaking Italian / Japanese in class, rather than interacting with the character.	3.84 (1.717)	3.92 (1.783)
I would be happy to interact with the character again.	5.63 (1.195)	5.50 (1.289)
I felt that this interaction was useful for my learning of Italian / Japanese.	5.92 (.850)	5.82 (1.023)
I felt I always understood what the character said.	5.53 (1.409)	5.49 (1.254)
I felt that the character did not understand what I said.	3.71 (1.873)	3.79 (1.773)
I felt I always knew how to respond to the character.	4.47 (1.672)	4.79 (1.472)
I felt that this dialogue was too easy for me.	5.16 (1.151)	4.92 (1.133)
I felt that the level of the language was too difficult for me to understand.	5.03 (1.365)	5.03 (1.405)
I felt the character was difficult to understand.	5.42 (1.244)	5.51 (1.233)
I felt the subtitles were distracting.	5.84 (1.346)	6.10 (1.231)
I felt that the subtitles interfered with my learning of Italian / Japanese.	6.11 (1.060)	5.95 (1.146)
I felt that the subtitles helped me understand the character.	6.11 (.727)	6.21 (.656)
I thought the subtitles helped me respond to the character.	6.16 (1.001)	6.05 (.857)

The overall mean score of the first usage of O-O2 was 5.15; the mean score of the second usage was 5.24.

Although the second usage was higher, this difference was not significant.

Investigation into differences between the two subtitle types was made. Table 4 details the mean scores for each individual usability statement for the second one-to-one scenario by subtitle type.

A repeated measures ANOVA was computed on participant responses to each of the individual statements in the O-O2 attitude questionnaires for the two subtitle types, with language of study, order of variable presentation and participant gender as between-group variables as before. Overall, no significant main effects between the default and extended subtitle conditions were found. With regards to those individual attributes in the questionnaire that focussed specifically on the subtitles, the results did not suggest that participants had a preference for one subtitle type over the other.

Further, investigation was made on any between-subjects effects of language of study, order of presentation or participant gender across all individual attributes in the questionnaire. Significant main effects of language were found on a number of attributes in the questionnaire. Those attributes which showed a significant effect are summarised in Table 5.

TABLE 5
BETWEEN SUBJECTS EFFECTS OF LANGUAGE O-O2

Questionnaire Attribute	Type III Sum of Squares	df	F	Sig.
Prefer speaking <IT/JP> in class	24.588	1	4.631	.040
Stressed when talking to character	19.759	1	4.372	.045
Level of language too difficult to understand	14.258	1	4.511	.042
Character difficult to understand	11.525	1	6.214	.019

Between-subjects effects were found for a preference for *speaking the target language in class* (p=.040), feeling *stressed when talking to the character* (p=.045), feeling that *the level of the language was too difficult to understand* (p=.042) and feeling that *the character was difficult to understand* (p=.019). In looking at the estimated marginal means, it was found that in each case, the Italian group scored these attributes higher than the Japanese group.

With regards to order of presentation of the experimental variables, significant main effects of order were found on three attributes in the questionnaire, again as with O-O1, each of them related specifically to the subtitles. Between-subjects effects were found for a *subtitles helped me understand the characters* [F=13.934, df=1, p=.001], *subtitles were distracting* [F=14.063, df=1, p=.008] and *subtitles interfered with my learning* [F=18.178, df=1, p=.001]. As with O-O1, the estimated marginal means found that in each case the variables experienced in the order of extended subtitle version followed by default subtitle version scored significantly higher than the reverse order.

Finally, one significant effect of gender was found for feeling that the *subtitles interfered with learning* [F=7.500, df=1, p=.025]. Again as with O-O1, females were significantly less likely to feel that the subtitles interfered with their learning, across both conditions, than the male participants.

B. Interview Data

Following the experience with the CALL program, each participant engaged in a verbal interview with the researcher, in order to collect data on their opinions of the program and in particular of the subtitle functionality. The data was later compiled and rated by one rater.

Participants were asked about the subtitle types in the experiment. Most participants (76%) stated that they did not notice a difference between the two subtitle types when they were interacting with the characters. During the interview, the researcher then explained and showed an example of the two subtitle types to the participants. At this point, when asked to choose which subtitle type they would prefer, the majority of participants (71%) gave the extended subtitle type. Some reasons given for this were:

- *It's easier if the subtitles stay up for longer as you can see the structure*
- *I could go over it if I didn't catch it all when he (the character) was talking.*
- *If you didn't know how to respond you could work it out from the subtitles*
- *It gives you a chance to refer back to the question to get the right vocabulary for the answer.*

However, other participants expressed that they felt the extended subtitles do not help to push them in their learning. Some participants comments from this perspective were:

- *It gives you a chance to do it yourself rather than rely on subtitles.*
- *With the subtitles that disappear, it feels like you are having a proper conversation.*
- *It's more of a challenge when they disappear.*
- *It gives you more time to think of your own answer - nothing on screen to distract you.*

These opposing views on this help strategy highlight the individual learning styles of these learners which indicate that some learners like to receive the additional help whereas other learners prefer the challenge of working it out with less help. Participants were also aware that such help strategies could be useful for some learners but not all, and that additional help is useful at different stages of learning. One participant commented "*it should have both, for different levels. First experience the subtitles stay on. Once you get used to it, they disappear.*"

For each subtitle version, participants were also asked whether they thought it helped them respond to the character and whether they thought it was useful for their learning. With regards to the default version of subtitles, 30% of participants stated that the default subtitles helped

them to respond to the characters; however, 84% stated that they felt the default subtitles were useful for their learning. One participant commented: "*Makes it harder - so you have to think more. You get more experience of listening, rather than reading all the time."* Another participant commented on the immersive aspect of the interaction: "*Helped you learn what it's like to be placed in a real situation e.g. in Japan. Instead of just reading*." With regards to the extended version of subtitles, 92% of participants stated that the extended subtitles helped them to respond to the characters, with many comments such as: "*You could see their question and turn it round into your answer*." For the extended subtitle type, 81% stated that they felt they were useful for their learning, a similar figure as given to the default version. These comments suggest that participants felt both subtitle types were useful for learning the language, but that the extended subtitle type was seen as helping learners in making their responses to the characters in the interactions.

Interestingly, some participants commented that it was unnecessary to give the subtitles a longer duration as the functionality of the system allows the user to pause the interaction during a scenario and restart, which offers the user the extra help that the text provides.

Finally, participants also expressed some general opinions of the CALL program. 89% of participants felt that the program helped them when they didn't understand something. Participants gave examples of ways in which the program helped them which included the vocabulary and grammar information available in the program. In addition, participants commented on the reactive help strategies such as pop-up information boards within the virtual scene and the *reformulations* of questions asked by the virtual characters. 97% of participants felt that the program was a useful learning tool. Some participants commented that the program showed them how they could answer the questions; others commented that the program gave them the opportunity to practise in the given situation. One participant commented that the program "*helps you to learn Italian and I feel I want to learn it more here rather than just sitting in class*". Another participant commented that "*it helps you if you make a mistake – it goes over it and tells you the right way to say it."* This refers to the *recast* component in the scenarios which offers implicit feedback to the learner when they have made a grammatical error. This participant's comment is interesting as it highlights that the participant is aware of the implicit feedback and in this case has noticed the 'corrected' form of the utterance.

A total of 92% of participants stated that they enjoyed using the program. This is an encouraging result as enjoyment in using an application can have an effect on the learner's motivation to use an application, and motivation is a key factor in successful language learning. When asked to elaborate on the reasons for their answer, some participants stated that they enjoyed the immersive aspect of the program; some stated that they enjoyed it because it was different, and others stated that they enjoyed speaking with the characters. One participant commented, "*in class, there are other people. This is paying attention to you the whole time- helps you out"* which highlights the individual attention that the program can offer learners. Another participant commented "*you don't get embarrassed if you get something wrong"*. This comment highlights an issue often experienced by language learners, that of anxiety often associated with speaking the target language in front of others.

C. User Response Type

Participants' utterances when interacting with the system were automatically recorded in the program, stored in log files in the system, and later transcribed for analysis of response type as well as recognition accuracy. The system also logged the recognition results at each stage of the dialogue. Participants' utterances were categorised into four response types. As the interaction between characters and user is a series of question and answer pairs, the shortest response type that facilitates the conversation is "answer only". This often is a one word answer (e.g., *Osaka*), but may also be a two word noun compound (e.g., *ju ji*). The second response type used is a "phrase", which constitutes a number of words but does not contain a main verb (e.g., *uno per favore*). For the purposes of the response type analysis, any utterance which contained the copula verb was classified as a phrase (e.g., *Tokyo desu*). The third response type employed is a "sentence" which contains a main verb (e.g., *io desidero andare a Milano*). The fourth response type is "verbal non answer". This final category constitutes responses where the user has made an utterance (which triggers the recogniser), but does not answer the question. For example, mutterings, thinking aloud in English, verbal hesitations and non lexical noises (e.g., coughs) are included in the "verbal non answer" category.

24 learners of Italian took part in the evaluation, all of whom completed the two one-to-one scenarios twice, providing a total of 731 utterances for analysis. 17 learners of Japanese took part in the evaluation, 13 of whom completed two of the one-to-one scenarios twice, providing a total of 466 utterances for analysis.

Tables 8 and 9 detail the response types of the participants in the evaluation.

TABLE 8
USER RESPONSE TYPE – JAPANESE (N=17)

nterac-tion	Utts	Answer only	Phrase	Sentence	Verbal non answer
O-O1	240	12.1%	64.6%	5.8%	17.5%
O-O2	226	34.5%	49.1%	6.2%	10.2%

TABLE 9
USER RESPONSE TYPE – ITALIAN (N=24)

Interac-tion	Utts	Answer only	Phrase	Sentence	Verbal non answer
O-O1	392	17.6%	9.4%	71.4%	1.5%
O-O2	339	23.6%	13.6%	59.3%	3.5%

In the Italian group, the majority of responses overall were full sentence attempts. Only a small minority of responses from the Japanese group were full sentence attempts. The majority of responses were in the phrase category, in this case '<destination> *desu*' or '<time> *desu*'; this was potentially due to the Japanese group having lower ability and confidence in the language, therefore preferring one-word and phrase responses across the scenarios.

The difference of response types between the two language groups may be in part due to either the linguistic competence which the participants hold, or their confidence in using the language. However, responding with an 'answer only' or 'phrase' is still facilitative to the ongoing discourse as the program is designed in order to allow learners to respond as they wish, without forcing them to make full sentence responses.

Investigation was then made on the response type data with respect to the subtitle type. The response type data for the one-to-one scenarios for each of the subtitle types are described in Tables 10 and 11.

TABLE 10
USER RESPONSE TYPE PER SUBTITLE TYPE – JAPANESE (N=17)

Subtitle Type	Utts	Answer only	Phrase	Sentence	Verbal non answer
Default	201	25.4%	55.7%	8.5%	10.4%
Extended	265	21.1%	58.1%	4.2%	16.6%

TABLE 11
USER RESPONSE TYPE PER SUBTITLE TYPE – ITALIAN (N=24)

Subtitle Type	Utts	Answer only	Phrase	Sentence	Verbal non answer
Default	380	17.6%	11.3%	69.2%	1.8%
Extended	351	23.4%	11.4%	62.1%	3.1%

In both language groups, only very slight differences were found for the types of response types across the two subtitle types. For example, the both groups gave a slightly higher percentage of full sentence responses with the default subtitle type and fewer responses classed as 'verbal non answer'. However, overall the subtitle type did not have an effect on the response types made by the users in this study. This indicates that with regards to assisting learners' to formulate fuller responses in an open dialogue design, the longer subtitle duration (subtitle_extended, which remains visible after the oral question from the characters has been made and into the user's conversational turn) does not have an affect. It should be noted, however, that in this study, the ability to 'pause' the interaction when necessary was not disabled. The implications of this are discussed in this next section.

V. DISCUSSION

Attitudes towards the use of the interactive CALL program were very positive, indicating that this kind of dialogue system is potentially very useful for language learners. From the interview data, a total of 92% of participants stated that they enjoyed using the program. Reasons include its immersive nature, its uniqueness in comparison to other language learning tools and the interactivity involved in speaking to the characters. Detailed analysis was made of the speech recognition component which is beyond the scope of this paper. However, as a summary, the analysis of the speech recognition component found an overall word-for-word accuracy of 73.1% for the Japanese group and 57.1% for the Italian group. User attitude results indicate a high level of engagement and enjoyment with using the system, despite misrecognitions made by the system. This is in accordance with research into the use of automatic speech recognition in CALL applications [20], which found that despite the limitations of the speech recogniser and the misrecognitions it generated, end users enjoy the interactions with the system and would prefer a speech interactive component to be included in the CALL application.

User response data found that the Japanese group tended to give shorter responses than the Italian group who attempted full sentences. This may, in part, explain the recognition accuracy being higher for this group. The Japanese group may have opted for shorter responses due to their more limited prior exposure to the target language than the Italian group. However, the Japanese group also scored lower on affective issues (embarrassment, and feeling under stress). The questionnaires highlight that the Italian group, possibly due to their greater exposure to the target language, felt more confident with interacting with the characters and using the program. This may have impacted on the way they responded to the characters, attempting fuller responses in their interaction with the characters.

Indeed, it has been suggested [18] that there may be a correlation between the degree of anxiety felt by the language learner and the complexity of the target language output which the learner produces. They suggest that anxiety "can affect the communication strategies students employ in language class. That is, the more anxious student tends to avoid attempting difficult or personal messages in the target language" (p.126). This correlation between response type and anxiety levels was found in the study reported here. The Japanese group, whose attitude results indicated a higher anxiety factor, tended to use shorter responses with the characters. Some

anxiety issues may have been due to the researcher being present while the learners were interacting with the program and these affective issues may have negatively impacted on the desire to try full sentence responses. In real usage of such a program, the learner would not be observed using the system, which may alleviate some of their feelings of anxiety.

Participants' subjective opinions as expressed in the interview suggested that the extended subtitle type helps the learners to respond to the characters. Over 90% of participants expressed that the extended subtitle helped them to responses and only 30% of participants stated that the default subtitle type helped them to make their responses. However, investigation of the subtitle functionality found that the longer duration of subtitle (subtitle type 'extended') did not have an effect on the users' response types such that users attempted longer utterances with the extended subtitles. In this study, participants who were prone to make full sentence responses did so in both subtitle conditions. Further, users' attitudes towards the subtitles, gathered from the Likert style questionnaire, were not significantly different for the two subtitle types. However, participants indicated in the verbal interview that in general the subtitles are a necessary help aid.

This experiment compared two subtitle conditions, where the duration of the subtitle was the variable. A limitation to the design of this experiment was that a 'no subtitle' condition was not included. This was due to a number of factors. Previous evaluations of the software described here showed that the subtitles were heavily relied upon by users, almost all participants in a previous evaluation opted to switch on the subtitles during their interactions. As the participants were from different schools, the 'lessons' used in the software evaluations were not designed to follow any current school lessons. Thus, the participants were not aware or primed of the lesson goals prior to their taking part in the research. Therefore, it was felt that it would be unfair to the participants to disable the help feature which they most relied upon. In addition, as this was a within-subjects design and each participant was asked to complete a number of different scenarios, fatigue and time issues would prevent introducing another subtitle condition. Therefore, although the experiment described here details the comparison between two subtitle duration conditions, the lack of a 'no subtitle' condition is a considerable limitation to this work.

A further limitation of this study was that the 'pause' functionality in the scenarios was not disabled. It is necessary to pause the scenario in order to access the other features of the program, such as the dictionary or grammar help files. (When interacting in the scenario, in order to strengthen the immersive quality of the interaction, the scenario becomes full screen). Almost all participants in the study accessed other information at some point in the interaction, therefore employing the pause functionality which in effect freezes the given subtitle displayed on the screen. It was acknowledged by some participants in the interview data that users have the control to pause the dialogue if more time is required to formulate a response, or to comprehend the question.

In this study, the extended subtitle type had limited impact on either user responses or on user attitudes towards the interactions. It was found that a longer duration of subtitle, that is one which remains after the system prompt and into the user's conversational turn, when compared with a subtitle which is only visible during the speaker's turn (in this case, the virtual character) did not positively impact upon a lengthier or more complex user response to the system's question. Nor was the extended subtitle perceived differently to the default subtitle duration by users of the system. Therefore, the subtitle_default was restored, and all further implementations of lessons in the interactive CALL program described here used this default subtitle type.

ACKNOWLEDGEMENT

The authors wish to thank Scottish Enterprise and Nuance Communications.

REFERENCES

[1] J. T. Pujolà,. "CALLing for help: Researching language
[2] Learning strategies using help facilities in a web-based multimedia program." *ReCALL*, 14 (2), 235-262, 2002.
[3] I. Borras, and R. Lafayette. "Effects of multimedia courseware subtitling on the speaking performance of college students of French." *The Modern Language Journal,* 78 (1), 61-75, 1994.
[4] M. Grgurović, and V. Hegelheimer. "Help options and multimedia listening: students' use of subtitles and the transcript." *Language Learning & Technology, Vol 11, Num 1*, 45-66, 2007. http://llt.msu.edu/vol11num1/grgurovic/ Accessed 10th September, 2010.
[5] H. Mitterer and J.M. McQueen. "Foreign subtitles help but native-language subtitles harm foreign speech perception." *PloS ONE*, *4* (11): e7785, 2009.
[6] H. Morton, N. Davidson, and M. A. Jack. "Evaluation of a Speech-Interactive CALL system." In F. Zhang & B. Barber (eds.) *Handbook of Research on Computer-Enhanced Language Acquisition and Learning,* Hershey, PA: Idea Group Publishing, pp. 220-240, 2008.
[7] H. Morton, and M. A. Jack." Speech interactive CALL: A cross-cultural evaluation." *Computer Assisted Language Learning, 23* (4): 295-319, 2010.
[8] M. H. Long. "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, 1996.
[9] T. Pica, T. "Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes?" *Language Learning, 44*: 493-527, 1994.
[10] S. Gass. *Input, Interaction, and the Second Language Learner.* Mahwah, NJ: Lawrence Erlbaum Associates, 1997.
[11] M. Swain. "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, 1985.

[12] M. Swain. "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, England: Oxford University Press, 1995.
[13] K. Wachowicz and B. Scott. "Software that listens: It's not a question of whether, it's a question of how." *CALICO Journal 16* (3): 253-276, 1999.
[14] H. Morton and M. A. Jack. "Scenario-based spoken interaction with virtual agents." *Computer Assisted Language Learning, 18* (3): 171-191, 2005.
[15] P. Allum. "Principles applicable to the production of CALL-ware: learning from the field of Human Computer Interaction (HCI)." *ReCALL 13* (2): 146-166. Cambridge University Press, 2001.
[16] J. Karat. "Software Evaluation Methodologies." In Helander, M. (Ed.), *Handbook of Human Computer Interaction*, Amsterdam: North-Holland, pp. 891-903. 1988.
[17] J.C. Foster, R.T. Dutton, M.A. Jack, S. Love, I.A. Nairn, N.A. Vergeynst and F.W.M. Stentiford. Intelligent dialogues in automated telephone services. In C. Baber and J.M. Noyes, (Eds), *Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computer,* London, Taylor and Francis, pp.167-175. 1993.
[18] S. Love, R.T. Dutton, J.C. Foster, M.A. Jack and F.W.M. Stentiford. Identifying salient usability attributes for automated telephone services. *Proceedings of International Conference on Spoken Language Processing*, pp.1307-1310, 1994.
[19] E. K. Horwitz, M. B. Horwitz and J. Cope. "Foreign language classroom anxiety." *The Modern Language Journal*, 70, 125-132, 1986.
[20] R. Likert. "*A Technique for the Measurement of Attitudes"*, New York, Columbia University Press, 1932.
[21] V. M. Holland, J. D. Kaplan and M. A. Sabol. "Preliminary Tests of Language Learning in a Speech-Interactive Graphics Microworld." *CALICO Journal, Vol 16, Num 3, 339-359*, 1999.

Hazel Morton received her Ph.D. from the University of Edinburgh in 2008 on the topic of Speech Interactive Computer Assisted Language Learning. Her research interests lie predominantly in the areas of speech recognition, speech synthesis and embodied conversational agents for eCommerce and eLearning applications.

Nancie Gunson received her Ph.D. from the University of Edinburgh in 2007 on the topic of multimodal spoken language dialogue services. Her principal research interest lies in the design and usability evaluation of spoken language dialogue services for human-computer interaction.

Mervyn Jack is Professor of Electronic Systems at the University of Edinburgh. A Fellow of the Royal Society of Edinburgh, Mervyn leads a multi-disciplinary team of researchers investigating usability engineering of eCommerce services and eBanking services. With 35 years experience of research in his field, his long-term research interests are in dialogue engineering and virtual reality systems design for advanced eCommerce and consumer applications.