Spoken Grammar Practice in an ASR-based CALL system

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

Centre for Language and Speech Technology, Radboud University Nijmegen, the Netherlands

{B.Penningdevries, S.Bodnar, C.Cucchiarini, W.Strik, R.vanHout}@let.ru.nl

Abstract
In this paper we present a computer assisted language learning (CALL) system that is developed to practice grammar in spoken language. To enable this, the system uses Automatic Speech Recognition (ASR) to process the L2 learner’s responses. We investigate the possibility of providing corrective feedback (CF) on learner errors, and compare that with self-monitored language learning through output practice. In this paper we present the comparison of the two conditions 1) one group of learners received oral practice and CF on spoken performance, and 2) the other group received oral practice and no CF on spoken performance. We found that our system is successful for L2 speaking practice. The main finding is that both groups show learning after treatment. Between the groups, we did not find a learning difference, but the groups’ sessions proceeded differently. Additionally we found that the CF group was more positive about the system than the NO CF group.

Index Terms: second language acquisition, corrective feedback, speech recognition, CALL

1. Introduction
A requirement for improving second language (L2) spoken proficiency is to practice speaking the language. For L2 learners, the opportunities for spoken practice are often not optimal: in language classrooms practice is limited, and the instruction is not tailored to individual needs.

CALL systems can create more suitable learning situations by offering individualized instruction, increased learner control, reduced anxiety and no time limits. Existing CALL systems, however, focus mainly on written language, whereas the benefits of CALL practice could also be applied to improve spoken practice. This requires use of automatic speech recognition (ASR). Currently, there are no systems that practice grammar in the spoken modality. For this reason we investigated the possibility of developing and testing such a system. In our study we investigate two versions of this system: spoken practice with and without automatic corrective feedback (CF). We present the results of an experiment on grammatical accuracy in oral production, and discuss the results.

2. Research background
2.1. Computer Assisted Language Learning (CALL)
Several studies have shown that CALL can be effective for language learning, and may outperform classroom instruction (e.g. [1]). The reasons for its effectiveness are availability, the amount of individual practice, and the learner’s level of control over the learning experience. Practice with a computer is also found to result in reduced anxiety about making mistakes -which can be a restraining factor in classroom or face to face interaction-and this results in the learners producing more, and more varied, output.

When examining the CALL literature, we find that the majority of studies address written production [2]. However, the differences between written production and spoken production, for instance the increased cognitive load required for speaking and the involvement of the articulatory system, make it interesting to examine oral proficiency.

The rapid advances in the field of ASR technology make it possible to implement focused exercises for spoken production [3]. A review of CALL systems using ASR reveals that these generally address communicative skills or pronunciation, while no systems offer grammar practice in the oral modality [4]. Since developing grammar is important for improving spoken proficiency, there are reasons to develop systems that push learners to practice their grammar in the spoken modality.

2.2. Second Language Acquisition (SLA)
In our study we adopt the interactionist view on SLA, which assumes that interaction and the accompanying feedback serve an important function in developing the interlanguage (e.g. [5], [6]). This is in contrast with proponents of Universal Grammar-based theories (e.g. [7], [8]), who argue that exposure to input is sufficient for language acquisition. These opposing viewpoints have implications for language learning, particularly with respect to the role of output and CF. In designing the learning conditions in our CALL system, we took these issues into account.

2.2.1. Output
The role of output in L2 learning is subject to discussion, with most of the evidence pointing toward a positive effect of output [9] (but see [10] for an opposing view). A theoretical basis for the role of language production for SLA is outlined in the Output Hypothesis [6], which indicates three possible functions of output: 1) a noticing function, 2) a hypothesis testing function and 3) a metalinguistic function. By referring to skill-acquisition theory [11], [12] added the function of enhancing fluency through practice, thus stressing the importance of producing spoken output to improve speaking proficiency.

2.2.2. Corrective Feedback (CF)
Though the role of CF is still controversial, a considerable body of research has shown that CF has an effect on language learning (e.g. [13], [14]). Factors that are found to influence CF effectiveness are educational setting, type of CF [14], and learner differences [15].

In an overview of CF research [16] lists the conditions for CF to be effective for interlanguage development: CF needs to be systematic and consistent, clear enough to be perceived as CF, it
should allow for time and opportunity for self-repair and modified output, the intention of CF should be clear, and the learner should be ready for the feedback. This overlaps with the objections that [17] raises against grammar correction in oral practice, who claims that it is impossible to achieve the ideal characteristics of individualized CF in classroom environments and therefore argues for abandoning grammar correction altogether. He admits at the same time that "the possibility remains that some untested combinations of these variables could produce successful feedback, while avoiding (or minimizing) the accompanying problems." (451)

2.3. Research Questions

Since grammatical accuracy is an important part of proficiency, we designed a system that provides learners with the opportunity to practice and internalize grammar rules and possibly receive CF based on ASR. The demands on our ASR are high, since it has to parse non-native speech accurately and provide accurate CF. Our design takes the limitations of ASR into account, but we need to test whether the system is effective in improving learners' proficiency. By implementing two learning conditions, one based on practicing with spoken output and opportunities for self-correction, and the other based on practicing with spoken output and automatic CF, we can test the working of the system and compare its performance in two learning conditions. Thus we address the following research questions:

- Is it possible to develop an ASR-based CALL system that can detect grammatical errors in spoken performance and provide appropriate corrective feedback?
- Is there a difference of effect between practicing with spoken output and self-monitoring (NO CF group), and practicing with spoken output and automatic CF (CF group)?

3. Method

This section describes our ASR-based CALL system that offers grammar practice in spoken production. The learning exercise we developed is embedded in an experiment to measure learning outcomes by applying pre- and post-tests, and learner appreciation through a post questionnaire.

3.1. The GREET system architecture

In Figure 1, we give a schematic overview of the practice system. The learner interacts with the system through the GUI (a screenshot is given in 3.5, Figure 2), and is presented with a task from the courseware database. The learner is given a question and ‘word blocks’ to construct a sentence. This restricts the number of possible responses and contributes to higher ASR accuracy.

For each question in the exercise, the language model contains all answer possibilities (i.e. all possible sentences) that can be created with the word blocks. These are tagged with meta-information, which in the current experiment is simply whether that answer-sentence is grammatically correct or incorrect. In later experiments, this setup can include more detailed messages.

When the learner records an utterance, the speech recognizer outputs a recognition result and a confidence level. If the confidence level is below a preset threshold, the system assumes that the learner did not record a valid attempt: it does not try to detect errors, but instead the learner is asked to re-record. Otherwise the recognizer maps the utterances to a sentence in the

language model, and error detection sends to the courseware engine the appropriate message regarding the sentence's grammaticality (for more detail see [18]). The final step is the presentation of a feedback message to the learner. Taking the input from error detection, the definition in the courseware engine determines what type of feedback is presented, and how it is presented (see section 3.5.1).

The experiment is run through a website and the learner interacts with the system through a web browser. All recognition is performed on the web server. This allows us to run more experiments at the same time and at different locations, and to store all data centrally.

![Figure 1. A schematic overview of the GREET system](image)

In our experiment the system behaves differently in the two conditions. For both groups, the speech is processed through the recognizer, and the recognition result is logged. However, for one group the answer is explicitly evaluated on the screen, and they can advance or retry according to the system; the other group receives no information on the recognition result, and can advance or retry according to their own preference.

Throughout the experiment, our system logs the learner-system interactions: this allows us to look in detail at learner behavior, and inspect the logs for irregular behavior (cf. [19]). Relevant for the current paper are the interactions with respect to questions, the number of attempts, and when a correct response was given after an error (i.e. repair).

3.2. Procedure

Participants signed up for two sessions of one and a half hours, at times of their choosing. In the first session, they completed a pre-test questionnaire on personal data, two pre-tests (described in section 3.5), and a treatment session (section 3.4). Within a week of the first test day they completed the second session: the second part of the treatment, two post-tests, and finished with a post-test questionnaire (section 3.6).

The experiment was run through a website. Participants logged in and the website gave them their task in a step-by-step fashion. Each time they completed a task, a new one was shown. Before the proficiency tests, instructions were shown in a slide show webpage, and the participants did three practice questions before moving on to the real test.
3.3. Participants

We recruited 29 adult participants from Dutch language courses at A2 and B1 (CEF level) at the Radboud University Nijmegen in the Netherlands. They were offered 15 Euros for participating. The participants were randomly assigned to the CF group, or to the NO CF group. Random assignment and drop-outs resulted in an uneven division in groups: 12 participants in the control group (6 male, 6 female), and seventeen in the experiment group (5 male, 12 female).

In total, there were 13 different L1s in our sample: Arabic, Chinese, Dari, English, French, German, Indonesian, Italian, Russian, Luganda, Polish, Romanian, Spanish (NO CF group: 7 L1s, CF group: 11 L1s). The participants' education was all over eight years of formal education after primary school, with most at university level. Their mean age was 31 years with a range of 22 to 48 years old.

3.4. Target structure

The syntactic feature under investigation was inversion as a result of Dutch verb second (V2). Dutch V2 means that the finite verb appears in the position following the first constituent in the main clause. In Dutch, the subject precedes the verb. Inversion of this order occurs when the first constituent of the sentence is not the subject: the V2 principle requires the verb to remain in second position, thereby forcing the subject to take the position following the verb.

a. Subject-initial main clause
Melvin koopt morgen bloemen
'Subject buys:3SG tomorrow flowers'

b. Inversion clause
Morgen koopt Melvin bloemen
'Tomorrow Melvin buys flowers'

The acquisition of inversion is problematic for L2 learners of Dutch [20]. A reason for this may be that violation of inversion does not necessarily affect meaning. This makes inversion an appropriate feature to study the effect of CF, as CF is likely to benefit errors that do not affect meaning, do not typically lead to communication breakdown, or that lack a clear form-meaning relationship [21].

3.5. Treatment design

Participants practiced implicitly with the target structure. After watching a short (approximately 35s) clip of an ongoing story, a 'teacher' on screen asked them questions about the content. To answer, the participants had to construct a sentence using 'word-blocks': parts of a sentence that need to be combined to form one sentence. The participants recorded their answer by speaking into the microphone.

The treatment was spread over two sessions of 45 minutes each. In total there were 120 questions (63 questions in session 1, 57 questions in session 2), of which 37 were target questions (19 in session 1, 18 in session 2). If the participant completed all the questions within 45 minutes, they went back to start at the beginning, to control for time-on-task.

3.6. Proficiency tests

Two proficiency tests were selected to measure knowledge of the target grammatical feature (accuracy). We selected two tests for cross-task comparison and validation [23]: a timed grammaticality judgment task (GJT) and a discourse completion test (DCT). The tests were selected based on the psychometric study by [24]. The tests are distinct in that the one is a receptive reading task, and the other a spoken production task, but at the same time they are complementary because they measure the same aspect of language competence.

3.6.1. Grammaticality judgment task (GJT)

In the GJT participants judged 40 sentences within a time limit, set at 12 seconds based on pilot versions. The test had an equal number of target and filler sentences, and grammatical and ungrammatical sentences were also equally distributed.

Pre- and post-test versions were counterbalanced, and the order of item presentation was randomized per subject. Correct
judgments scored 1, incorrect judgments scored 0. A response outside the time limit was scored as incorrect.

3.6.2. Discourse completion task (DCT)

The discourse completion task (DCT) elicits oral production. The target structure under investigation is easily avoided in Dutch, so we restricted the possibilities for output, and modified the design to make inversion obligatory in target sentences (see [25] for a similar task design for written production). Participants saw the beginning of a sentence which they were required to complete. To establish some context for the task, they were given a lead-in sentence, one or two hint words, and a picture. To answer, the participant pressed the record button and spoke a full sentence. In this task, there was a time limit of 30 seconds.

The test was counterbalanced, and the order of item presentation was randomized for each participant. Each test version was made up of 32 items of which half were targets, and half were fillers. The recordings were transcribed and scored for correct use of inversion.

3.7. Post-test questionnaire

After the test, the participants were asked their opinion of the system. Subjects indicated on a five-point Likert scale whether they agreed or disagreed with statements about the experiment system. Questions concerned if they felt that their level of Dutch had improved, their self-confidence to speak Dutch had grown, whether it was a good system for learning Dutch, and if they had enjoyed using it. They were also asked open questions to elicit opinions and suggestions about the program and CF.

4. Results

4.1. System performance

In a pilot study we found the feedback accuracy rate of our system to be 96% (accepts) and 97% (rejects) [26]. In this experiment, we examine several indicators such as practice logs and questionnaire responses to see if the CF provided by the ASR was as expected.

Analysis of individual sessions revealed that for two subjects several questions elicited more attempts than normal. They were receiving many 'I don't understand' feedback messages, which does not allow you to advance. When inspecting these cases, we found that for these subjects the sound recording was inferior, which was caused by either the wireless connection, or hardware sound recording issues. Some other subjects also experienced network related sound problems, but for them it did not affect CF accuracy. There seemed to be a particular combination of accent, sentence content, and a female voice with the sound problem that resulted in the ASR being unable to confidently process the response. Overall, we did not see evidence of proficiency level as measured by the pre-test (or CEF level), or L1 background on system behavior.

In total, 23 participants were able to complete the full experiment without assistance from the experimenter, which meant they could practice with the system individually. The only issues that caused disruptions for the other six were some wireless network related problems, learner errors using the recording interface or microphone, and one server crash. These problems are unfortunate, but can largely be avoided in future experiments. However, though unrelated to the ASR technology, they affected the learner's experience with the system.

4.2. Proficiency tests

As a first step in our analysis we inspected the item scores in both proficiency tests. The internal reliability scores (Cronbach's alpha) of the GJT was 0.86, and for the DCT 0.95. The two tests show a high correlation (pre- plus post-test: r=.781, p=.000). However, they are both qualitatively different and measure competence in a different modality, so we take the tests to be complementary in providing us a better picture of the learner’s language level. As a result, we present the test results separately and combined.

Table 1 shows the data from the pre- and post-tests. The CF group received immediate CF after each utterance; the NO CF group had the possibility to re-record their answer after each utterance.

In Table 1 both groups seem to improve. There is no difference immediately visible between the two groups. We turn to an analysis of variance to see if there was a difference of interaction of groups with the treatment. The filler data of the GJT is also included to see if non-target items were learned.

<table>
<thead>
<tr>
<th>Group</th>
<th>DCT Pre M</th>
<th>DCT Pre SD</th>
<th>GJT Pre M</th>
<th>GJT Pre SD</th>
<th>GJT Post M</th>
<th>GJT Post SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF group</td>
<td>.64</td>
<td>.35</td>
<td>.64</td>
<td>.22</td>
<td>.70</td>
<td>.19</td>
</tr>
<tr>
<td>NO CF group</td>
<td>.71</td>
<td>.27</td>
<td>.77</td>
<td>.31</td>
<td>.66</td>
<td>.22</td>
</tr>
</tbody>
</table>

Table 1. Mean and Standard Deviation of proficiency tests

In Table 2 we see that treatment had a significant effect p<.05 in the GJT test, and for the DCT and GJT combined, and a p<.1 for the DCT. We did not find an effect for the fillers of the GJT, which shows that the subjects improved only on the target sentences, and have improved their accuracy of Dutch V2 as a result of treatment.

There is no evidence of an interaction of group with treatment. This indicates that the CF group did not have an additional learning effect as a result of prompt feedback.

4.3. Log data

The system logs interactions with the participant, which is inspected for information on learner behavior during the test. First we excluded ceiling participants, because they have already mastered the rule, and the log data cannot show us learning behavior among these participants. We excluded participants who scored higher than .93 on the combined pre-tests. Then we
The CF group had significantly more attempts. When comparing the NO CF and the CF groups, we found that there is a significant difference for the number of attempts in session 1 ($t$ ($df=27$) = -4.285, $p=.000$), and for session 2 ($t$ ($df=27$) = -3.027, $p=.000$). The CF group also had significantly more counts of successful repair in both sessions: session 1: $t$ ($df=27$) = -4.496, $p=.000$, session 2: $t$ ($df=27$) = -4.666, $p=.000$. (If we take a strict significance level of 0.05, we find that in session 1 for 86% of the attempts, and in session 2, 99% of the attempts, the system is performing as expected. If we take these numbers, there is still a significant difference between the CF and the NO CF group with respect to Attempts and Repair).

Interesting to note is the almost identical number of questions for the CF and the NO CF group with respect to Attempts and Repair. This suggests that CF might be necessary at lower levels of proficiency [cf. 27].

4.4. Post-test questionnaire

In a post-test questionnaire we asked the participants about their practice session in 17 questions. Participants responded using a 5-point Likert scale (1 = very negative, 5 = very positive), from which we calculate their attitude towards the system. The data is given in Table 3 below.

The mean for both groups is above three, which shows that the participants were moderately positive about using the system (3 = neutral). The CF group is more positive about the system: they show a higher appreciation score ($t$ ($df=26$) = 2.193, $p=0.037$).

The participants were also asked open questions. Analysis of these questions revealed that some participants in the CF group indicated a wish to receive more specific CF, or to receive CF on pronunciation. In the NO CF group, most participants indicated they would like to receive feedback, though they did not specify what kind of feedback they would prefer.

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>N</th>
<th>Std. Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO CF</td>
<td>3.34</td>
<td>12</td>
<td>.347</td>
<td>2.50</td>
<td>3.79</td>
</tr>
<tr>
<td>CF</td>
<td>3.68</td>
<td>16</td>
<td>.445</td>
<td>2.94</td>
<td>4.31</td>
</tr>
<tr>
<td>Total</td>
<td>3.53</td>
<td>28</td>
<td>.434</td>
<td>2.50</td>
<td>4.31</td>
</tr>
</tbody>
</table>

Table 3. Descriptive statistics of the post-test questionnaire

5. Discussion

We found that both the CF group and the NO CF group show improved accuracy after practice with GREET. We did not observe an additional effect of providing automatic CF. Though there is a larger number of utterances (attempts) and repairs found in logs for the CF group, we did not find a higher overall improvement. An explanation for this may be that the NO CF group had to pay more attention to their utterance, because they had to self-monitor their answer, making the task was more cognitively demanding. Additionally, we may assume an effect of structured input, since the logs show that both groups received an equivalent amount of input (cf. [25]).

Output practice without CF seems to be effective only when the learner has prior knowledge of the grammatical structure practiced in the treatment. A closer investigation of pre-test proficiency level indicates that NO CF group participants with a low entry level were unable to improve their accuracy on the target structure. This suggests that CF might be necessary at lower levels of proficiency [cf 27].

It is important to note that the NO CF group was told that their answers were recorded and evaluated by the system, and they would receive their score after the experiment. In a pilot experiment where we did not promise a score, the learners indicated that they felt they were not practicing effectively. This leads us to the conclusion that a valuable contribution of ASR in our system is that learners feel that somebody (i.e. the CALL system or an interlocutor) is listening to (the accuracy of) their utterances. The learner seems to need a sense of interaction to produce output that is beneficial for L2 development.

In addition, we found that the learners indicated a preference for receiving CF from the system. This may tie in with the idea that the learner appreciates the feeling of interaction: the immediate response of the system on their utterance. Relevant may also be the role of the positive feedback, in the form of a green check mark.

It is also likely that the (positive and negative) CF had an effect on learner confidence. They may feel more confident that they are learning, because they can only proceed when they provide a correct Dutch sentence. This is also an important aspect of the learning system, because it may prompt learners to use their language more often, and thus practice more.

Though the GREET system worked well overall, we found that there were some improvements possible for the experiment design and the CF design. With respect to experiment setting, a reliable internet connection is required; and we intend to include a microphone test for the learner to check if his/her recording sounds proper before commencing the experiment. With respect
to CF design, it seems that the check for the valid utterance is set too strict for the ASR (confidence level threshold), resulting in too many 'I cannot understand' messages. This may confuse the CF reception by the learner. A reason why we can lower the threshold is because learners are generally found to be motivated and cooperative and trying to record a valid utterance.

6. Conclusions

To improve L2 learning possibilities, we developed an ASR-based CALL system that offers spoken grammar practice. The system worked successfully. L2 learners who practiced their spoken grammar by using the system improved their grammatical accuracy. We obtained learning gains for a group of 29 participants practicing for 90 minutes with our system. In the group that received CF, the ASR component of our system successfully interacted with learners with eleven different L1s. Moreover, we found that the group receiving CF evaluated the system more positively than the group that did not receive CF. The learners of the CF group were positive that the system was a good way to learn Dutch, and that their Dutch had improved as a result of working with the system. This suggests that our current setup is a good way to practice grammar for oral proficiency.

In developing this system, we succeeded in creating learning situations in which we can monitor the learner’s behavior during practice. Additionally, the system interacted with the learner by providing CF on spoken grammar. The system we have developed lends itself well to further language learning research, as we can continue to run experiments under various learning conditions and with an increasing number of participants. Experiments with more learners will eventually shed light on the role and impact of individual learner differences.

7. Acknowledgements

We would like to thank our colleague Joost van Doremalen for developing the ASR component of the CALL system used in this experiment. 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).

8. References