An Experimental Study of Effective Feedback Strategies for Intelligent Tutorial Systems for Foreign Language*

Anita Ferreira

Universidad de Concepcion, Concepcion, Chile
aferreir@udec.cl

Abstract. This paper aims to inform the design of feedback strategies in ITS for Foreign Language. We explore empirical evidence about effectiveness of feedback strategies used in an experimental study in which students interacted with a web-based tutoring program. Results suggest that an ITS for a foreign language should implement feedback which prompts students for answers with grammar errors.

1 Motivation

Most research in ITS has investigated feedback and guidance moves (such as prompting, hinting, scaffolding, and pumping) in teaching procedural skills in domains such as algebra, geometry, physics and computer programming [11,2]. However, little effort has been put into areas such as ITS for foreign languages. The research on tutoring effectiveness has focused on identifying the repertoire of tactics or moves available to tutors [6] such as giving explanations, giving feedback, and scaffolding. In particular, these studies have tried to determine how tutors decide and choose among these different tactics, how they generate explanations and feedback, and what variety of hints they use.

Empirical studies on human tutoring have also been carried out to analyze the tutor responses to error-ridden student contributions. Graesser et al. [3] stated some relevant facts on the complexity of this kind of dialogue move. An effective tutor should give the student feedback in relation to the student's contributions so the tutor can handle the errors by acknowledging that the error occurred, identifying where the error occurred, instructing the student how to repair the error, diagnosing the bugs and misconceptions that generated the error and setting new goals that remediate the error, bugs, and misconceptions,etc. The tutors' feedback moves are responsible for students' learning in a procedural skill. Hume et al. [4] states that the tutors' desire to encourage active learning convinces them to prompt the student with hints. Hinting or reminding is a

* This research is sponsored by the National Council for Scientific and Technological Research (FONDECYT, Chile) under grant number 1040500 "*Effective Corrective-Feedback Strategies in Second Language Teaching with implications for Intelligent Tutorial Systems (ITS) for Foreign Languages (FL).*"

J.S. Sichman et al. (Eds.): IBERAMIA-SBIA 2006, LNAI 4140, pp. 27–36, 2006.

strategy that stimulates the recall of inert knowledge or activates the inferences needed in the completion of a task. In trying to put Hume et al.'s theory of hints into practice in the *CIRCSIM* tutor, Zhou and colleagues [11] found the theory to be too broad and too hard to simulate. There is a significant evidence that a great deal of tutors' tactics can be reframed as prompting or encouraging students to construct knowledge, either through the use of content-free prompts or scaffolding prompts [2]. Scaffolding (or scaffolding episode) has been considered to be a pivotal kind of adult-child interchange in which the adult "guides" the child to develop and achieve to the child's fullest potential.

We explore empirical evidence about the type, frequency and effectiveness of feedback strategies based on studies involving three different learning contexts: an observational study of face-to-face classroom interactions, a case study of one-on-one tutorial interactions, and an experimental study in which students interacted with a web-based tutoring program. The results of our empirical studies were similar with regard to the type, frequency and effectiveness of corrective feedback. In this paper, we focus on an our experimental study of feedback strategies implemented in a web-based computer tutor program. We propose here that the incorporation of effective teaching strategies into ITS for Spanish as a foreign language can be informed by the analysis of effective feedback strategies used by students in a a web-based tutoring program.

1.1 Feedback in ITS for Second Language Learning

ITS for FL have incorporated NLP techniques to analyze learners' language production or model their knowledge of a foreign language, in order to provide learners with flexible feedback and guidance in their learning process. These systems use parsing techniques to analyze the student's response and identify errors or missing items. This allows systems, such as those of [10,5,7], to produce sophisticated types of feedback, such as meta-linguistic feedback and error reports, to correct particular student errors. Sams [10] included information about the students' errors as a type of feedback in the BRIDGE ITS, a multimedia tutoring system for German. When using BRIDGE, students receive feedback about the correctness of their responses for all exercise types.

Levin and Evans [5] also argue that ITS for FL can benefit from NLP technology, enabling systems to produce error feedback based on linguistic analysis. They developed the ALICE-chan system which can identify the location of errors and explain the errors in terms of linguistic relations. However, the feedback provided by ALICE-chan is not pedagogically optimal because it uses technical terms which may be confusing to the student.

Although ITS for FL have been developed, there have been few empirical studies demonstrating the effectiveness of feedback in these systems. Nagata [7] investigated the effectiveness of two types of CALL feedback: traditional feedback that indicates only missing or unexpected words in the learner's response, and feedback that provides further information about the nature of the errors in the form of meta-linguistic rules. The results of an achievement test, followed by a retention test three weeks later, showed that the second type of feedback

was more effective than the first for improving the grammatical proficiency of learners of Japanese as L2 in the use of complex structures.

2 The Experimental Study: Analysis of Effective Feedback Strategies

The results of two previous observational and tutorial studies suggest that for grammar, an ITS for a foreign language should implement ways to prompt students' answers using meta-linguistic cues, elicitation and clarification-requests. There is a tendency for Prompting Answer Strategies (PAS) to be more effective than Given Answer Strategies (GAS) for dealing with grammar errors. Indeed, the prompting strategies seem to promote more constructive student learning in a tutorial context [2] because they encourage the student to respond more constructively than when the teacher gives a simple repetition of the answer or a correction of the error. In second language teaching, **Corrective Feedback** is an indication to a learner that his or her use of the target language is incorrect. In our studies, we classified corrective feedback strategies identified in the SLA literature into two groups:

1. **Giving-Answer Strategies (GAS):** Types of feedback moves in which the teacher directly gives the target form corresponding to the error in a student's answer, or shows the location of the student's error. These include:
   (a) **Repetition** of the error or the portion of the learner's phrase containing the error, using stress or rising intonation to focus the student's attention on the problematic part of the utterance. E.g., S: "Future" (Incorrect tense); T: "¿Future?"
   (b) **Explicit correction:** The teacher provides the correct target form. E.g., S: "Cuando ella andó." *(When she went)*; T: "andaba." This differs from recast because the teacher directly corrects the error without rephrasing or reformulating the student's answer.
2. **Prompting-Answer Strategies (PAS):** Types of feedback moves in which the teacher pushes students to notice a language error in their response and to repair the error for themselves. We have called this group prompting answer strategies because of the similarity these strategies bear to the notion of "prompting" described in [2]. This group includes two types of strategies:
   (a) **Meta-linguistic cues:** The teacher provides information or asks questions regarding the correctness of the student's utterance, without explicitly providing the target form. E.g., S: "Compra" (to buy); T: "Tienes que poner un condicional." *(You have to use a conditional.)*
   (b) **Elicitation:** The teacher encourages the student to give the correct form by pausing to allow the student to complete the teacher's utterance, by asking the student to reformulate the utterance, or by asking questions to elicit the correct answer, such as "How do we say that in Spanish?" E.g., T:Dónde está Jorge (Where is Jorge?) S: "Jorge está ..." (Jorge is...); T: "En la..... (In the....)"

By distinguishing the PAS and GAS groups in our study, we hope to gain further insight into the relative merits of feedback strategies that encourage students to attempt to generate or construct the correct form themselves (PAS), versus those in which the teacher resolves the language error either by indicating the location of the error or providing the target form (GAS). This distinction is motivated by Chi et al.'s [2] study arguing for the benefits of constructive learning, in which the student as an active learner constructs an understanding by interpreting new material in the context of prior knowledge by, for example, making inferences, elaborating the material, by integrating material, and so forth. Knowledge construction may occur as a result of self-explaining (either spontaneously or as the result of elicitation), asking questions, responding to the teacher's questions, etc. As in many previous empirical studies in the ITS literature, Chi et al. [2] studied one-on-one human tutorial interactions, in an attempt to identify the features that make tutoring such an effective learning intervention. Their studies suggest that one-on-one tutoring is effective because it provides students with opportunities to engage in constructive activities, and they argue that ITSs should implement ways to elicit constructive responses from students.

To design the materials and procedures for this experimental study, we developed a teaching component concerning aspects of the subjunctive mood that would help learners to improve their grammatical skills in various ways. This component considered PAS (elicitation and meta-linguistic cues) and GAS (repetition of error and correction) for corrective feedback. For correct answers, positive acknowledgments were considered. The present study addresses the research question: *Are PAS or GAS feedback strategies more effective for teaching the Spanish subjunctive mood for foreign language learners?* by addressing the following hypotheses:

- *Hypothesis 1:* Learners who receive PAS after their subjunctive errors will show greater ability to produce this mood correctly, as measured by pre-test post-test gain scores, than learners not exposed to this feedback.
- *Hypothesis 2:* Learners who receive GAS after their subjunctive errors will show greater ability to produce this mood correctly, as measured by pre-test post-test gain scores, than learners not exposed to this feedback.
- *Hypothesis 3:* Learners who receive PAS after their subjunctive errors will show greater ability to produce this mood correctly, as measured by pre-test post-test gain scores, than learners who receive GAS after their errors with this mood.

2.1 Methodology

Participants. Two groups of students participated in the experimental study. The first group was composed of 6 adult students from a Scottish College. They were all enrolled in the first term of an advanced Spanish course. There was 1 male and 5 females, ranging in ages from 25 to 50 years old (average=34.8).

Participants in the second group were 18 young Jamaican university students. All were enrolled in the second year of a high-intermediate Spanish as a FL course. There were 4 males and 14 females, ranging in age from 18 to 21 years old (average=19.5).

The vast majority of the participants reported English as their first language. The exceptions were two students from the Scottish College, whose L1 languages were French and Portuguese. It is important to note that students were not paid for their participation in the study.

2.2 Design of the Experiment for Spanish Subjunctives

In order to determine the effectiveness of feedback strategies in the context of Spanish as a FL, it was necessary to employ a pre-test post-test control group design. Participants were randomly assigned to form three groups, each one containing eight students:

- **The PAS group:** After the pre-test, the participants received PAS feedback for dealing with the incorrect answers and positive acknowledgements for correct answers during the three treatment sessions, and then the post-test.
- **The GAS group:** After the pre-test, the participants received GAS feedback for dealing with the incorrect answers and positive acknowledgements for correct answers during the three treatment sessions, and then the post-test.
- **Control group:** After the pre-test, the participants received only positive and negative acknowledgements after their answers during the three treatment sessions, and then the post-test.

The students did not receive instruction on the subjunctive mood immediately before the experiment. All students did same sessions and worked with the same material, the only difference was about feedback strategies that they received (PAS, GAS or acknowledgment). The instructional tasks were designed to elicit planned writing production and the recognition of two aspects of the subjunctive mood. The students were invited by email to participate in a series of activities to practise their Spanish and get useful tourist information about Chile.

3 Web-Based Computer Tutor Interface

Each test and activity was constituted by ten exercises, eight about subjunctive and two about indicative mood. These last were included in order to keep the participants on their toes. In order to carry out the experiment for the different activities, a simple web interface was designed to allow students to do the tests asynchronously, that is, it is available to the student anytime, anywhere. The interaction begins with a starting page giving instructions about the experiment as a whole and a personal data entry form. As a student first enters his/her personal details, these are registered and the student is automatically assigned to one of the three experimental groups: GAS, PAS or Control. In order to keep

a balanced sample as much as possible, students are assigned to groups in the following way. The first one receives PAS, the second one receives GAS, the third receives Control feedback, the fourth one receives PAS, and so on.Next, the student starts answering the 10-question pre-test in which the first 5 are fill-in-the-blank questions, and the rest are multiple-choice questions.Since the students finish the activities at different rates, an internal state register is enabled to allow the students to move on at their own rhythm. Once a student starts answering a question, the feedback is provided depending on the type of error that occurred.

4 Analysis and Results

The experiment entailed an independent variable and a dependent variable. The independent variable was the group: (1) the PAS group; (2) the GAS group; (3) the control group. The dependent variable was the difference scores that participants earned between their pre-test and post-test scores. With regard to our first hypothesis *"Learners who receive PAS after their subjunctive errors will show a greater ability to produce this mood correctly, as measured by pre-test post-test gain scores, than learners not exposed to this feedback"*, Tables 1 and 2 show the different scores between the pre-test and post-test for the PAS and control groups. As can be seen, the progress made by the PAS group was much more substantial. This suggests that the participants showed steady improvement in accuracy on the subjunctive mood in the syntactic frames involved.

Table 1. Pre-test and Post-test Results of the PAS Group

	Advanced		High-intermediate							
Test	S1	S2	S3	S4	S5	S6	S7	S8	Total	$\bar{Sc}$
Pre-test	1	2	3	3	5	5	7	4	$\frac{30}{80}(38\%)$	3.8
Post-test	9	8	5	7	6	8	8	7	$\frac{58}{80}(73\%)$	7.3
Difference	8	6	2	4	1	3	1	3	28	3.5

For the purpose of statistically measuring gain in learning, we calculated the differences between the average scores of the post-test and that of the pre-test for each group (Tables 1 and 2), which indicates either an actual gain (if the difference is positive), or a loss (if the difference is negative). The analysis shows that gain scores between the pre-test and post-test for the participants who received PAS feedback were statistically more reliable (average score ($\bar{Sc}$)=3.5; $t - test = 4.04, df = 7, p < 0.005$) than those of control group participants ($\bar{Sc}$=0.7; $t - test = 3, df = 7, p < 0.02$) as predicted by hypothesis 1. In addition, the differences between PAS-gain and control-gain ($t - test = 3.422, df = 7, p < 0.02$) were slightly significant.

Concerning our second hypothesis, *"Learners who receive GAS after their subjunctive errors will show a greater ability to produce this mood correctly,*

Table 2. Pre-test and Post-test Results of the Control Group

	Adv.		High-intermediate							
Test	S17	S18	S19	S20	S21	S22	S23	S24	Total	$\bar{Sc}$
Pre-test	1	5	7	3	3	5	5	4	$\frac{33}{80}(41\%)$	4.1
Post-test	2	5	7	3	4	7	6	5	$\frac{39}{80}(49\%)$	4.9
Difference	1	0	0	0	1	2	1	1	6	0.6

as measured by pre-test-post-test gain scores, than learners not exposed to this feedback", the results in Table 2 and 3 suggest that the progress made by the GAS group was better than the control group. However, gain scores between the pre-test and post-test for the participants who received GAS feedback were significantly less reliable ($\bar{Sc}$=1.9; $t-test = 2.53, df = 7, p < 0.04$) than those of the control group ($\bar{Sc}$=0.7; $t-test = 3, df = 7, p < 0.02$). In addition, the differences between GAS-gain and control-gain scores were weakly significant ($t-test = 2.312, df = 7, p < 0.06$).

Table 3. Pre-test and Post-test Results of the GAS Group

	Adv.		High-intermediate							
Test	S9	S10	S11	S12	S13	S14	S15	S16	Total	$\bar{Sc}$
Pre-test	2	5	7	3	6	5	6	3	$\frac{37}{80}(46\%)$	4.6
Post-test	8	5	7	4	7	9	7	5	$\frac{52}{80}(65\%)$	6.5
Difference	6	0	0	1	1	4	1	2	15	1.9

In accordance with hypothesis 3, *"Learners who receive PAS after their subjunctive errors will show greater ability to produce this mood correctly, as measured by pre-test-post-test gain scores, than learners who receive GAS after their errors with this mood"*, Tables 1 and 3 show the difference scores between pre-test and post-test for the PAS and GAS groups. As can be seen, the progress made by the PAS group was better than the GAS group. This suggests that participants who received PAS after their subjunctive errors showed an improvement in accuracy with this mood.

Gain scores between the pre-test and post-test for the participants who received PAS feedback were statistically more reliable ($\bar{Sc}$=3.5; $t-test = 4.04, df = 7, p < 0.005$) than those who received GAS feedback ($\bar{Sc}$=1.9; $t-test = 2.53, df = 7, p < 0.04$) as predicted by hypothesis 3. In addition, considering the small numbers of data, the differences of PAS-gain with GAS-gain were suggested to be somewhat significant ($t-test = 2.56, df = 7, p < 0.04$).

5 Implications for the Design of ITS for FL

The implementation of strategies in our PAS group requires that the ITS for FL be able to carry on an appropriate interaction with the student. Although

unconstrained conversation of the type that human teachers employ is beyond reach, recent advances in tutorial dialogue systems research make the interactive techniques we propose more feasible than they were at the time many ITS for FL were designed. This research shows that sophisticated interactions can be carried on in domains for which rich underlying models have been developed [8,12], or for which possible correct and incorrect responses have been enumerated and feedback moves for each case have been authored [3,9]. In addition, recent work has shown that an ITS in which students produced self explanations by making selections from a menu led to learning outcomes that were equivalent to a version of the system in which students explained their problem-solving steps in their own words [1]. This result suggests that full-blown natural language understanding may not be required in order to support interactions that evoke knowledge construction.

Based of the results of this study and two previous. We have defined a model for the design of a feedback component for ITS for Spanish as a foreign language (Figure 1), which takes into account:the type of error the learner has made (grammar, vocabulary or pronunciation error), and the learner's level of proficiency (beginner, intermediate, advanced). In our model, we assume that error analysis is performed by an interpreter/analyzer. As noted above, prior ITS for FL have made successful use of parsing technology to identify grammar errors, and recent research on a reading tutor has shown that given good expectations about what the student is trying to say, automatic speech recognition can be used to identify pronunciation errors. In this model, we also assume that the learner's level of proficiency is given.

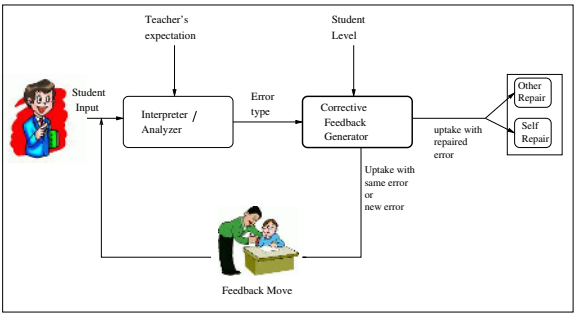

Fig. 1. The Process of Error Treatment and Feedback Generation for an ITS for FL

In our model, the feedback sequence starts when a student's answer contains at least one error. If the answer contains more than one error, the system must determine which error should be treated first, and in our model this decision is based on the learner level. For beginners, grammar and pronunciation errors are the most frequent, and thus we suggest that priority should be given to the treatment of these types of errors. For intermediate and advanced learners, grammar and vocabulary errors should be addressed first. Once an error is identified, a

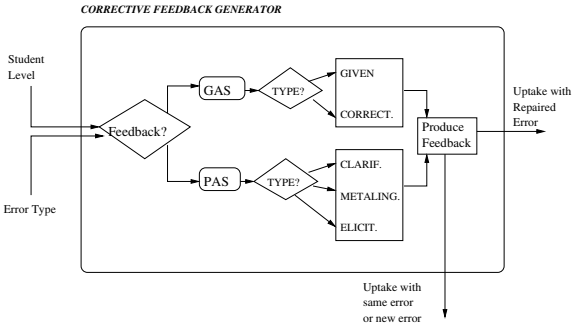

Fig. 2. Model of Corrective Feedback Generation

feedback strategy must be chosen, as shown in the model for feedback generation in Figure 2. After the feedback has been generated, the student may produce several types of responses (uptake):

1. An immediate uptake in which the student modifies his/her answer correctly, either by self-repair (if PAS was generated) or by other-repair (if GAS was generated). This indicates that the student has noticed the error and the given assistance, and the correct answer may indicate a first step towards improvement.
2. An uptake which still contains the error. This may occur because the student did not notice the target form provided by the teacher's feedback or the student does not know how to correct the error. In cases such as this, our human teachers either try an alternative feedback strategy or continue the discussion with the next question, an accept turn, or a domain turn.
3. An uptake in which the student repairs the original error, but his/her answer contains another error. In this case, a feedback strategy is selected according to the algorithm for presenting the first corrective feedback move over an error given above.

A remaining issue that must be addressed in any implementation of our model is how the feedback strategies should be realized in natural language and presented to the student. This will depend on aspects of the overall ITS for FL, such as whether the student interacts with the system using speech or typing (or a combination), whether the ITS for FL is taking a Focus on Forms or Focus on Meaning approach, and so on.

6 Conclusions and Further Issues

This paper addressed the research question of whether feedback strategies (PAS or GAS) are more effective for teaching of the Spanish subjunctive mood for foreign language learners. Overall, the PAS gain was found better than the GAS gain for supporting the process of practicing/learning some aspects of the

subjunctive. After three weeks of the treatment process, learners who received prompting strategies about the sequence of tense and clauses of the subjunctive were significantly more capable of producing the correct forms and of identifying contexts in which the use of subjunctive was appropriate.

Overall, our research approach has been enriched by the research in several disciplines, including Second Language Acquisition, Intelligent Tutoring Systems. These diverse perspectives lead to general questions about how ITS for FL can contribute to alleviating the limitations or disadvantages presented by classroom mode in the treatment of errors, such as giving more opportunity for interaction, prompting student-generated repair. Moreover, the necessity of implementing feedback strategies in ITS for FL can expand our understanding of this key issue and enable us to envisage the kind of contribution that can be useful for ITS for FL systems, as well as teaching training in the context of foreign language instruction.

References

1. V. Aleven, Popescu, and O. Ogan. A formative classroom evaluation of a tutorial dialogue system that supports self-explanation. In V. Aleven, editor, *Procs. of the 11th Int. Conf. on Artificial Intelligence in Education*, pages 345–355, 2003.
2. Michelene T. H. Chi, Stefanie Siler, H. Jeong, T. Yamauchi, and Robert G. Hausmann. Learning from tutoring. *Cognitive Science*, 25(3):471–533, 2001.
3. A.C. Graesser, S. Lu, G.T. Jackson, H. Mitchell, M. Ventura, A. Olney, and M.M. Louwerse. Autotutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, 36:180–193, 2004.
4. G. Hume, J. Michael, A. Rovick, and M. Evens. Hinting as tactic in one-on-one tutoring. *Journal of Learning Sciences*, 5(1):23–47, 1996.
5. L. Levin and D. Evans. ALICE-chan: A case study in ICALL theory and practice. In V. Holland, J. Kaplan, and M. Sams, editors, *Intelligent Language Tutors: Theory shaping technology*. Lawrence Erlbaum Associates, 1995.
6. D. Merrill, B. Reiser, and S. Merrill. Tutoring: Guided Learning by Doing. *Cognition and Instruction*, 13(3):315–372, 1995.
7. N. Nagata. The effectiveness of computer-assisted metalinguistic instruction: A case study in japanese. *Foreign Language Annals*, 30(2):187–199, 1997.
8. E. Owen and K. Schultz. Empirical foundations for intelligent coaching systems. In *Procs. of the Interservice/Industry Training, Simulation and Education Conference*, Orlando, Florida, 2005.
9. C. P. Rosé, P. Jordan, and M. Ringenberg. Interactive conceptual tutoring in atlas-andes. In J. D. Moore, editor, *AI in Education: AI-ED in the wired and wireless future*, pages 256–266. IOS Press, 2001.
10. M. Sams. Advanced Technologies for Language Learning: The Bridge project. In V. Holland, J. Kaplan, and M. Sams, editors, *Intelligent Language Tutors: Theory shaping technology*. Lawrence Erlbaum Associates, 1995.
11. Y. Zhou and R. Freedman. What Should the Tutor do When the Student Cannot Answer a Question? *Procs. of the Twelfth Florida Artificial Intelligence Symposium (FLAIRS'99), Orlando*, 1999.
12. C. Zinn, J. Moore, and M. Core. Intelligent information presentation for tutoring systems. In O. Stock, editor, *Multimodal Intelligent Information Presentation*, volume 27, pages 227–254. Kluwer Academic Publishers, 2005.