ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

Feedback in an ASR-based CALL system for L2 syntax: A feasibility study

Stephen Bodnar, Bart Penning de Vries, Catia Cucchiarini, Helmer Strik, Roeland van Hout

Centre for Language and Speech Technology, Radboud University Nijmegen
{s.bodnar,b.penningdevries,c.cucchiarini,h.strik,r.vanhout}@let.ru.nl

Abstract

In the FASOP project, we are interested in using CALL systems to study issues in foreign language learning, particularly those related to the acquisition of spoken proficiency. Specifically, we use ASR in a CALL system that provides oral practice exercises to investigate the role of corrective feedback in syntax acquisition of Dutch L2. To investigate the feasibility of this proposal we built and tested a CALL system in a pilot experiment. This paper analyses the accuracy of the feedback in the system, showing that it performs accurately. We also provide an overview of learner behaviour in the exercises, with a view to understanding how learners respond to feedback, and to system errors. Based on our analysis, we suggest ways of improving the feedback that might increase training effectiveness without exceeding the limits of the ASR technology.
Index Terms: CALL, ASR, grammar, corrective feedback

1. Introduction

In the FASOP (Feedback and the Acquisition of Syntax in Oral Proficiency) project [1], we study the effectiveness of different forms of corrective feedback (CF) on speaking proficiency through the use of a Computer-assisted Language Learning (CALL) system that employs Automatic Speech Recognition (ASR) [2]. A considerable body of Second Language Acquisition (SLA) research indicates that a clear understanding of the role of corrective feedback in oral proficiency still eludes us [2, 3, 4]. There is evidence that this might be due to the fact that it has not hitherto been possible to create appropriate research conditions to offer feedback that is systematic, consistent, intensive, and clear enough to be perceived as such, and that provides opportunity for self-repair and modified output [5].

The rationale behind the FASOP project is that an ASR-based CALL system makes it possible to approximate such conditions. First, learners can produce spoken output, which is important for language learning [6]. Linguistic knowledge underlying language proficiency is thought to be skill-specific, so that gains from written practice may not transfer to oral proficiency [7]. Second, spoken utterances can subsequently be automatically analysed in order to provide feedback to the learner. Feedback provided by the system can indeed be systematic, consistent, intensive, clear enough to be perceived as such, while opportunities for self-repair and modified output can also be offered. Furthermore, the feedback can in principle be tailored to the learners needs [2, 3]. Finally, an additional advantage of using a CALL-system is that all learner-system interactions can be logged and are thus available for investigating the effect of feedback and the learner reactions from different perspectives.

In this paper we report on a first feasibility study we conducted within the FASOP project to test different aspects of our experimental setup. After describing the content and procedure of the practice session and the instruments developed for measuring feedback effectiveness through the pretests and posttests, we go on to describe part of the results of this study. We first pay attention to the accuracy of the ASR system and then go on to the analyses of the logged learner-system interactions to determine how feedback was reacted to and whether it was effective in terms of uptake (deployment of the correct forms). Subsequently we present the results concerning feedback effectiveness as measured by the pre-tests and post-tests and, finally, the results pertaining to the learners evaluation of the system.

2. FASOP: A First Feasibility Study

In FASOP the central question is whether and how corrective feedback contributes to the acquisition of verb second, a syntactic feature that appears to be problematic even for advanced Dutch L2 learners.

As a first step towards creating a viable language learning environment for CF research, we built and tested a prototype CALL system in a pilot experiment with Dutch L2 learners. In this pilot study we wanted to check the feasibility of the ASR-based syntax exercises, the accuracy of ASR, the feasibility of measuring feedback effectiveness in an ASR-enabled CALL system and the learner impressions of the usefulness, enjoyability and accuracy of the system.

In early spring 2011, we asked a small number of volunteers to participate in an experiment with the prototype system. Five volunteers trained with the CALL system for 90 minutes, and also completed pre- and posttests and a questionnaire that gathered subjective impressions. This number of participants was sufficient for testing that the experimental setup functions correctly before conducting larger studies.

3. Materials and Methods

This section presents materials used in the pilot experiment: a CALL system for verb second inversion training, pre- and posttests measuring changes in proficiency related to the target structure, and a questionnaire gathering participants' impressions of the training. A description of the experimental procedure is also included.

3.1. A CALL system for verb second inversion training

We built an ASR-driven CALL system designed to provide practice for verb second in Dutch L2 spoken production. The activity allowed learners to speak their answers into a microphone, the system analyzed the utterances and provided corrective feedback. A screenshot of the activity is shown in Figure 1. The activity centers around a series of movie clips devel-

oped by an educational publisher for L2 learners of Dutch. After each clip, the system quizzes the learners on events in the story. To answer a question, participants speak an utterance to the system. Not all utterances are permitted. Instead, the learners must make use of a number of answer blocks (see bottom of Figure 1). This is done to constrain the learners' spoken output so that it can be more easily processed by the ASR system. By carefully considering the contents of the blocks, the learner can determine the correct order and record their utterance for system evaluation. Each question is associated with a single target-answer (hidden from the user) which specifies the correct block ordering. To force verb second formulations, some questions contain a fixed answer-block. In total, participants complete 117 questions, 35 of which target verb second formulations while the remaining 82 are included as filler material.

Figure 1: A screenshot of the CALL application. Learners answer questions posed by a virtual tutor about a film clip. In this example, the tutor is asking 'What does it say on the box that Melvin has packed his things in?'. To answer, learners compose an utterance using the prompt and required blocks (marked with 'Allemaal') and one of the optional blocks (marked with 'Eentje') in the bottom half of the screen. A correct order is given by the block sequence 'Op de verhuisdoos van Melvin staat zijn naam' (On the box belonging to Melvin is his name); a typical inversion error is 'Op de verhuisdoos van Melvin zijn naam staat' (On the box belonging to Melvin his name is).

Important in the design of this exercise is the system's ability to provide automatic corrective feedback on learner errors. Three types of feedback are provided, depending on learner output. In cases where the learner's utterance is identical to the target block sequence, the system provides successful-attempt (OK) feedback by displaying a large green checkmark then automatically advancing to the next question.

A second type of feedback is provided when the learner's recording differs significantly from the types of utterances expected in the exercise (e.g. the learner accidentally stops recording partway, or when a required block is absent from the utterance). These types of utterances are detected by comparing the confidence score output by the ASR while processing the utterance with a pre-set threshold, set at a level to be lenient enough to allow reasonable attempts through for error analysis (for instance utterances with a deviant pronunciation, as is often the case with L2 learners). If the ASR confidence score is below the threshold, the system responds with a message equivalent to 'Sorry, I didn't understand. Please try again' (hereafter referred to as did-not-understand (DNU) feedback). It is intended to be a neutral message that simply states that the system did not understand the utterance.

A third type of feedback is used for cases where the system successfully recognizes a valid block sequence, but when the ordering of the blocks in the recognized utterance differs from the target sequence for the question. For these learner outputs, the system responds with an error message stating 'Your answer was incorrect. Please try again'. We refer to this type of feedback as wrong-sequence (WS) feedback. After displaying the feedback message, the system prompts the learner to continue by selecting the next block in the target sequence and placing it to the right of the lead-in and or existing blocks. In this way the system gradually fills in the blanks of the sentence to assist the learner in producing the target block sequence.

To ensure continuation even in case of perceived system errors, learners have access to a skip button that moves them to the next question. In the pilot experiment, participants were instructed to try each question at least three times before using this button.

3.2. Pre- and posttests

Language and proficiency testing is an actively researched area in SLA literature (see [8, 9] for a discussion). To measure changes in accuracy with respect to verb second, we selected two complementary proficiency tests from a collection of tests investigated in [10]. A timed grammaticality judgement test (GJT) measured receptive accuracy, and a discourse completion task (DCT) measured production accuracy. Both tests included time pressure to ensure that participants cannot reflect on the test items, and must reply intuitively.

The GJT employed in our experiment is a sentence rating task. Participants judge a sequence of sentences, indicating for each, whether it is grammatically correct or incorrect. Content for pre- and posttests consisted of identical collections of syntactical structures with different lexical content. Additional sentences containing common language errors made by L2 Dutch learners were included as distractors.

The DCT requires participants to produce spoken language. The participants are given the beginning of a sentence which they have to complete. This constrains the learner into using the target construction in their speech, or otherwise make an error. Additional on-screen information, such as pictures and words, are given to add context.

3.3. Learner questionnaire

We designed a questionnaire that asked participants for feedback on four different aspects of the system: overall impression, videos, practice activity and feedback. Each section contained statements about the system and students indicated whether they agreed or disagreed with each statement using a 5-point Likert scale. The phrasing of the statements was balanced to include an equal number of positive and negative items. Students completed the questionnaire at the end of the second session.

3.4. Procedure

Before the experiment, volunteers were asked to select two days from a two-week period to visit the lab and train with the sys-

	System		
Annotator	OK	WS	total
OK	499	25	524
WS	6	263	269
total	505	288	793

Table 1: Comparing feedback output as expected by an annotator (left-most column) with actual system output (top-most row). Feedback types are successful-attempt feedback (OK) and wrong sequence (WS).

tem. Experiment sessions were designed to take approximately 90 minutes each, with approximately 45 minutes of training, 15 minutes for each pre- and posttest, and 15 minutes for questionnaires and instruction; in practice, they were able to complete the work in the time allotted. In session 1, they completed a background questionnaire, pretests, and practiced with the system. In session 2, they again trained, and next completed posttests and filled out the second questionnaire. Time between the two sessions varied between 2 and 7 days for each learner. During the experiment, the system recorded sound files for each utterance and logged ASR and feedback output.

4. Results

To evaluate the feedback provided by the system, this section presents an objective evaluation of feedback accuracy, a report on question completion rates, data from proficiency tests designed to measure the effects of training with the system, and participants' subjective evaluations of the feedback.

4.1. Feedback accuracy

We evaluate feedback accuracy by comparing system feedback given during training sessions with feedback supplied by an annotator who was a Dutch native speaker during an utterance review task. In the review task, we instructed the annotator to listen to each sound file and determine whether the utterance matched the target block sequence specified for the question. Table 1 compares expected system feedback with actual system feedback recorded in the log files.

The results in Table 1 suggest that the system provides accurate feedback, with the accuracy and F-score of the system being .96 and .97 respectively. Additionally, we can evaluate performance in terms of false rejection and false acceptance rates. In this data, false acceptances make up 6 out of 269 utterances, or 2.2 %. False rejections occur more frequently, accounting for 25 of 524 utterances, or 4.8 %. An examination of the false rejections shows 9 instances which appear to have been caused by non-standard pronunciation that the system should have accepted. The remaining 16 can be attributed to errors unrelated to ASR performance (see section 5 for examples). If we exclude these cases, ASR accuracy rises to .98.

A second type of false rejection (not shown in Table 1), pertaining to OK utterances that were misclassified as DNU utterances, accounted for an additional 18 or 3.3 % of the misclassified OK utterances.

4.2. Feedback and question completion

An important requirement of the training session is that the participants are able to produce the target utterance for each training question. If their first attempt is unsuccessful, the system gives them CF for each new attempt until they complete the question. Learners also had the option of leaving the current question uncompleted and advancing to the next question by using a skip button if they experienced unanticipated difficulties. To assess whether the CF achieves its intended function, we calculate the percentage of successful outcomes of questions in which CF is provided (both DNU and WS feedback) one or more times. Results indicate that in most cases participants who receive CF were eventually able to produce the target sequence, with completion rates varying between 79 and 90 %.

4.3. Training effects

In addition to analysing how learners used CF during training, we are also interested in whether training with the system was successful in increasing learner proficiency in utterances requiring verb second inversion. For information on this, we turn to our pre- and posttests and analyse two outcomes: (1) the number of items correct in the DCT, and (2) the number of items correct in the GJT. For both outcomes we expected improvement in performance between the pre-and posttests outcomes. The outcomes are given in Table 2.

	GJT score		DCT score	
Subj.	pre	pos	pre	pos
1	11	14	13	15
2	14	14	11	14
3	11	9	5	7
4	9	10	1	3
5	11	15	9	12
Avg.	11.2	12.4	7.8	10.2

Table 2: Results of the timed GJT and DCT tests.

Both tests show an improvement in the direction expected, with the exception of the GJT test for subject 3. On the level of the separate tests, the DCT is the only one which produces a significant result in a paired T-test($F(1,4) = 96.000$, $p = .001$). Obviously we cannot draw serious conclusions on the basis of such a small number of subjects and in the absence of a control group, but these results are nevertheless encouraging.

4.4. Learner impressions

In addition to these quantitative measures, we also collected participants' subjective views on the feedback provided during training. Participants were asked how strongly they agree with a number of statements concerning feedback and responded using a 5-point Likert scale. Table 3 shows the ratings for each statement.

Participants all agreed that the feedback was easy to understand, and only one participant found the feedback annoying. Most participants agreed that the provision of corrective feedback is necessary. Concerning their overall impression of accuracy, the accuracy scores (column A in Table 3) for subjects 1 - 4 appear similar, with accuracy at approximately 96 - 98 percent, and yet the subjective impression of the accuracy of the feedback differs among the participants, with results that are positive but mixed (see Table 3, column Q4): three subjects indicated that they agreed or strongly agreed with the statement, one selected a neutral value, and another disagreed with the statement.

Subj.	A	Q1	Q2	Q3	Q4
1	0.970	4	5	5	5
2	0.982	4	3	5	4
3	0.965	4	4	2	4
4	0.961	4	3	-	3
5	0.918	5	2	5	2
Avg.		4.2	3.4	4.25	3.6

Table 3: Accuracy (A) of feedback provided, together with participants impressions of the feedback as recorded in a posttest questionnaire that employed a 5-point Likert scale. Higher numbers indicate more agreement. The statements were: *The feedback is easy to understand* (Q1); *The feedback is not annoying* (Q2); *Receiving corrective feedback is necessary* (Q3); and *The feedback given was correct* (Q4).

5. Discussion and Conclusions

The results of the feasibility study presented in Section 4 indicate that the approach proposed here is indeed feasible, although a number of improvements can be made to the experimental setup. Regarding feedback accuracy, the data suggests that system performance is quite good. In general, in language learning applications false rejections (FR) are considered to be more harmful than false acceptances (FA). In our data we have two kinds of FRs, OK-WS and OK-DNU. The first type is more serious, because telling the learner their speech is wrong when it is not is misleading, and is likely to have negative pedagogical effects. In the present study we have found this value to be suitably low. Considering how many instances occurred in each training session, on average, session 1 contained 1.8 OK-WS instances (distributed over 63 questions), while session 2 contained 3.2 OK-WS instances (distributed over 72 questions).

For DNU feedback, FRs are arguably less serious because the actual message presented to the learner does not state that the learner's speech is incorrect, but only that the system had difficulty understanding the utterance. It is often the case in human-human conversation that well-formed utterances are simply not processed by the receiver and that repetition is necessary. Looking at the results section, we also find that this number is suitably low. On average, session 1 contained 2 OK-DNU instances (distributed over 63 questions), while session 2 contained 1.6 OK-DNU instances (distributed over 72 questions).

For questions where CF was provided, we find that in general learners were able to produce the target utterance. In cases which necessitated use of the skip button, we find that the lack of success is due to a combination of a number of different issues that are, for the most part, unrelated to ASR performance. For example, we observed that occasionally some learners began speaking before recording started or prematurely stopped their recording. A small number of other issues related to the system user interface or authored content (such as questions with blocks containing phonologically similar content) seemed to have affected ASR performance. Overall, the problems mentioned here seem to be of the type expected in a pilot of a prototype system, and can be easily resolved in future experiments.

Comparing feedback accuracy with participants' impressions, we see that overall the impressions of the accuracy are high. Only subject 5 disagrees with the view that the feedback provided by the system is accurate. This may be explained when we look at the interaction with the system: subject 5 had two questions which required 10 attempts, and in one of the questions the learner was unable to produce the answer. Both questions were in the second session, after which the participants filled out the questionnaire. These turns may have caused the subject to respond negatively on the questionnaire.

The scores on the pre- and posttests also show an increase in proficiency by the subjects, suggesting that the CF treatment had a positive effect. This is especially true of the DCT results, which produced a significant result. It is encouraging to observe that improvements in verb second inversion accuracy may be observable after a relatively short training period (90 minutes). An interesting task for the future will be to investigate the learning gains in longer training periods.

Taken together, the results from this pilot study are promising. The ASR-based CALL system developed for the FASOP project appears to be suitable for studying the effect of corrective feedback on the acquisition of syntax in Dutch L2 speakers. Following a small number of improvements, we can be optimistic that the system will serve as a useful platform for researching the contribution of corrective feedback, the relative effectiveness of different types of feedback, and the effect of individualized feedback on L2 syntax acquisition.

6. Acknowledgements

We would like to thank our colleague Joost van Doremalen for developing the ASR-based CALL system used in this experiment and our three anonymous reviewers for their helpful comments. This work is part of the research program 'Feedback and the acquisition of syntax in oral proficiency' (FASOP), which is funded by the Netherlands Organisation for Scientific Research (NWO).

7. References

[1] [Online]. Available: http://lands.let.ru.nl/~strik/research/FASOP.html

[2] B. Penning de Vries, C. Cucchiarini, H. Strik, and R. van Hout, "The role of corrective feedback in second language learning: New research possibilities by combining call and speech technology," in *Proceedings of the L2WS SLaTE conference, Tokio*, 2010.

[3] Y. Sheen, "Introduction," *Studies in Second Language Acquisition*, vol. 32 (2), pp. 169–179 (10), 2010.

[4] R. Lyster and K. Saito, "Oral feedback in classroom sla," *Studies in Second Language Acquisition*, vol. 32 (2), pp. 265–302 (37), 2010.

[5] M. E. Tatawy, "Corrective feedback in second language acquisition," *Working papers in TESOL and Applied Linguistics*, vol. 2 (2), pp. 1 – 19 (19), 2002.

[6] M. Swain, *Input in Second Language Acquisition*. Rowley MA: Newbury House, 1985, ch. Communicative competence: some roles of comprehensible input and comprehensible output in its development, pp. 235 – 253.

[7] K. DeBot, "The psycholinguistics of the output hypothesis," *Language Learning*, vol. 46, pp. 529 – 555 (26), 1996.

[8] R. Ellis, *Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching*. Multilingual Matters, 2009.

[9] J. Norris and L. Ortega, *The Handbook of Second Language Acquisition*. Blackwell Publishing, 2003, ch. Defining and Measuring SLA, pp. 538 – 571(34).

[10] R. Ellis, "Measuring implicit and explicit knowledge of a second language: A psychometric study," *Studies in Second Language Acquisition*, vol. 27 (2), pp. 141–172 (31), 2005.