Informatics Research Proposal
Intelligent Language Tutoring System for Passive Voice in French

Demetrios G. Eliades
MSc in Artificial Intelligence
Department of Informatics
University of Edinburgh
s0457090@sms.ed.ac.uk

1 Motivation
1.1 Project Purpose
The purpose of the project is to design, implement and evaluate an Intelligent Language Tutoring System for French Passive Voice. The system should be easily accessible through the Web, capable of handling various language tasks and of providing individualized feedback to the learner.

1.2 Selection of Passive Voice
The motive for selecting Passive Voice as the topic for specification in the project is that it is considered as one of the areas which students have difficulty with. Common errors in the formation of passive voice are to be found in verb tense and agreement. In general, areas of Passive Voice in French that need to be explicitly clarified for the native speakers of English are:

1. Past participle agreements
2. Verb agreements
3. Spotting were the direct object is
4. Formation of the past participle
5. Getting the tense right

1.3 Importance of study
Research suggests that Intelligent Language Tutoring Systems (ILTS) are efficient when dealing with second language acquisition. Previous work by [Ferreira-Cabrera, 2003], [Nagata, 1993] and [Nagata, 1995], along with [Heift, 2001] and others has shown that there is positive impact on a student’s learning when ILTS are used. Furthermore, intelligent feedback which is customized to the student’s level of language and takes into account the weaknesses and the preferred style of being taught, can enhance the learning process and help in a better language acquisition.

Pursuing further study in this field, we explore the effects of using an ILTS for teaching an area of French Grammar to learners, and more specifically, the Passive Voice in French, for Anglophone students. Positive results will confirm the assistance of ILTS and as a result, the methodology used can be further applied in other grammatical topics of language learning. It will thus, contribute to the limited literature on Intelligent Language Tutoring Systems, based on a robust framework of experimental methodology.

2 Background
2.1 Feedback in Tutoring Systems
Nagata [1993] cites four types of feedback that tutoring systems use in relation to feedback messages:

1. Present correct answer
2. Pinpoint location of error with pattern matching method
3. List of anticipated errors and their feedback
4. Perform linguistic analysis of the student’s response, comparison with the relevant grammatical rules and lexicon and identify problematic or missing items in answer.

Traditional systems concentrate on the first three types usually utilizing a ‘wrong try again’ approach of interaction, but cannot analyse the students’ response from a linguistic perspective.

2.2 Intelligent Feedback Approaches
Nagata [1993] investigated the effect of metalinguistic feedback in comparison with limited feedback. As limited feedback Nagata perceived as giving the errors found in the answer in their most generic form, instead of going more in-depth and being more specific along with a natural language response, as the metalinguistic does. In general the ‘limited feedback’ is similar to what a traditional tutoring system would provide. She has shown that there is a significant difference between them, favouring the intelligent feedback.

Heift [2000] provided different feedback based on the learners’ level. Beginners were returned the most detailed feedback, whereas the Intermediate level learners were informed that e.g. a certain agreement error occurred. The Advanced level learners merely received a notification that there was an error in their answer. The general idea behind this was that the better the language skill, the less the feedback was needed.

Ferreira [2003] has also explored various intelligent feedback strategies. Her results suggest using a combination of feedback strategy, depending on the error (i.e. if it is pronunciation, grammatical or vocabulary) and on the level of

---

1From personal communication with Dr. Peter Dayan, University of Edinburgh
the learners (i.e. Beginners, Intermediate and Advanced). Ferreira’s findings suggest that a good feedback approach for grammar would be using Prompting Answers Strategies (PAS). There are three different approaches to PAS: Elicitation of the error, Clarifications, and use of Metalinguistic information. For the Beginner level she proposes as first feedback, Elicitation of the student’s answer, and as second feedback, Clarification requests. The first feedback for the Intermediate level should be Clarification, and then Elicitation. For the Advanced Level she suggests Metalinguistic clues as first feedback and Elicitation as the second feedback.

2.3 Evaluation Methodology

Evaluation is essential for reasoning about the use of Intelligent Tutoring Systems. Furthermore it helps to improve the usability of the system, enhance learning outcomes and increase learning efficiency. There are two main types of evaluation [Mark and Greer, 1993]:

- Formative Evaluation is the process of on-going evaluation during the design, in order to recognize errors and problem. This kind of evaluation often appears in an informal manner.
- Summative Evaluation is performed at the end of the design process to test the usability and the effectiveness. It involves statistical methodologies and is confirmed if it succeeds or not in the specifications set in the initial design.

In Tutoring Systems we are seeking for improvements in various fields, either in learning or in some other reduction of time and money. For the purposes of this project, these will be pre-specified as they will determine the experimental design. In general, the observed variables may be among others [Ainsworth, 2003]:

- Learning Gain, by performing Pre-test and Post-Tests
- Learning Efficiency, by observing if it reduces time spent in learning
- Practical Use, by tracking which features are used by the users User’s attitude towards the system
- Cost savings
- How well learners teach what they have learned

The experimental design must include a statement of a causal hypothesis, a variable to be manipulated, randomly assignment of subjects to groups and use of systematic procedures to test the hypothesized causal relationships. Possible problems the design might need to face are for instance the subject performing better than what they would under different circumstances, because of the attention given to them.

There are various methods for performing the evaluation tests of the prototype, each with advantages and disadvantages. For this project the most suitable will be selected, based on the number of the subjects and their time consumption. Some of these methods in related research are [Ainsworth, 2003]:

1. Experiment, Post-test. Quick but does not consider prior knowledge.
2. Pre-test, Experiment, Post-test. Better than previous, as it explains why some improve more than others, takes into concern prior knowledge and can associate similar subjects. It does not check for long term effects.
3. Pre-test, Experiment, Post-test, Delayed Post-test. Checks if improvement maintained, but usually impractical
4. Interrupted time-series. The repeated checks reduce ceiling effects, but effects may appear because of this repetition. It is as the previous, time-consuming.
5. Cross-over of the subjects with the methods tested. Reveals order effects and controls differences between subjects, but is very complex and needs double the number of subjects.

Statistical Methods

In order to explore the meaning of this projects results, statistical analysis of the data will be performed in order to reveal the significance of the findings. In general the following statistical tests are used in the previous paragraph’s methods of evaluation:

- T-test or One way ANOVA: Two group posttest only randomized experimental design
- ANCOVA: Pretest-posttest randomized experimental design

A discussion on the test design is in [Dimitrov and Rumrill, 2003], where they compare two-groups testing with four-groups testing, with randomize subject selection or not. The four-groups measures more information than the two-group evaluation. They propose that ANCOVA statistical analysis should be preferred for pre-test/post-test data, as well as ANOVA for gain scores. They disagree in using ANOVA for residual scores and in repeated measures ANOVA with the pre-test and post-test. Furthermore, they note that use of ANCOVA reduces error variance in randomized test designs, whereas in non-randomized designs ANCOVA is used to adjust the post-test means for pre-test differences among intact groups.

ITS Evaluation

In relevance to the methodology we will follow for this project, we need to specify in between which types of systems are we evaluating the ILTS. From the following forms of intelligent tutoring systems evaluation, we will choose the one that satisfies our project goals and is applicable within the time space [Ainsworth, 2003]:

- No in-between comparison, only use the developed intelligent system
- Intelligent against a system that does not intervene
- Intelligent system against the real classroom
- Two versions of intelligent systems with different tutoring methods
- Intelligent system against an ablated version (with some of the features not in use)
- Against mixed models of difficulty, amount, strategy etc.
2.4 Evaluation Paradigms in ILTS Research

In this section we will analyse two different approaches of ILTS evaluation, one by [Nagata, 1993] and one by Heift [Heift, 2001]. These ILTS were selected to be part of this section, as both of their researches have presented their work in detail and clarity for us to follow. The evaluation was broken into parts such as the number of students, types of exercises and their number, types of system under evaluation, student modelling, feedback strategies, internal system design, data collection and analysis, results and their significance.

Nihongo-CALI, [Nagata, 1993]
- Subjects: 34 second year university students of a Japanese course in University of Pittsburgh
- Exercise Format: A real-life scenario is described to the student. The context is in English and a Japanese sentence is given for the learners to respond to.
- Number of exercises: 90
- Evaluation Systems: Two versions, T-CALI and I-CALI, the first with limited feedback as in similar traditional systems.
- Student Model: No Student Modelling was used.
- Feedback strategy: In the first two tries the student was given information in particle and verbal predicate errors; in the third the answer could be selected from an options menu. All errors are given to the student immediately.
- ILTS System Design: The answer is sent to the NLP parser to get grammatical information. The parser can analyse non-grammatical sentences and produce output messages. These are checked with a pattern matching algorithm to see if the response was appropriate for the question. Final feedback is produced.
- Collection of Data: Pre-test on general grammatical knowledge to pair subjects in first class session. Questionnaire in the end of the computer session. Post-test in class, during the final session.
- Follow-up: One week after experiment, speaking practice on the same topic. Three weeks after experiment, retention test.
- Prior knowledge: Students were given brief explanation on passive structures in first session.
- Provided to Learners: Dictionary, Grammar notes
- Method of statistical analysis: T-Test
- Results: Significant results on post-test ($p < 0.05$) as well as for the retention test (but because of speaking practice in between, the latter cannot be fully justified on I-CALI)
- Issues/Problems: Slow speed of system response

A set of the tests is presented in [Nagata, 1995] during a repeat of the experiment. The pre-test form was with 10 Cloze-type (fill in the blanks) sentences and 2 questions to construct sentences. The questionnaire was with 23 statements graded with Likert scale, i.e. 1-strongly disagree to 5-strongly agree. The post-test was again 10 Cloze-type exercises, plus 9 sentence constructions.

German Tutor [Heift, 2001]
- Subjects: 33 students from two introductory German classes
- Duration: 3 one-hour sessions
- Exercise Format: Create a grammatical German sentence using all provided words.
- Number of exercises: 120
- Evaluation System: One version of the system is evaluated. The system adopts its feedback based on each student’s skill on a task.
- Student Model: It represents the skill level of grammar and vocabulary. The score in each node is increased or decreased according to the student’s performance.
- Feedback strategy: Based on the student model, feedback message is selected based on the system’s assumption about the student’s level. Remedial exercises are provided depending on the state of the Student Model.
- ILTS System Design: The answer is passed to the NLP parser, which is consisted by the grammar information and the parser. Violations are passed to Student model.
- Collection of Data: Computer log were used to collect information on student computer interaction.
- Follow-up: Not known.
- Prior knowledge: The linguistic structures of the experiment had already been practiced before in communicative class activities. Students familiar with grammatical terminology used by the system.
- Provided to Learners: Inflections.
- Method of statistical analysis: Not known
- Results: Performed an error analysis and typology based on students’ level.
- Issues/Problems: Not known issues addressed. No significance tests in evaluation.

Unfortunately Heift didn’t provide significance values for her findings; instead, she performed frequencies analysis of the various error types. As a result an error typology derived, which is also useful but cannot support that the ‘German Tutor’ has actually influenced the students’ learning curve. This raises a general methodological issue of certain projects, that although they produce a respective amount of work (design and implementation), they fail to report significance tests. Instead, if they measure something at all, they may measure parameters of different purpose (as in the above paradigm).

2.5 Hypothesis

The task of the project is to build an Intelligent Language Tutoring System for French Passive Voice. Therefore we wish to demonstrate that the system is indeed capable of providing intelligent feedback and enhance learning performance. The most plausible formation of hypotheses would be to search for learning gain between two learner groups, which are taught with an intelligent and a basic system. The
basic system represents the knowledge a learner would accumulate if s/he were to solve exercises in a textbook. Hence, we perceive the null hypothesis and the hypothesis to be:

\[ H_0: \text{There is no significant difference when comparing post-test results between two groups of learners, when the one is tutored by an ILTS and the other (control group) by an ablated tutor.} \]

\[ H_1: \text{There is significant difference in learning gain between a group that is tutored by an ILTS and a control group that is tutored by an ablated tutor, favouring the first group.} \]

3 Methods
Testing the hypothesis involves two steps: (a) the system design and building and (b) the experimental design and execution. In this section, the stages of the development system will be addressed and the experimental design will be sketched; further details will be discussed in the next paragraphs.

3.1 Software Implementation
Our French Passive Voice ILTS is set on a Client-Server model. The Web system accessible to the student is considered as the Client; the Tutor model, the NLP analyser and the Domain Knowledge, as well as the Student Model are part of the Server. The building of the system will be divided into two parts: the Basic System which will be used for the formative evaluation, mentioned in the next section, and the Full System on which the final evaluation will be performed.

The basis of the implementation will derive from ideas by the ‘German Tutor’, but will adopt feedback strategies from ‘Nihongo-CALI’, both systems mentioned in the previous section. In [Heift and Nicholson, 2000] an explicit technical description of the German Tutor is provided. In short, the students run their exercises on Java applets, the responses are sent to a web-server, where the Java/HTML pre-processing occurs. The sentences are analysed on a dedicated Prolog-server. As described in the same paper, German Tutor passes the students’ input through various checks: Spelling, Example (pre-stored responses), Missing Word, Extra Word, Grammar and Punctuation. On each error, the check sequence is interrupted and related feedback is given.

System Architecture
The development of the ILTS is built on the ‘HTML-CGI’ architecture. The learner interacts from the HTML pages and communicates with the learner through a standard Web browser. The solutions and the actions of the user are sent to the Web server, which forwards the messages to a CGI script for processing and action. The control is passed to the Tutor Model and when the feedback is returned, the CGI script creates the new HTML pages and sends them back to the student’s browser. In that way, it will be easy to update the system, since it is only installed on the server, but perhaps a possible problem is that if the server is very busy, there processes will be delayed.

Tutor Model
This part is responsible for checking the students’ answer and deciding whether to respond with a proper feedback message, or to choose to continue with the next exercise. This model relies heavily on the expert teachers’ experience, for what is important in tutoring and what is not. Within the Tutor model, the answers of the student are checked for spelling errors, missing or extra words and for their grammaticality, as discussed in the next subsection. According to the mode of operation (fully functional or ablated) the feedback is generated. Consideration will be given on communication with the Student Model in a potential update, as discussed in the last subsection.

Knowledge Representation
As a knowledge base of the grammatical and lexical information, Prolog programs will be used. Spelling checking will be either performed with an external spelling program, or within the Prolog code, through a lexicon. Missing word checking will make sure if all words are in the student’s answer. Extra word checking will confirm that the student has not included a new word in the sentence, thus making the grammatical analysis task unhandy.

Lastly, the grammatical parser in Prolog will analyse the sentence, consult its database and return information about the kind of metalinguistic feedback which will be used. The parser will be able to handle ungrammatical sentences and return specific error values, that correspond to each of the possible errors. The metalinguistic feedback, depending on the error, will formulate an appropriate response to the learner, within a simple natural language generator or with templates.

Student Model
With each interaction with the system, the student model will be updated. For this project the Student Model will follow the methodology proposed by [Beck et al., 1997]. The student model for the passive voice tasks for each student will be represented by difficulty-level lists. In this way, the lowest level of knowledge will be represented as [1,0,0,0,0] and the highest as [0,0,0,0,1] for each passive voice task. Therefore, a vector \([0.2, 0.2, 0.4, 0.2, 0.0]\) for a task will imply that the student is by 20% chance at level 1, 2 or 4 and by 40% chance at level 3. The model is updated when an error occurs and feedback is returned, or when an exercise of a specific level is solved in the correct way. Although the Student Model will not be used during the evaluation, the data gathered at that point will form a basis for further research.

3.2 Experimental Design
The most suitable method for evaluating the hypothesis is the ‘Intelligent Tutor vs Ablated-Intelligent Tutor’. For the test selection, the most suitable one appears to be a pre-test/post-test analysis. All the other methods have been rejected as they are either impractical or time consuming, or need a large number of subjects.

3.3 Exercises
Exercises are expected to be discussed with teachers of French. The exercises will include French - French transformation to passive voice of varying difficulty and topic of consideration. The learners are expected to have already been taught the Passive Voice. Previous knowledge acquired by the learner in previous exercises is expected to be used.
Effort will be made to ensure the simplicity of the formation of the exercises, in order for them to be accessible to every teacher, providing the basics for developing an authoring system in a future research.

4 Evaluation

4.1 Formative Evaluation Methodology

As mentioned in a previous paragraph, the formative evaluation is performed during the design of the system. Its purpose is to reveal errors in the system, buggy behaviours and other problems. During this stage, feedback from students and teachers will be requested. For this we need to develop beta-versions of the ILTS, even if not all of the features are fully functional. This procedure is important, as experts and students provide necessary information for improving the evaluation version of the tutoring system.

4.2 Prime Summative Evaluation Methodology

Two separate groups of learners are needed to interact with the system for a specific period of time and for a limited number of exercises. A Pre-test will provide the initial information of each learner’s skills on grammar, and thus be paired with another student of similar competence. Allocation to each group will be based on the results of the pre-test. A one-hour evaluation of the system will be conducted, with a series of exercises. The ILTS group will receive Intelligent feedback, while the Control group will only be notified as to whether the exercise was correct or not. After the evaluation, a Post-test of the same type as the pre-test was, will be performed to measure the improvement of the subjects. Both Pre-test and Post-test will include Cloze-type (fill in the blanks) and a number of sentence writing exercises. Although not in the scope of the project, a questionnaire will be provided at the end to comment on the interaction with the system.

At all times, all the interaction of the students with the system will be anonymously recorded for a posterior analysis of the data.

4.3 Backup Evaluation Methodology

In case of human-caused problems during the evaluation (i.e. not obtaining a student group or insufficient number of Anglophone students) we propose to follow a different evaluation methodology. The group subjects will be volunteers, members of the students’ e-mailing lists, and must be persons who have once studied French and would like to recap their knowledge in the passive voice. The pre-test will be conducted through the web. A selection of subjects will be performed and those interested in continuing the experiment will be allocated in one of the two groups. The experiment will be performed in asynchronous manner, i.e. each student in their own time. After finishing the experiment within a specific time, they will complete the post-test. A questionnaire will be sent via email.

Unfortunately web experiments have the risk of deception by participants, or the subjects might not perform as they would under normal tutoring conditions. If this methodology is to be followed, those reservations will be adequately pointed out.

5 Expected Outputs

As an outcome of the project, a prototype system will be implemented and tested upon a group of students. At the end of the evaluation, we will have gathered a number of pre-test and post-test scores. We expect that the pre-tests and post-tests will affirm that the two groups have acquired knowledge in a different manner, favouring those participating in the ILTS group. The prototype will be easily integrated into a more grammatically diverse ILTS.

Analysis of the data will be performed with statistical methods and provide the significance, which is expected to be \( p < 0.05 \), thus rejecting the null hypothesis. The results of the evaluation, positive or negative, will contribute to the active research of Intelligent Tutoring Systems in Language Learning.

6 Research Plan

To achieve the objectives defined in the previous paragraphs the tasks were broken into eight parts. All the tasks are depended from their previous, except the T8, which starts during the middle of T6. As first day of the plan we perceive the 29th of May and as last the 15th of August. The eight tasks are given in the list below:

- **T1: Expert Knowledge Acquisition (1 week)** The project’s preliminary task involves discussing and receiving feedback from experienced teachers of French. The information acquired will become the basis for the knowledge base.

- **T2: Design (1 week)** The task involves setting the system specifications, in respect of software and knowledge engineering. It will take into account the different steps needed in order to build the prototype, and at a later stage, the final implementation.

- **T3: Prototype Build (2 weeks)** The basic system will be implemented in this step. This will be the basis for the final implementation.

- **T4: Formative Evaluation (1 week)** During this step, the prototype will be given to teachers and students for testing. Feedback is expected in order to fine-tune the project, as well as for debugging purposes. The final design of the system will be completed in this step.

- **T5: Intelligent Tutor Build (1 week)** Based on the results from previous steps, the Intelligent Tutor will be implemented and published on the internet. A lesser version of the system will be developed for use with the test group, as described in previous paragraphs.

- **T6: Evaluation (2 weeks)** Evaluating the system involves three separate activities with the human subjects: conducting a pre-test in order to classify the subjects according to their skills in the two test groups, the actual system testing, and lastly, the post-test.
**T7: Results Analysis (1 week)** The data of the pre-test and post-test will be analysed by a statistical package.

**T8: Dissertation (20 days)** The final part of the project entails the finishing of the write-up of the MSc dissertation and the presentation of the results.

**References**


