Error Profiling: Toward a Model of English Acquisition for Deaf Learners

Lisa N. Michaud and **Kathleen F. McCoy**
Dept. of Computer and Info. Sciences, University of Delaware, Newark, DE 19716, USA
{michaud,mccoy}@cis.udel.edu
http://www.eecis.udel.edu/research/icicle

Abstract

In this paper we discuss our approach toward establishing a model of the acquisition of English grammatical structures by users of our English language tutoring system, which has been designed for deaf users of American Sign Language. We explore the correlation between a corpus of error-tagged texts and their holistic proficiency scores assigned by experts in order to draw initial conclusions about what language errors typically occur at different levels of proficiency in this population. Since errors made at lower levels (and not at higher levels) presumably represent constructions acquired before those on which errors are found only at higher levels, this should provide insight into the order of acquisition of English grammatical forms.

1 Introduction

There have been many theories of language acquisition proposing a stereotypical order of acquisition of language elements followed by most learners, and there has been empirical evidence of such an order among morphological elements of language (cf. (Bailey et al., 1974; Dulay and Burt, 1975; Larsen-Freeman, 1976)) and some syntactic structures (cf. (Brown and Hanlon, 1970; Gass, 1979)). There is indication that these results may be applied to any L1 group acquiring English (Dulay and Burt, 1974; Dulay and Burt, 1975), and some research has focused on developing a general account of acquisition across a broad range of morphosyntactic structures (cf. (Pienemann and Håkansson, 1999)). In this work, we explore how our second language instruction system, ICICLE, has generated the need for modeling such an account, and we discuss the results of a corpus analysis we have undertaken to fulfill that need.

1.1 ICICLE: an Overview

ICICLE (*Interactive Computer Identification and Correction of Language Errors*) is an intelligent tutoring system currently under development (Michaud and McCoy, 1999; Michaud et al., 2000; Michaud et al., 2001). Its primary function is to tutor deaf students on their written English. Essential to performing that function is the ability to correctly analyze user-generated language errors and produce tutorial feedback to student performance which is both correct and tailored to the student's language competence. Our target learners are native or near-native users of American Sign Language (ASL), a distinct language from English (cf. (Baker and Cokely, 1980)), so we view the acquisition of skills in written English as the acquisition of a second language for this population (Michaud et al., 2000).

Our system uses a cycle of user input and system response, beginning when a user submits a piece of writing to be reviewed by the system. The system determines the grammatical errors in the writing, and responds with tutorial feedback aimed at enabling the student to perform corrections. When the student has revised the piece, it is re-submitted for analysis and the cycle begins again. As ICICLE is intended to be used by an individual over time and across many pieces of writing, the cycle will be repeated with the same individual many times.

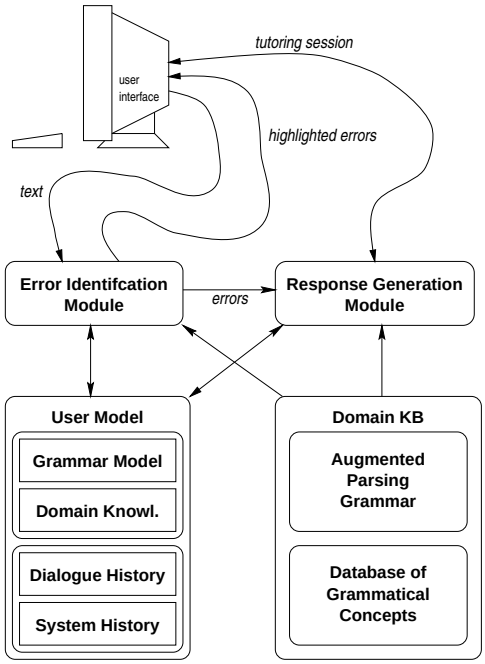

Figure 1: ICICLE system architecture.

Figure 1 contains a diagram of ICICLE's overall architecture and the cycle we have described. It executes between the User Interface, the Error Identification Module (which performs the syntactic analysis of user writing), and the Response Generation Module (which builds the feedback to the user based on the errors the user has committed). The work described in this paper focuses on the development of one of the sources of knowledge used by both of these processes, a component of the User Model representing the user's grammatical competence in written English.

1.2 A Need for Modeling L2A Status

What currently exists of the ICICLE system is a prototype application implemented in a graphical interface connected to a text parser that uses a wide-coverage English grammar augmented by "mal-rules" capturing typical errors made by our learner population. It can recognize and label many grammatical errors, delivering "canned" one- or two-sentence explanations of each error on request. The user can then make changes and resubmit the piece for additional analysis. We have discussed in (Schneider and McCoy, 1998) the performance of our parser and mal-rule-augmented grammar and the unique challenges

"She is teach piano on Tuesdays."
Beginner: Inappropriate use of auxiliary and verb morphology problems. "She teaches piano on Tuesdays."
Intermediate: Missing appropriate +ing morphology. "She is teaching piano on Tuesdays."
Advanced: Botched attempt at passive formation. "She is taught piano on Tuesdays."

Figure 2: Possible interpretations of non-grammatical user text.

faced when attempting to cover non-grammatical input from this population.

In its current form, when the parser obtains more than one possible parse of a user's sentence, the interface chooses arbitrarily which one it will assume to be representative of which structures the user was attempting. This is undesirable, as one challenge that we face with this particular population is that there is quite a lot of variability in the level of written English acquisition. A large percentage of the deaf population has reading/writing proficiency levels significantly below their hearing peers, and yet the population represents a broad range of ability. Among deaf 18-year-olds, about half read at or below a fourth grade level, while about 10% read above the eighth-grade level (Strong, 1988). Thus, even when focused on a subset of the deaf population (e.g., deaf high school or college students), there is significant variability in the writing proficiency. The impact of this variability is that a particular string of words may have multiple interpretations and the most likely one may depend upon the proficiency level of the student, as illustrated in Figure 2. We are therefore currently developing a user model to address the system's need to make these parse selections intelligently and to adapt tutoring choices to the individual (Michaud and McCoy, 2000; Michaud et al., 2001).

The model we are developing is called SLALOM. It is a representation of the user's ability to correctly use each of the grammatical "features" of English, which we define as incorporating both morphological rules such as plural-

izing a noun with **+S** and syntactic rules such as the construction of prepositional phrases and **S V O** sentence patterns. Intuitively, each unit in SLALOM corresponds to a set of grammar rules and mal-rules which realize the feature. The information stored in each of these units represents observations based on the student's performance over the submission of multiple pieces of writing. These observations will be abstracted into three tags, representing performance that is consistently good (*acquired*), consistently flawed (*unacquired*), or variable (*ZPD*[1]) to record the user's ability to correctly execute each structure in his or her written text.

1.3 An Incomplete Model

A significant problem that we must face in generating the tags for SLALOM elements is that we would like to infer tags on performance elements not yet seen in a writer's production, basing those tags on what performance we have been able to observe so far. We have proposed (McCoy et al., 1996; Michaud and McCoy, 2000) that SLALOM be structured in such a way as to capture these expectations by explicitly representing the relationships between grammatical structures in terms of when they are acquired; namely, indicating which features are typically acquired *before* other features, and which are typically acquired *at the same time*. With this information available in the model, SLALOM will be able to suggest that a feature typically acquired before one marked "acquired" is most likely also acquired, or that a feature co-acquired with one marked "ZPD" may also be something currently being mastered by the student. The corpus analysis we have undertaken is meant to provide this structure by indicating a partial ordering on the acquisition of grammatical features by this population of learners.

1.4 Applications

Having the SLALOM model marked with grammatical features as being acquired, unacquired, or ZPD will be very useful in at least two different ways. The first is when deciding which possible parse of the input best describes a particular sentence produced by a learner. When there are multiple parses of an input text, some may place the "blame" for detected errors on different constituents. In order for ICICLE to deliver relevant instruction, it needs to determine which of these possibilities most likely reflects the actual performance of the student. We intend for the parse selection process to proceed on the premise that future user performance can be predicted based on the patterns of the past. The system can generally prefer parses which use rules representing well-formed constituents associated with "acquired" features, mal-rules from the "unacquired" area, and either correct rules or mal-rules for those features marked "ZPD."

A second place SLALOM will be consulted is in deciding which errors will then become the subjects of tutorial explanations. This decision is important if the instruction is to be effective. It is our wish for ICICLE to ignore "mistakes" which are slip-ups and not indicative of a gap in language knowledge (Corder, 1967) and to avoid instruction on material beyond the user's current grasp. It therefore will focus on features marked "ZPD"—those in that "narrow shifting zone dividing the already-learned skills from the not-yet-learned ones" (Linton et al., 1996), or the frontier of the learning process. ICICLE will select those errors which involve features from this learner's learning frontier and use them as the topics of its tutorial feedback.

With the partial order of acquisition represented in the SLALOM model as we have described, these two processes can proceed on the combination of the data contained in the previous utterances supplied by a given learner and the "intuitions" granted by information on typical learners, supplementing empirical data on the specific user's mastery of grammatical forms with inferences on what that means with respect to other forms related to those through the order of acquisition.

2 Profiling Language Errors

We have established the need for a description of the general progress of English acquisition as determined by the mastery of grammatical forms.

[1] Zone of Proximal Development: see (Michaud and McCoy, 2000) for discussion. These are presumably the features the learner is currently in the process of acquiring and thus we expect to see variation in the user's ability to execute them.

We have undertaken a series of studies to establish an order-of-acquisition model for our learner population, native users of American Sign Language.

In our first efforts, we have been guided by the observation that the errors committed by learners at different stages of acquisition are clues to the natural order that acquisition follows (Corder, 1967). The theory is that one expects to find errors on elements currently being acquired; thus errors made by early learners and not by more advanced learners represent structures which the early learners are working on but which the advanced learners have acquired. Having obtained a corpus of writing samples from 106 deaf individuals, we sought to establish "error profiles"—namely, descriptions of the different errors committed by learners at different levels of language competence. These profiles could then be a piece of evidence used to provide an ordering structure on the grammatical elements captured in the SLALOM model.

This is an overview of the process by which we developed our error profiles:

Goal : to have error profiles that indicate what level of acquisition is most strongly associated with which grammatical errors. It is important that the errors correspond to our grammar mal-rules so that the system can prefer parses which contain the errors most consistent with the student's level of acquisition.

Method :

1. Collect writing samples from our user population
2. Tag samples in a consistent manner with a set of error codes (where these codes have an established correspondence with the system grammar)
3. Divide samples into the levels of acquisition they represent
4. Statistically analyze errors within each level and compare to the magnitude of occurrence at other levels
5. Analyze resulting findings to determine a progression of competence

In (Michaud et al., 2001) we discuss the initial steps we took in this process, including the development of a list of error codes documented by a coding manual, the verification of our manual and coding scheme by testing inter-coder reliability in a subset of the corpus (where we achieved a Kappa agreement score (Carletta, 1996) of $K = .78$)[2], and the subsequent tagging of the entire corpus. Once the corpus was annotated with the errors each sentence contained, we obtained expert evaluations of overall proficiency levels performed by ESL instructors using the national Test of Written English (TWE) ratings[3]. The initial analysis we go on to describe in (Michaud et al., 2001) confirmed that clustering algorithms looking at the relative magnitude of different errors grouped the samples in a manner which corresponded to where they appeared in the spectrum of proficiency represented by the corpus. The next step, the results of which we discuss here, was to look at each error we tagged and the ability of the level of the writer's proficiency to predict which errors he or she would commit. If we found significant differences in the errors committed by writers of different TWE scores, then we could use the errors to help organize the SLALOM elements, and through that obtain data on which errors to expect given a user's level of proficiency.

2.1 Toward an error profile

Although our samples were scored on the six-point TWE scale, we had sparse data at either end of the scale (only 5% of the samples occurring in levels 1, 5, and 6), so we concentrated our efforts on the three middle levels (2, 3, and 4), which we renamed *low, middle*, and *high*.

Our chosen method of data exploration was Multivariate Analysis of Variance (MANOVA). An initial concern was to put the samples on equal footing despite the fact that they covered a broad range in length—from 2 to 58 sentences—and there was a danger that longer samples would tend

[2]We also discuss why we were satisfied with this score despite only being in the range of what Carletta calls "tentative conclusions."

[3]Although these samples were relatively homogeneous with respect to the amount of English training and the age of the writer, we expected to see a range of demonstrated proficiency for reasons discussed above. We discuss later why the ratings were not as well spread-out as we expected.

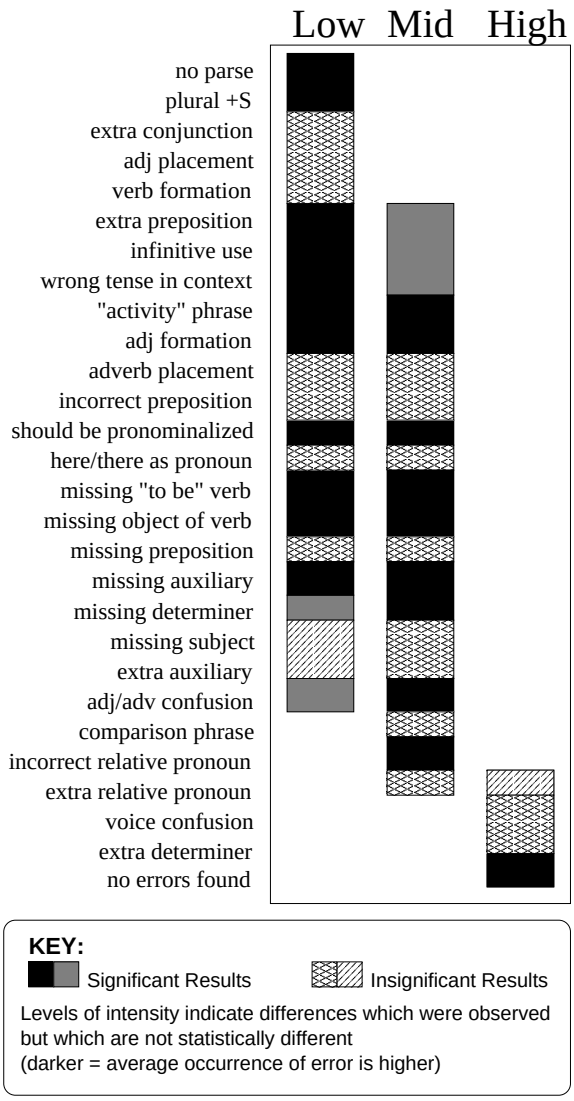

Figure 3: Illustrating the errors each level is most likely to commit.

to have higher error counts in every category simply because the authors had more opportunity to make errors. We therefore used two dependent variables in our analysis: the TWE score and the length of the sample, testing the ability of the two combined to predict the number of times a given error occurred. We ran the MANOVA using both sentence count and word count as possible length variables, and in both runs we obtained many statistically significant differences between the magnitude at which writers at different TWE levels committed certain errors. These differences are illustrated in Figure 3, which shows the results on a subset of the 47 error code tags for which we got discernible results[4].

In the figure, a bar indicates that this level of proficiency committed this type of error more frequently than the others. If two of the three levels are both marked, it means that they both committed the error more frequently than the third, but the difference between those two levels was unremarkable. Solid shading indicates results which were statistically significant (with an omnibus test yielding of significance level of $< .05$), and intensity differences (e.g., black for *extra preposition* in the low level, but grey in the middle level) indicate a difference that was not significant. In the example we just mentioned, the low-level writers committed more *extra preposition* errors than the high-level writers with a significance level of 0.0082, and the mid-level writers also committed more of these errors than the high-level writers with a significance of .0083. The comparison of the low and middle levels to each other, on the other hand, showed that the low-level learners committed more of this error, but that the result was strongly insignificant at .5831.

The cross-hatched and diagonal-striped results in the figure indicate results which did not satisfy the cutoff of $< .05$ for significance, but were considered both interesting and close enough to significance to be worth noting. The diagonal stripes have "less intensity" and thus indicate the same relationship to the cross-hatched bars as the gray does to the black—a difference in the data which indicates a lower occurrence of the error which is not significantly distinguished (e.g., high-level learners committed *extra relative pronoun* errors less often than mid-level learners, and both high- and mid-level learners committed it more often than the low-level learners), but, again, not to a significant extent.

Notice that the overall shape of the figure supports the notion of an order of acquisition of features because one can see a "progression" of errors from level to level. Very strongly supportive of this intuition are the first and last errors in the figure: "no parse," indicating that the coder

[4] "Activity" refers to the ability to correctly form a gerund-fronted phrase describing an activity, such as "I really like walking the dog;" "comparison phrase" refers to the formation of phrases such as "He is smarter than she;" "voice" refers to the confusion between using active and passive voice, such as "The soloist was sung."

was unable to understand the intent of the sentence, statistically more often at the lowest level than the at the other two levels, while "no errors found" was significantly most prevalent at the highest level (both with a significance level of $< .0001$).

Other data which is more relevant to our goals also presents itself. The lowest level exhibited higher numbers of errors on such elementary language skills as putting plural markers on nouns, placing adjectives before the noun they modify, and using conjunctions to concatenate clauses correctly. Both the low and middle levels struggled with many issues regarding forming tenses, and also exhibited "ASLisms" in their English, such as the dropping of constituents which are either not explicitly realized in ASL (such as determiners, prepositions, verb subjects and objects which are established discourse entities in focus, and the verb "TO BE"), or the treatment of certain discourse entities as they would be in ASL (e.g., using "here" as if it were a pronoun). While beginning learners struggled with more fundamental problems with subordinate clauses such as missing gaps, the more advanced learners struggled with using the correct relative pronouns to connect those clauses to their matrix sentence. Where the lower two levels committed more errors with missing determiners, the highest level among our writers had learned the necessity of determiners in English but was over-generalizing the rule and using them where they were not appropriate. Finally, the upper level learners were beginning to experiment with more complex verb constructions such as the passive voice. All of this begins to draw a picture of the sequence in which these structures are mastered across these levels.

2.2 Discussion

While Figure 3 is meant to illustrate how the three different levels committed different sets of errors, it is clear that this picture is incomplete. The low and middle levels are insufficiently distinguished from each other, and there were very few errors committed most often by the highest level. Most importantly, many of the distinctions between levels were not achieved to a significant degree.

One of the reasons for these problems is the fact that our samples are concentrated in only three levels in the center of the TWE spectrum. We hope to address this in the future by acquiring additional samples. Another problem which additional samples will help to solve is sparseness of data. Across our 106 samples and 68 error codes, only 30 codes occur more than 25 times in the corpus, and only 21 codes occur more than 50 times. Most of our insignificant differences come from error codes with very low frequency, sometimes occurring as infrequently as 7 times.

What we have established is promising, however, in that it does show statistically significant data spanning nearly every syntactic category. Additional samples must be collected and analyzed to obtain more statistical significance; however, the methodology and approach are proven solid by these results.

3 Future Work: Performance Profiles

If we were to stop here, then our user model design would simply be to group the SLALOM contents addressed by these errors in an order according to how they fell into the distribution shown in Figure 3, assuming essentially that those errors falling primarily in the low-level group represent structures that are learned first, followed by those in the low/middle overlap area, followed by those which mostly the mid-level writers were struggling, followed finally by those which mostly posed problems for our highest-level writers.

Given this structure, and a general classification of a given user, if we are attempting to select between competing parses for a sentence written by this user, we can prefer a sentence whose errors most closely fit those for the profile to which the user belongs. However, up until now we have only gathered information on the *errors* committed by our learner population, and thus we still have no information on a great deal of grammatical constructions. Consider that some types of grammatical constructions may be avoided or used correctly at low levels but that the system would have no knowledge of this. By only modeling the errors, we fail to capture the acquisition order data provided by knowing what structures a writer can *successfully* execute at the different levels. Therefore, the sparse data problems we faced in this work are only *partly* explained by the small corpus and some infrequent error codes.

They are also explained by the fact that errors are only one half of the total picture of user performance.

Although we experimented in this work with equalizing the error counts using different length measures, we did not have access to the numbers that would have provided the most meaningful normalization: namely, the number of times a structure is attempted. It is our belief that information on the successful structures in the users' writing would give us a much clearer view of the students' performance at each level. Tagging all sentences for the *correct* structures, however, is an intractable task for a human coder. On the other hand, while it is feasible to have this information collected computationally through our parser, we are still faced with the problem of competing parses for many sentences. Our methodology to address this problem is to use the human-generated error codes to select among the parses trees in order to gather statistics on fully-parsed sentences.

We have therefore created a modified version of our user interface which, when given a sample of writing from our corpus, records all competing parse trees for all sentences to a text file[5]. Another application has been developed to compare these system-derived parse trees against the human-assigned error code tags for those same sentences to determine which tree is the closest match to human judgment. To do this, each tree is traversed and all constituents corresponding to mal-rules are recorded as the equivalent error code tag. The competing lists of errors are then compared against the sequence determined by the human coder via a string alignment/comparison algorithm which we discuss in (Michaud et al., 2001).

With the "correct" parse trees indicated for each sentence, we will know which grammar constituents each writer correctly executed and which others had to be parsed using our mal-rules. The same statistical techniques described above can then be applied to form *performance profiles* for capturing statistically significant differences in the grammar rules used by students within each level. This will give us a much more detailed description of acquisition status on language elements throughout the spectrum represented by our sample population.

The implication of having such information is that once it is translated into the structure of our SLALOM user model, performance on a previously-unseen structure may be predicted based on what performance profile the user most closely fits and what tag that profile typically assigns to the structure in question; as mentioned earlier in this text, features typically acquired before a structure on which the user has demonstrated mastery can be assumed to be acquired as well. Those structures which are well beyond the user's area of variable performance (his or her current area of learning) are most likely unacquired. Since we view the information in SLALOM as projecting probabilities onto the rules of the grammar, intuitively this will allow the user's mastery of certain rules to project different *default* probabilities on rules which have not yet been seen in the user's language usage.

With this information, ICICLE will then be able to make principled decisions in both parsing and tutoring tasks based on a hybrid of direct knowledge about the user's exhibited proficiency on grammatical structures and the indirect knowledge we have derived from typical learning patterns of the population.

4 Conclusion

In this paper we have addressed an empirical effort to establish a typical sequence of acquisition for deaf learners of written English. Our initial results show much promise and are consistent in many ways with intuition. Future work will apply the same methodology but expand beyond the analysis of user errors to the analysis of the complete image of user performance, including those structures which a user can successfully execute. When completed, our model will enable a complex tutoring tool to intelligently navigate through multiple competing parses of user text and to focus language instruction where it will do the most good for the learner, exhibiting a highly desirable adaptability to a broad range of users and addressing a literacy issue in a population who could greatly benefit from such a tool.

[5]Thanks are due to Greg Silber for his work on revising our interface and creating this variation.

Acknowledgments

This work has been supported by NSF Grants #GER-9354869 and #IIS-9978021. We would like to thank the readers at the English Language Institute for their expert judgments and Dr. H. Lawrence Hotchkiss at Research Data Management Services at the University of Delaware for his help with statistically analyzing our data. We would also like to thank the other members of the ICICLE group, including Matt Huenerfauth, Jill Janofsky, Chris Pennington, Litza Stark (one of our coders), and Greg Silber.

References

N. Bailey, C. Madden, and S. D. Krashen. 1974. Is there a 'natural sequence' in adult second language learning? *Language Learning*, 24(2):235–243.

C. Baker and D. Cokely. 1980. *American Sign Language: A Teacher's Resource Text on Grammar and Culture*. TJ Publishers, Silver Spring, MD.

Roger Brown and Camille Hanlon. 1970. Derivational complexity and order of acquisition in child speech. In John R. Hayes, editor, *Cognition and the Development of Language*, chapter 1, pages 11–54. John Wiley & Sons, Inc., New York.

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

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

Heidi C. Dulay and Marina K. Burt. 1974. Errors and strategies in child second language acquisition. *TESOL Quarterly*, 8(2):129–136, June.

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

Susan Gass. 1979. Language transfer and universal grammatical relations. *Language Learning*, 29(2):327–344.

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

Frank Linton, Brigham Bell, and Charles Bloom. 1996. 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. UM96, User Modeling, Inc.

Kathleen F. McCoy, Christopher A. Pennington, and Linda Z. Suri. 1996. 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. UM96, User Modeling, Inc.

Lisa N. Michaud and Kathleen F. McCoy. 1999. Modeling user language proficiency in a writing tutor for deaf learners of English. In Mari Broman Olsen, editor, *Proceedings of Computer-Mediated Language Assessment and Evaluation in Natural Language Processing, an ACL-IALL Symposium*, pages 47–54, College Park, Maryland, June 22. Association for Computational Linguistics.

Lisa N. Michaud and Kathleen F. McCoy. 2000. 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. FLAIRS.

Lisa N. Michaud, Kathleen F. McCoy, and Christopher A. Pennington. 2000. 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. SIGCAPH.

Lisa N. Michaud, Kathleen F. McCoy, and Litza A. Stark. 2001. Modeling the acquisition of English: an intelligent CALL approach. In *Proceedings of the Eighth International Conference on User Modeling*, Sonthofen, Germany, July 13-17.

Manford Pienemann and Gisela Håkansson. 1999. A unified approach toward the development of Swedish as L2: A processability account. *Studies in Second Language Acquisition*, 21:383–420.

David Schneider and Kathleen F. McCoy. 1998. 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, Université de Montréal, Montréal, Québec, Canada, August 10-14. COLING-ACL, Morgan Kaufmann Publishers.

M. Strong. 1988. A bilingual approach to the education of young deaf children: ASL and English. In M. Strong, editor, *Language Learning and deafness*, pages 113–129. Cambridge University Press, Cambridge.