Modeling the Acquisition of English: an Intelligent CALL Approach *

Lisa N. Michaud, Kathleen F. McCoy, and Litza A. Stark

{michaud,mccoy,stark}@cis.udel.edu
Department of Computer and Information Sciences
University of Delaware, Newark, DE 19716
http://www.eecis.udel.edu/research/icicle

Abstract. In this paper, we present a methodology for the development of a user model for CALL which captures various levels of language acquisition using individualized overlays supported with stereotypes. Our current focus is the empirical analysis of the order of written English grammatical structure acquisition in our learner population used to develop stereotype layers in our model.

1 ICICLE: An Introduction

We are currently developing the system ICICLE (*Interactive Computer Identification and Correction of Language Errors*) to take a Computer-Assisted Language Learning (CALL) approach toward tutoring deaf students on their written English [12, 13]. Our target learners are native or near-native users of American Sign Language (ASL), a language entirely distinct from English, so we approach the acquisition of skills in written English as a second language (L2) acquisition task. This allows us to develop a system for a population which can greatly benefit from access to individualized, adaptive tutoring [13] while contributing to the field of CALL a design and methodology which can be also generalized to other L2 learners.

Our system is a writing tutor intended to supplement classroom instruction by providing students with detailed feedback on the grammatical errors in their English compositions. Its primary concerns are the correct analysis of student-generated language errors and the production of tutorial feedback to student performance which is both correct and tailored to the student. Its interaction with a user begins when the user submits a composition to the system, which is analyzed for grammatical errors. The system then responds with tutorial feedback aimed at enabling the student to perform corrections. When the student has revised the piece, it can be re-submitted for analysis. As ICICLE is intended to be used by an individual over time and across many pieces of writing, the system should learn as much as it can about a user to perform the language analysis and constructive feedback as accurately as possible.

* This work has been supported by NSF Grants #GER-9354869 and #IIS-9978021.

1.1 User Modeling in SLALOM

ICICLE's current implementation is a windows-based application with a text parser that uses an English grammar augmented by a bug catalogue or "mal-rules" capturing typical errors made by our learner population [15]. The system recognizes and marks many grammatical errors, delivering "canned" one- or two-sentence explanations of each error on request. When the system finds more than one possible analysis of the grammar structure underlying a user's sentence, it currently chooses arbitrarily. Since the selection of parses may determine which error(s) it assumes the user has made, and the instructive text is canned, the system currently lacks adaptivity to its user. To change this, we intend to implement in ICICLE a model of the system user, based on our view of the student as an L2 learner.

Originally proposed in [11], SLALOM (Steps of Language Acquisition in a Layered Organization Model) captures the user's ability to use each of the grammatical "rules" of English in a hybrid of an overlay model and a stereotype-based model; each of the rules in the model is marked based on the system's observations of student performance, and this information will be supplemented by stereotype information about typical language learners in this population. The model will therefore reflect those rules which the student uses in his or her language production—correct English rules from those English structures the user has acquired, and mal-rules for those structures which the user has not yet learned. In addition, there may be some structures which exhibit variation between correct and incorrect form, introducing competing rules (both standard and mal-rule) into the model. We consider the realm of grammar covered by these competing rules to correspond to Vygotsky's *Zone of Proximal Development* (ZPD), essentially that subset of language which the learner is about to master [17]. Krashen's observation that at each step of language learning there is some set of grammar rules which the learner is "due to acquire" [8], and the fact that elements which are on the verge of being acquired vacillate between correct and incorrect applications (cf. [7]), effectively reinforce the application of this concept to our domain.

The system's decision on how to interpret a user's text when there are multiple possibilities must depend upon the proficiency level of the learner and which rules are present in the user model. As this model represents those rules and mal-rules the user typically uses, parses can be selected whose hierarchical structure is composed of rules which most closely mirror those in the user model.

Once the text has been analyzed, ICICLE must generate a tutorial session, beginning by determining which of the errors will be the subjects of tutorial explanations. This decision is important if the instruction is to be effective. There should be a distinction in how the system addresses language "errors," which reflect grammatical competence, versus "mistakes," which are merely slip-ups [3]. We also want to avoid generating instruction which is beyond the user's understanding. The concept of "learnability" in second language instruction constrains knowledge that can be assimilated to those concepts the learner is ready to acquire [6]. We therefore focus instruction on that "narrow shifting zone dividing

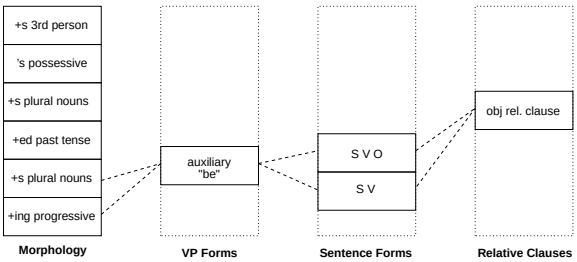

Fig. 1. SLALOM: Steps of Language Acquisition in a Layered Organization Model.

the already-learned skills from the not-yet-learned ones" [10], or the frontier of the learning process. ICICLE will select those errors which involve items in the ZPD and ignore both simple inadvertent mistakes by the user and errors that are beyond the student's frontier of learning.

2 Compensating for Partial Evidence: Stereotypes in SLALOM

We have presented a model reliant on the recorded tendencies of the user with respect to grammatical forms in English. This knowledge will sometimes be only partial, particularly at the beginning of the system's interaction with a user. We therefore must establish a method by which the system can infer a fuller description of user proficiency than is directly displayed in his or her past use of language forms. We propose to do this using a structure of relationships between language concepts based on language learning stereotypes.

We first proposed a structure for SLALOM in [11]. Here we discuss how a more mature version of this model can be used by ICICLE to fill in the gaps in user observations. We partition grammatical knowledge into knowledge units where each unit corresponds to a set of grammar rules (both standard English and mal-rules) that realize a grammatical structure. A simplified representation can be seen in Fig. 1.

Steps of Language Acquisition refers to our intention to capture the order of acquisition of these knowledge units. There is empirical support for stereotypical sequences of English acquisition [5, 9, 8], and our model reflects this stereotype by grouping units into related "hierarchies" (such as morphology markings, verb phrase constructions, and relative clause formation), each of which has an order represented in the figure by a vertical relationship; "easier" items which are typically acquired earlier sit below those acquired later. The example hierarchy in Fig. 1 demonstrates a possible SLALOM hierarchy based on the results of empirical work by [5] on the acquisition order of morphology markers.

The *Layered Organization Model* part of SLALOM's design is shown in the figure by the dashed lines coordinating the acquisition steps across the hierarchies by indicating a "layer" of concurrent acquisition; elements connected at the

same layer are acquired at about the same time. Intuitively, one layer represents the structures currently being learned by the user. Because of the ordered nature of the hierarchies, those items below that layer have typically already been acquired, while those above have not been acquired.[1]

With the assistance of this stereotype information, ICICLE's syntactic analyzer will be enabled to make parse selections even when data on this individual does not cover all of the rules in the parsing grammar. A grammatical structure corresponding to a knowledge unit below those considered "acquired" or "ZPD" in the model can be considered previously acquired, and thus covered by a standard English rule in the grammar; conversely, those structures above the user's frontier of learning should be covered by mal-rules, and those in the same layer as the ZPD should be flexible, covered by either correct or incorrect rules.

Since this model is based on observations of an individual user, it will be initialized following the first analysis of a new user's writing. As actual data on the user's performance is delivered to the model over subsequent analyses, the inferences provided by the stereotype will be overwritten to reflect the individual, allowing the system to adapt to users who may learn outside of the stereotypical sequence due to different instructional programs. The model is also dynamic, maintaining statistics on user performance within a certain window of the present, including enough information from previous writing samples to give us a more complete view of user performance while excluding errors that are no longer being made. Because SLALOM's ordering of layers represents an expected order of acquisition, the most likely path of the ZPD layer is to move "up" in the model as the user learns.

3 Populating the User Model: An Empirical Study

A question not addressed in [11] has to do with "populating" the SLALOM model, determining where the knowledge units of the model go in our representation of stereotypic acquisition order. As mentioned in Section 2, support for the existence of such an order can be found in empirical studies on the acquisition of English. However, although there is support for a universal order of morpheme acquisition regardless of L1 (first language) [5], no existing work proposes an order for the entire body of grammatical structures covered in SLALOM and none addresses our learner group in particular.

The remainder of this paper describes how we have taken on the task of empirically deriving a stereotypic acquisition order for our own user population, whose L1 is American Sign Language. We seek to identify: (1) which aspects of English are mastered in what order, and (2) what groups of items are learned around the same time. This will allow us to group SLALOM contents into layers and impose upon those layers some stereotypic order.

[1] Note that what is considered a "layer" may be much larger than just one item per hierarchy. The important aspect of the definition is that each layer represents a grouping of language rules acquired at about the same time. The layer in the figure is for example purposes only and does not reflect any empirical findings.

id	Incorrect Determiner	mds	Missing Dummy Subject
md	Missing Determiner	ids	Incorrect Dummy Subject
sv	Subject/Verb agreement	mo	Missing Object
ht	Here/There as a Pronoun	bh	Be/Have Confusion
nf	Noun Formation	ii	Incorrect Intensifier

Fig. 2. Examples of codes from our error coding manual.

Toward that end, we have been examining a corpus of 106 samples of writing by deaf college students at various levels of English proficiency. Since these samples represent different levels of grammatical ability, obtaining syntactic analyses of the samples would ideally provide us with information on which aspects of English are being executed with what levels of success across the different competence levels because the resulting parse trees would explicitly contain correct or mal-formed structures. However, we are still faced with the problem that the current analyzer (without a completed SLALOM) has no intelligent way to select between alternative parses. We therefore first need to provide the benefit of human intuition in distinguishing what interpretation of each sentence is the most accurate. From this, we can build a profile of performance across the levels represented by our corpus.

3.1 Grading User Performance

To provide the system with human judgments on interpretations of user text, we marked up our corpus by hand to indicate what errors had occurred in each sentence. For this task, we developed a taxonomy of "error codes," each of which addresses errors typical of our user population, based on the initial corpus analysis presented in [16, 11]. Our codes have a close relationship to the mal-rules which the parser uses to recognize errors. The completed taxonomy had 68 error codes, primarily covering syntactic structures. Some example codes are shown in Fig. 2.[2]

We then divided the 106 samples between two coders, each receiving approximately half, with an overlap of 23 samples which both individuals coded in order to demonstrate consistency between the two perspectives.

To illustrate that our two coders were operating consistently with respect to each other, we needed to determine some measure of agreement between them. One difficulty we faced in this task was that each sentence was tagged with a string of codes (see Fig. 3) which was organized linearly according to the order in which the errors occurred in the sentence but were not otherwise connected to the specific incidents they recorded. Furthermore, the strings tended to have gaps; when structures involved gray areas of grammaticality and were ungrammatical

[2] "Dummy Subject" refers to subjects which are not referents: There are books on the table / It is nice to see you.

Example:
Those who argue that it will have less hazing incidents here on campus if the abolishment of fraternities and sororities are done.
Coder 1: (mds ids bh ii sv) *There are* those who argue that *there* will *be fewer* hazing incidents here on campus if the abolishment of fraternities and sororities *is* done.
Coder 2: (mds ids sv) *There are* those who argue that *there* will *be* less hazing incidents here on campus if the abolishment of fraternities and sororities *is* done.

Alignment:

mds	ids	bh	ii	sv
mds	ids	*	*	sv

Fig. 3. Examples of coders' tags and alignment program output.

to one coder but acceptable to the other, a code in one string would not have a correspondent in the other. Therefore, we faced a task of comparing for each sentence in our corpus two lists of error codes without knowing which pairs of codes referred to the same error.

For this problem we borrowed a page from bioinformatics research, which has developed many algorithms to compare two strings of DNA in order to determine the best alignment. In particular, we adapted the Smith-Waterman algorithm, which computes a matrix of alignment scores where matches are rewarded and mismatches or gaps introduced in one or the other string are penalized. After we tuned the penalties and rewards to reflect the relative importance of matches, mismatches, and gaps in our particular problem, we applied this algorithm to the codes for the 23 overlap compositions. The alignment program produced a very accurate alignment between the two coders' lists, consistent with human judgment. An example output can be seen in Fig. 3.

Disagreement on the grammaticality of user text was the largest source of conflict between our coders, and because both were equally qualified judges and either interpretation could be correct, we discounted gaps as disagreements. In some of the gap cases, one coder had a different semantic interpretation of the changes required to bring the sentence into consistency with the surrounding discourse, such as the tense context (should these verbs have been in present or past tense?) or the discourse entities (should "friend" be "friends?"). In others, including the example in Fig. 3, a code was absent from one string because of a confusion on the definition of what needed to be explicitly coded. Fig. 4 shows agreement between the coders when both coders agreed an error occurred (i.e., ignoring gaps). The number of "possible" codes for each sentence was determined by the total length of the alignment sequence.

Measuring the frequency with which each of the coders used each of the 68 error codes, we determined the likelihood of chance agreement over this task and calculated an adjusted "Kappa" value [2] of .78. Although this is just inside the margin of what Carletta names the range of tentative conclusions ($.67 < K < .8$), we are satisfied with the agreement level because it is strong enough to indicate

794	Total Possible Errors
432	Both Coders Agreed Were Errors
352	Both Coders Assigned Same Code
81%	Bare Agreement
78%	Adjusted Agreement (Kappa)

Fig. 4. Agreement statistics on overlapped coding.

that our coders are providing our study with similar "human intuitions" in a domain which has a lot of gray area, and if our parser sided with either coder's interpretation, it would still be judging consistently with an experienced human; that is the best that we could require.

3.2 Clustering User Groups

The next step in our corpus analysis was to determine whether the coded samples were stratified in any way that corresponded to a judgment of general proficiency level. An important step in this was to have levels assigned by experienced judges. We distributed the samples to four instructors at the English Language Institute of the University of Delaware who are trained to grade compositions according to the standards of the national Test of Written English (TWE). Each sample was given a TWE score from 1 to 6 by two different judges, and a third judge arbitrated in the case of disagreement. We then prepared "error count" sets for each sample, indicating the number of occurrences of each of our error codes, normalized by the number of sentences in the sample in order to compensate for the fact that our samples varied greatly in length (from 2 to 58 sentences). We applied several clustering algorithms on the normalized data, seeking out clusters in which samples were minimally distant from each other, and so should represent samples which have the same errors in approximately the same magnitude.[3]

We show the results of running the Ward clustering algorithm on our data in Fig. 5. Although we have a problem with sparse data (only 5% of the samples occurring at TWE levels 1, 5, and 6), there is a clear trend with lower and higher proficiency levels showing a preference to different clusters, overlapping in Cluster 2. Together with MANOVA analyses we have run which further indicate trends of change between those errors committed by students at different levels, we have confirmed Corder's statement [3] that errors are a clue to the learner's current state of acquisition; however, we have also discovered that errors alone are not enough to precisely characterize the learner.

[3] The ATULA-ATS system [14], uses statistical clustering to develop user groups into which to categorize a new user, based on non-knowledge attributes like attitude and background.

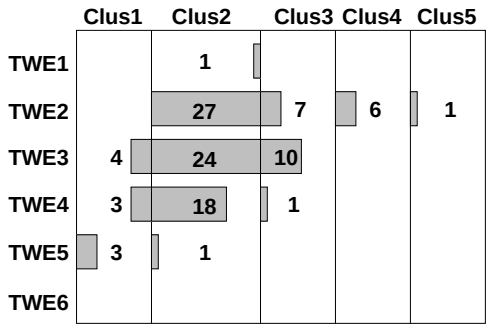

Fig. 5. Distribution of levels of proficiency across Ward clusters.

4 Future Work

Only by obtaining data on those structures a student can execute correctly will we be able to look at errors meaningfully by comparing the failed attempts to use a structure against the successes. We are currently in the process of adapting our syntactic analyzer so that we may explore the competing parse trees created for each input sentence in our corpus. Since we have specified the relationship between each of the mal-rules of our parsing grammar and our list of error codes, each parse tree can be converted to a string of error codes. These strings can then be compared against the human-generated codes to select which parse is the closest to a human's judgment using our alignment program.

With the "correct" parses indicated, we will know for each sentence which grammar constituents were formed correctly and which were made from mal-rules. From this we can derive a very detailed account of syntactic performance for each sample, supplementing our error information with the elements of English the user has executed *correctly* as well. Stratified into the levels determined by our expert judges, this will flesh out our view of syntactic performance as proficiency develops, and give us our partial orders of acquisition on which to base the inferencing relationships in SLALOM; structures mastered by the lowest-proficiency writers will be grouped in the lowest layers of the model, structures mastered only by the higher-proficiency writers grouped toward the top, and indications of transitional performance will indicate which structures will be considered part of a single layer.

5 Related Work

The overlay/stereotype hybrid student model design is not unheard of in ITS student model design. Desmarais et. al's POKS (Partial Order Knowledge Structures) approach [4] builds an inferencing network incorporating order of acquisition information derived automatically from training data. Their approach, however, is strongly influenced by the fundamental assumption that while the

user's mastery is measurable on any given knowledge unit, there is a poverty of such sampling because measurement must be done through explicitly querying the user. Our system, by contrast, will have access to large sets of user knowledge data via unobtrusive observation through the analysis process. SLALOM's partial ordering is both more robust (being based on a larger corpus with broadly-sampled data) and less relied upon, as the abundant individual data will always take precedence over inferences.

In the field of CALL, the student model Mr. Collins [1], has a design methodology similar to ours; the dynamic student model reflects more correct knowledge and fewer misconceptions as it moves toward the expert ideal. Also, a review of typical student errors in their learner group was applied toward the model's ability to expect future user performance and toward diagnosing the errors of the student. There are several distinctions between their approach and ours, particularly that Mr. Collins was developed in the restricted domain of the placement and use of Portuguese clitic pronouns, whereas the work we have described in this paper addresses student development over a broad range of grammatical structures. The system containing Mr. Collins is also designed to interact at a meta-level with the student, collaborating on the content of the model and explicitly providing tools for the student's learning strategies; since ICICLE does not have the liberty of interaction in the student's native language[4], our ability to communicate topics at the meta-level is perforce constrained.

6 Summary

We have embarked upon the implementation of a complex student model for a language instruction system and have presented a methodology for the empirical derivation of stereotypes for learners of written English as a second language. Since data on the specific user will be numerous, this stereotype information will only be used to supplement individual overlay data which will be more reliable as the system interacts with the individual. Since ICICLE is modular and only isolated aspects of the system (the mal-rules in the analyzer and the inferencing in SLALOM) are specific to the target user population, it is our intention that the methodology we have developed could be applied to adapt ICICLE to any L1 user group.

References

[1] Susan Bull, Paul Brna, and Helen Pain. Extending the scope of the student model. *User Modeling and User-Adapted Interaction*, 5(1):45–65, 1995.

[2] Jean Carletta. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254, June 1996.

[3] S. P. Corder. The significance of learners' errors. *International Review of Applied Linguistics*, 5(4):161–170, November 1967.

[4] We are currently investigating the integration of on-screen signed instruction.

[4] Michel C. Desmarais, Ameen Maluf, and Jiming Jiu. User-expertise modeling with empirically derived probabilistic implication networks. *User modeling and user-adapted interaction*, 5(3/4):283–315, 1996.

[5] Heidi C. Dulay and Marina K. Burt. Natural sequences in child second language acquisition. *Language Learning*, 24(1), 1975.

[6] Rod Ellis. The structural syllabus and second language acquisition. *TESOL Quarterly*, 27(1):91–113, Spring 1993.

[7] Rod Ellis. *The Study of Second Language Acquisition*. Oxford University Press, New York, 1994.

[8] Stephen D. Krashen. *Principles and Practice in Second Language Acquisition*. Pergamon Press, New York, 1982.

[9] Diane E. Larsen-Freeman. An explanation for the morpheme acquisition order of second language learners. *Language Learning*, 25(1):125–135, June 1976.

[10] Frank Linton, Brigham Bell, and Charles Bloom. The student model of the LEAP intelligent tutoring system. In *Proceedings of the Fifth International Conference on User Modeling*, pages 83–90, Kailua-Kona, Hawaii, January 2-5 1996. UM96, User Modeling, Inc.

[11] Kathleen F. McCoy, Christopher A. Pennington, and Linda Z. Suri. English error correction: A syntactic user model based on principled mal-rule scoring. In *Proceedings of the Fifth International Conference on User Modeling*, pages 59–66, Kailua-Kona, Hawaii, January 2-5 1996. UM96, User Modeling, Inc.

[12] Lisa N. Michaud and Kathleen F. McCoy. Supporting intelligent tutoring in CALL by modeling the user's grammar. In *Proceedings of the 13th Annual International Florida Artificial Intelligence Research Symposium*, pages 50–54, Orlando, Florida, May 22-24 2000. FLAIRS.

[13] Lisa N. Michaud, Kathleen F. McCoy, and Christopher A. Pennington. An intelligent tutoring system for deaf learners of written English. In *Proceedings of the Fourth International ACM SIGCAPH Conference on Assistive Technologies (ASSETS 2000)*, Washington, D.C., November 13-15 2000. SIGCAPH.

[14] Sue Milne, Edward Shiu, and Jean Cook. Development of a model of user attributes and its implementation with an adaptive tutoring system. *User modeling and user-adapted interaction*, 6(4):303–335, 1996.

[15] David Schneider and Kathleen F. McCoy. Recognizing syntactic errors in the writing of second language learners. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and the Seventeenth International Conference on Computational Linguistics*, volume 2, pages 1198–1204, Universite de Montreal, Montreal, Quebec, Canada, August 10-14 1998. COLING-ACL, Morgan Kaufmann Publishers.

[16] Linda Z. Suri and Kathleen F. McCoy. A methodology for developing an error taxonomy for a computer assisted language learning tool for second language learners. Technical Report TR-93-16, Dept. of Computer and Information Sciences, University of Delaware, 1993.

[17] Lev Semenovich Vygotsky. *Thought and Language*. The MIT Press, Cambridge, Massachusetts, 1986. Translation revised and edited by Alex Kozulin; originally published in 1934.