User Modeling and User-Adapted Interaction (2005) 15: 55–84
DOI 10.1007/s11257-004-6262-0

A Model to Disambiguate Natural Language Parses on the Basis of User Language Proficiency: Design and Evaluation

LISA N. MICHAUD[1], KATHLEEN F. McCOY[2] and RASHIDA Z. DAVIS[2]
[1]*Department of Mathematics and Computer Science, Wheaton College, Norton, MA 02766, USA*
[2]*Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA*

(Received: 31 October 2003; accepted in final form 30 October 2004)

Abstract. This paper discusses the design and evaluation of an implemented user model in ICICLE, an instruction system for users writing in a second language. We show that in the task of disambiguating natural language parses, a blended model combining overlay techniques with user stereotyping representing typical linguistic acquisition sequences captures user individuality while supplementing incomplete information with stereotypic reasoning.

Key words. computer-aided language learning, learner modeling, natural language, parse disambiguation

1. Introduction: The ICICLE System

The name ICICLE represents ‘Interactive Computer Identification and Correction of Language Errors’ and is the name of an intelligent tutoring system currently under development (Michaud et al., 2001; Michaud and McCoy, 2001, 2003, 2004). The system’s primary long-term goal is to employ natural language processing and generation to tutor deaf students on grammatical components of their written English. ICICLE’s interface is similar to that of a text editor, allowing the user to load in text files or to type directly in a large window. We envision it being used to complete general class writing assignments, such as those for which a typical student would use a commercial word processor with a grammar checker (Note: Grammar checkers in commercial word processors are generally inadequate for the task of handling the grammatical productions of non-native English users). ICICLE therefore accepts as its input free-written English texts (not constrained to translation of specific sentences or responses to a question prompt (Note: For examples of the kinds of sentences we have been working with, see the Appendix A at the end of this article) and then responds to the user by highlighting sentences with errors. Our system makes use of a user model to track the user’s level of competence in different English syntactic structures in order to help the system narrow down the most likely parses of an input sentence; the development of this model is discussed in other work. This paper overviews the basics of our model's design and describes how we have evaluated our model with respect to identifying user stereotypes, reflecting user individuality, adapting to changes in user language proficiency, and the effectiveness of such a model as one component in the disambiguation of natural language parses.

1.1. USER MODELING AND THE DISAMBIGUATION TASK

ICICLE uses a CFG grammar with features and a parser which is descended from that described in (Allen, 1995). The grammar consists of 321 rules modeling standard English usage, augmented with 93 error-production rules called *mal-rules*, which were derived from an error taxonomy compiled out of actual writing samples from deaf college students (Suri and McCoy, 1993), enabling the parser to recognize and identify potentially error-containing user-written English sentences. The coverage of our grammar over our target domain of writing is still developing, but was evaluated with respect to agreement errors earlier in (Schneider and McCoy, 1998). In conjunction with our parser, ICICLE uses the Kimmo morphological analyzer (Antworth, 1990) and the COMLEX Syntax 2.2 lexicon (Grishman et al., 1994) to recognize and process basic lexical items.

In the process of seeking a correct analysis of user errors, the ICICLE system needs to choose between multiple parses of each utterance. Some of these parses represent different structural representations of the text, and in the case of ungrammaticality, may place the 'blame' for the error on different constituents. As a step toward determining which is correct, it is necessary for the system to have at its disposal a model of the student's grammatical proficiency which indicates his or her mastery of the language rules involved. This knowledge aids in choosing between structurally differentiated parses by providing information on which grammatical constructs the user can be expected to use correctly or incorrectly.[1]

Other components required for the disambiguation process (e.g., the frequency with which particular rules are used, captured in typical probabilistic treatments) and their relationship to this component are discussed as future work in the Conclusion.

1.2. A MODEL OF GRAMMAR PROFICIENCY

The ICICLE user model, described in depth in (Michaud, 2002; Michaud and McCoy, 2001; Michaud et al., 2001), attempts to capture what we refer to as 'I_i,' or the user's current *Interlanguage* state. The concept of interlanguage is that a language learner is generating utterances from a hypothesized grammar I which approaches the language being learned over time (Selinker, 1972). At the current

[1] This is not to say that the user will not make mistakes in already-mastered material. What we wish to select is the most likely parse given the current mastery of the language.

step in the progression, I_i, certain constructs have been mastered, others are currently being learned, and some are still beyond the user's reach.

One component of the ICICLE user model is MOGUL (Modeling Observed Grammar in the User's Language), which captures what is known about the user's interlanguage grammar I_i through an overlay representation in which individual constructs of morphology and syntax (which we refer to as Knowledge Units, or KUs) are scored according to the system's observations of the user's success in executing those KUs in the writing he or she has previously produced. This model compares the number of times the KU has appeared correctly in the user's productions against the total number of times the KU has been attempted and summarizes this information into one of three tags: *Unacquired*, meaning the user has definitely not mastered the KU, *Acquired*, meaning the user consistently uses it correctly, and *ZPD*,[2] meaning the KU is currently being mastered by the user and is therefore exhibiting great variation in successful execution.

Since a user's performance is expected to change over time, part of the model incorporates the idea of a 'window' of current observations; instead of basing MOGUL tags on all of the user's performance to date, tags are based only on the most recent N writing samples, in order to ensure that past failures (for example) are no longer reflected in the model when the user is now performing at a higher level.

When the system has not yet gathered sufficient data on a specific KU, however, the result is incomplete knowledge in MOGUL. These 'gaps' in the profile of the user are filled using the information provided by the second component, Steps of Language Acquisition in a Layered Organization Model (SLALOM).

In a way similar to the double stereotype system in Knome (the User Modeling component in Chin's Unix Consultant (Chin, 1986, 1989)), grammatical information in SLALOM is grouped together into three levels: Easy (those grammatical constructs that are acquired first by second language learners), Medium (those constructs generally acquired only after the Easy constructs have been acquired), and Hard (those constructs that are generally acquired last). Thus, SLALOM organizes grammatical constructs into three layers where constructs within a layer (e.g., Easy, Medium, Hard) are generally acquired at the same time, while the Easy layer is generally acquired before the Medium layer, which is generally acquired before the Hard layer. KUs are also grouped into hierarchies of related structures such as those which realize NPs, those which realize VPs, and those which are Relative Clauses (indicated in Figure 1 by vertical groupings). Structures at a given layer, regardless of hierarchy, are typically learned before structures at a layer 'above' that layer. For more detailed information on the structure of the model, please see (Michaud and McCoy, 2004). Using this SLALOM model, we can define three different stereotypes of users as depicted in Figure 1: Low, Middle, and High. Within each of the stereotypes, the boxes are meant to represent groups of KUs.

[2] *Zone of Proximal Development* – see (Vygotsky, 1986) for a discussion.

'Easy' KUs are at the bottom of each figure, 'Medium' are in the middle, and 'Hard' are at the top. Horizontal lines connect KUs that are acquired at roughly the same time (i.e., occur in the same layer).[3] The three User Stereotypes can be described as follows:

Low: In a Low-level stereotype our expectation is that this learner is in the process of acquiring the Easy KUs and therefore the KUs are marked ZPD. All other KUs are marked Unacquired.

Middle: In the Middle-level stereotype the Easy KUs are now assumed to be Acquired, while the Medium KUs are now marked ZPD (and the Hard KUs remain Unacquired).

High: In the High-level stereotype all Easy and Medium KUs are now assumed Acquired while the Hard KUs have not moved into the ZPD.

The import of this representation is that user performance is directly recorded in the MOGUL portion of the model. At a given point in time, a user will also be classified as Low, Middle, or High depending on which of the SLALOM stereotypes the user's recorded performance most closely resembles. A 'blank' KU in MOGUL can then have its tag be inferred on the basis of these stereotypes; for example, if a user is in the Middle stereotype grouping, then a blank KU in MOGUL will be compared against the stereotype image for Middle and receive the tag which a Middle user typically has on that KU. That stereotype therefore provides probable tags for the KUs which have not yet been observed in sufficient quantity in the user's performance.

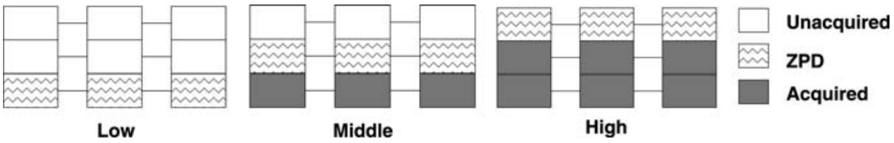

Figure 1. The progression of stereotypes in SLALOM.

1.3. IMPLEMENTING THE MODEL

This section describes how we have implemented the described user model within the current ICICLE instructional system.

1.3.1. *Determining the KU Placement: Easy, Medium, or Hard*

We are currently undertaking our own exploration of the typical linguistic structure acquisition order for our user population to determine placement of KUs

[3] Although not depicted in Figure 1, different numbers of KUs fall into different layers. In addition, a single KU may participate in multiple layers reflecting the fact that some structures may take longer than one stage to master.

into Easy, Medium and Hard; for a detailed discussion of our methodology, see (Michaud and McCoy, 2004). In this earlier work we have concentrated on the errors occurring in writing samples we have collected to represent our user population. We applied clustering algorithms to determine that the users in this population can be classified into three groups based on the errors that they commit in their writing; we have also executed MANOVA tests to see which errors are committed by each of the three groups. These are necessary first steps to establish which KUs go into which layers in the SLALOM model; our continuing efforts focus on acquiring information about structures executed both correctly and incorrectly at the different levels of language acquisition in our collection of writing samples.

While we are working on acquiring this linguistic sequence data, we have created a prototype layer determination for the KUs in the SLALOM structure using alternate data from a study of deaf individuals first conducted on a small scale by (Power and Quigley, 1973) and then extended to an investigation involving approximately 450 deaf students between the ages of 10 and 18; for summaries of this work, see (Quigley et al., 1977) and (Wilbur, 1977). This group of authors used an elicitation-based test battery called the Test of Syntactical Ability, or TSA (Steinkamp and Quigley, 1977), to determine acquisition data over a fairly broad range of syntactic structures. Since this work largely predates the overt acknowledgment of ASL as a language in its own right, no account of learner L1 background was made in these studies. The studies also did not cover all of the syntactic structures which form KUs in our user model. For these and other reasons, it had been deemed unsuitable for our long-term goals as a permanent basis for the acquisitional relationships in SLALOM; however, it provided an initial framework which could be extended ad hoc to cover all of the KUs for this prototype work, and thus it was suitable for evaluating the effectiveness of the MOGUL and SLALOM models for disambiguating parses and following the user as his/her proficiency changes over time. For details on the process by which we converted the Quigley group results for our own purposes, see (Michaud, 2002). The modular nature of the user modeling architecture will make it very straightforward to substitute our own results when they become available.

1.3.2. *A Database Implementation*

The graphical user interface of ICICLE is written in C++, and the Allen-designed parser which is used to process user input is implemented in LISP. The user modeling component of ICICLE has been implemented in C++ as a mediator standing between these two system layers. The user model itself has been realized as a series of connected MySQL database tables served off of a PC running Windows. It is accessed through wxODBC classes, which allow an interface between C++ object classes and the database through which one can execute database queries.

A portion of one of the database tables storing the MOGUL information is shown in Figure 2. Each entry in this table consists of a student identification number (St_id), then a rule name (rule_id), and finally a number indicating how many ‘hits’ that student has scored on that rule, i.e. how many times that rule has been used in parses of sentences written by that student. Another table associates these rules with the 114 KUs implemented in the system (these concrete rules are clustered into the KUs which represent them as abstractions, grouping together the rules and mal-rules which realize one of the 114 generalized grammatical structures both correctly and incorrectly). The concept behind this table is illustrated in Table I. This table shows two KUs, one which captures rules for NPs (Noun Phrases) with determiners and another for implementing third person present tense verbs. The KU for the NPs has one ‘good’ rule and five mal-rules (indicated by the ‘M’ in their identifier) associated with it, while the verb KU contains one good rule and two mal-rules. Also captured in our database are such elements as student login names and passwords, KU layer information, and canned error explanations for our mal-rules.

```
mysql> select * from student_rule_hits where hits > 0;
+-------+------------+------+
| St_id | rule_id    | hits |
+-------+------------+------+
|     3 | -8>        |   19 |
|     3 | -277>      |   19 |
|     3 | -30_1>     |   19 |
|     2 | -MN200_23> |   32 |
|     2 | -MN09>     |    2 |
|     2 | -MN01>     |    9 |
|     2 | -8>        |   46 |
|     2 | -6_0>      |   57 |
|     2 | -A64>      |    2 |
|     2 | -CN4>      |    1 |
|     2 | -P14_1>    |    1 |
....
```

Figure 2. Some of the entries in one of the MySQL tables representing part of the user model.

Table I. Rules contained in sample KUs

Determiner-NP	3rd person Verb +s
-200_3>	-3>
-MN01>	-MV01>
-MN05>	-MV01_2>
-MN06>	
-MN06_1>	
-MN09>	

1.3.3. *Creating a New User*

Although we do desire to have the MOGUL model reflect only KU tags which are based on observed user performance, every model must have a starting point. Specifically, ICICLE faces the ‘First Parse Problem,’ where the first analysis of a new user’s text must be done with an entirely empty MOGUL model.

We have decided to initialize the representation of a new user by asking the user to place himself or herself into the proficiency level Beginner (Low), Intermediate (Middle), or Advanced (High). This approach is based on similar approaches in other systems, for example the EDGE system (Cawsey, 1990, 1993). In ICICLE, we record the user’s self-selected stereotype level within the database in order to infer KU tags for parsing the first writing sample from this user. Instead of using the stereotype in the usual way – merely filling in tags for *some* blank KUs in MOGUL – the system uses the stereotype to infer all of them. Subsequent to the first sample, performance data can be seen (and recorded into MOGUL) from the structures used in that first sample and once this data reaches a certain threshold[4] the user’s stereotype level will be revised (if necessary).

1.3.4. *From Hits to MOGUL Tags*

Real data is recorded in the MOGUL model at the termination of each error analysis run (which typically consists of one essay put through ICICLE for analysis). In terms of the actual implementation, this process is extremely straightforward; the ‘Student Rule Hits’ table shown in Figure 2 is simply updated with incremented values for the ‘hits’ on all rules which occurred in parses selected by the system to represent the utterances written by that particular student. In our current prototype, the ‘current window of performance’ has not been implemented; all of the rule hits to date are recorded.[5]

This rule hit information is then processed to obtain a ‘hit ratio’ for each of the KUs in the database. The essence of this concept is to capture how frequently the user is able to execute the grammatical structure embodied by the Knowledge Unit correctly. This is calculated by a simple ratio:

$$\frac{\text{successful executions}(KU_k))}{\text{successful executions}(KU_k) + \text{errors}(KU_k)} \tag{1}$$

Executions of the grammar structure, both correct and error-containing, are counted via rule ‘hits.’ The ratio then becomes:

$$\frac{\text{correct rule hits}(KU_k))}{\text{correct rule hits}(KU_k) + \text{mal-rule hits}(KU_k)} \tag{2}$$

[4] The threshold is currently 10 ‘rule hits’ on a KU; this is discussed later.

[5] In future versions of the system, the window will be part of the implementation. However, determining the size of this window is an area of future research.

Table II. Correspondence between user model hit ratios and MOGUL Tags

Hit Ratio	Tag
0–39%	Unacquired
40–69%	ZPD
70–100%	Acquired

These ratios are associated with their respective KUs and stored in the database.[6]

Recall that MOGUL stands for Modeling Observed Grammar in the User's Language. The data on 'rule hits' and 'KU ratios' are the core of MOGUL; although the words for the MOGUL tags introduced earlier – Unacquired, ZPD, and Acquired – do not themselves appear in the database, the ratios are translated into MOGUL tags by the user model interface any time the data are retrieved. The translation is shown in Table II, which illustrates that KUs are marked Unacquired if used consistently incorrectly (correct less than 40% of the time), are marked Acquired if used consistently correctly (70–100% of the time), and are put into the ZPD if they are used with variation (defined as 40–69% correct).

This tag-based representation of the data is consulted for the next step of updating the user information, placing the user into a (possibly new) stereotype level.

1.3.5. *Determining the User Stereotype*

Every time MOGUL is changed (i.e., after each new essay is analyzed in ICICLE), the user's stereotype level must be updated to reflect the new data. We have chosen to place a user in a stereotype according to which stereotype's profile is closest in similarity to the user as represented in MOGUL. To determine the similarity, we have chosen the measure proposed by (Tversky, 1977), who described the *ratio* scale of similarity between objects a and b, defined by feature sets A and B respectively, as:

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)} \tag{3}$$

The function $f(A \cap B)$ represents those 'features' which sets A and B have in common; $f(A - B)$ and $f(B - A)$ represent the distinctiveness of A and B respectively, while α and β represent the weight these distinctions should be given with respect to how much they affect similarity.

[6] In the case of the number of positive and negative hits being so scarce on a KU as to provide insufficient data (our current threshold is 10 hits), a special value is recorded to indicate that there is no solid data on this KU yet.

Tversky goes further to describe the specific situation where a *variant*, or specific instance, is being compared for similarity against a stereotype.[7] He addresses the issue that the variant may have fewer features than the stereotype, which is a more complete image. For this reason, it is far more damaging to the level of similarity for the variant to have distinctive features from the stereotype than for the stereotype to have features which are simply not included in the variant's incomplete set. Therefore, for a variant/stereotype comparison where a is the variant and b is the stereotype, $\alpha > \beta$ in Equation 3.

This variant/stereotype description is a very appropriate match to our own similarity problem. We take object a to be the variant (our data on a particular user), while object b is the stereotype to which a is being compared. The feature set A is the set of MOGUL tags (essentially $[KU, tag]$ pairs) that we have recorded for the user a, while the feature set B is the set of tags associated with the stereotype b. Obviously, B is complete; in our stereotype image, we associate a tag with every KU. A, on the other hand, is incomplete; there will be some KUs for which we do not have data on this user. Since we do not want the incompleteness of our knowledge of this user a to affect the similarity of a to the stereotype b in any way, we set the weight β for the function $f(B-A)$ to 0. We set α to 1. We then explicitly define the function $f()$ in the following way:

$f(A \cap B)$ How many features a and b have in common, or the numerical count of how many KUs are tagged for user a identically to the stereotype b.

$f(A - B)$ How many features a has that are distinct from those of b, or the count of how many KUs have been calculated by user performance to be one tag, but tagged in the stereotype to be another.

We note that with $\beta = 0$ and $\alpha = 1$, Equation 3 becomes:

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + f(A - B)} \tag{4}$$

With the definition of $f()$ above, this is equivalent to

$$S(a,b) = \frac{\text{number of tags in } A \text{ correct for stereotype}}{\text{number of tags correct in } A + \text{number of tags incorrect in } A} \tag{5}$$

Since the denominator is clearly the total number of tags user a has marked, this reduces further to

$$S(a,b) = \frac{\text{number of tags in } A \text{ correct for stereotype}}{\text{total number of tags in } A} \tag{6}$$

[7] Tversky refers to the concept as a prototype. We will maintain the 'stereotype' term in this discussion so as to avoid confusion with the SLALOM prototype implementation.

This is equal to the percentage of MOGUL tags recorded for user *a* which are correct for the stereotype to which *a* is being compared, a very intuitive measure of *a*'s similarity to that stereotype. Therefore, whenever the stereotype for the user needs to be updated, the user is compared against each stereotype and the stereotype with whom the user has the greatest number of tags in common is selected and recorded as the user's new stereotype.[8]

Note that when the stereotype is updated (after each new analysis of a user's writing), there is no bias to the update decision which prefers that the stereotype be revised up, down, or at all. The user modeling component merely takes the measure of how the collection of KU tags representing this user measures against the Low, Middle, or High stereotype. Since the user's KU tags may have changed, one of these stereotypes may now 'fit' the user better than the one to which he or she is already assigned, or the stereotype may remain the same. Since we make no assumptions about the direction in which the user is developing, if at all, we allow for the stereotype assignment not only to transition with the user but also to correct for initial inaccuracies if needed (either overestimating or underestimating).

Note also that this approach to stereotype placement does not take into account unacquired KUs which are not marked for this user because the user has never attempted to execute the structure associated with that KU. A KU will only exist in set *A* with the Unacquired tag if the user attempts the structure and makes consistent errors in those attempts. The reason we cannot account for avoided structures is because it is impossible to distinguish (prior to stereotype placement) between structures which have been avoided because they are unacquired and those which the user has not had opportunity or reason to use. We must proceed using only that data which we have been able to derive from actual performance. Subsequent to stereotype placement, however, ICICLE will be able to infer Unacquired tags on some structures using the SLALOM architecture, as those KU which SLALOM indicates as beyond the user's current ZPD are most likely unacquired. This will be discussed more in the next section.

1.4. USING THE MODEL TO 'SCORE' A PARSE

This two-component model helps ICICLE to sift through the multiple syntactic analyses provided by its parser by indicating which are more likely given the success (or failure) of the attempted syntactic structures. The algorithm to accomplish this task was implemented with the following steps:

[8] At this time, no distinction is made if a user is only marginally a better fit for one stereotype as opposed to another. In future versions of the system, we may use this in conjunction with the *confidence* scores discussed later in Section 1.4. If the user does not closely match only one stereotype, and the multiple stereotypes in which the user could potentially be placed disagree on the tags for a given KU, that could weaken the strength of confidence the system has on the tag the chosen stereotype assigns.

(1) Obtain all possible parses for the input sentence. This involves processing in parallel all possible interpretations of each lexical item, including different possible syntactic categories and morphological analyses.
(2) Score each parse tree according to how likely it is given the user's current interlanguage state I_i (as captured in the user model). This scoring process is described below.
(3) Select a parse tree with maximal score, i.e. one made from rules most likely in the current interlanguage.

Determining a parse tree's compatibility to I_i is done as a two-step process. First, the tree is traversed so that a score for each node (which represents a rule used in the parse) is determined in the following manner:

(1) Determine the parsing rule used to construct the constituent represented by this node and the KU to which this rule belongs.
(2) Determine the tag on this KU. This will be Unacquired, ZPD, or Acquired. If there is insufficient data in MOGUL to supply this tag, the tag is inferred using the SLALOM information on typical performance for the user's stereotype level.
(3) Translate this marking into a score for this rule, giving high scores to those rules which should be in I_i given the tag on the KU, and low scores to those rules which are not expected to be in I_i.

The process of obtaining the score in Step #3 reflects an answer to the question: *Do we believe that this rule is in I_i?* If the answer is yes, the node receives a positive score of 1. If the answer is no, the node receives a negative score of -1. 'Unacquired' KUs imply that rules representing correct execution of the structure are not in I_i, but rules representing malformations of the structure are. Conversely, 'Acquired' KUs are represented by correct (regular) rules in I_i, not mal-rules. KUs in the ZPD represent structures realized by competing rules, both correct and incorrect, which result in the variation in ZPD-level performance; for that reason, both mal-rules and correct rules are believed to co-exist in I_i for those structures. Table III reflects the relationship between tag, the nature of the rule, and the score assigned.

Table III. Calculating node scores

Tag	Rule type	In I_i?	Score
Unacquired	Regular	no	-1
	Mal-Rule	YES	1
ZPD	Regular	YES	1
	Mal-Rule	YES	1
Acquired	Regular	YES	1
	Mal-Rule	no	-1

One of the benefits of scoring from the endpoints of a $[-1, 1]$ interval is the notion of 1 representing a 'strong positive' and -1 a 'strong negative.' The value 0, being in between, would reflect an inability to state a belief about the rule's status. Although 0 is never assigned to a rule node,[9] it is clear that future versions of the scoring mechanism may benefit from using the strength of evidence on a tag to produce a measurement of the *confidence* in the score. If this confidence measure were expressed in terms of a percentage, e.g. 90%, it could be applied against either 'yes' or 'no' scores by multiplication to affect the strength of the score. Scores would then range over the $[-1, 1]$ interval, with low confidence levels bringing the score closer to 0, the neutral statement.

Once all of the node scores for a tree are determined, these scores are combined to obtain an average score to represent the likelihood of the entire tree overall.

2. Evaluating the Model

This parse scoring mechanism and the user model on which it is based have been implemented within the ICICLE system. In order to demonstrate the efficacy of the implementation, we set out to show the following:

- Parse selection based on a stereotype successfully identifies parses which are the closest match to the 'expected performance' depicted in the stereotype image in SLALOM. That is, different parse trees are given maximal score when different stereotypes are used.
- When a user builds up a history of performance that deviates significantly from the assigned stereotype – for instance, when the student's proficiency changes because he or she is learning – the stereotype assignment is updated to better reflect the user.
- When a user is correctly placed in a stereotype and yet has individual deviations in his or her MOGUL tags from that stereotype,[10] representing a history of 'atypical' performance, the parse selector correctly scoring mechanism correctly recognizes the appropriateness of parse interpretations that are consistent with that user's individuality.

For this evaluation, we used a corpus of sentences contained in 106 samples of writing by deaf individuals at various levels of English proficiency.

2.1. PARSE SELECTION DEPENDING UPON THE STEREOTYPE

The parse selection process as it operates when all decisions are based on a selected stereotype level is consistent with the mode of operation with a new user,

[9] Except in a small number of situations where a rule in the parsing grammar does not represent a grammatical structure, as is the case with some 'fix-it' rules which merely insert certain key features. These nodes are not included in the tree average.

[10] Individual deviation from a stereotypic learner is common in all domains, including second language acquisition; a specific type of deviation which we used in our evaluation is discussed later in this paper.

and also reflects the system's ability to select parses which are consistent with a complete performance profile. To illustrate this process, we selected a stereotype level of 'Middle' for a hypothetical user, and we parsed the following sentence from our corpus:[11]

(1) I really like wrestling.

The parser found six possible trees to span this input. Several of these parses received low scores; they all involved a syntactic interpretation containing a dropped copula verb.[12] This interpretation would be consistent with reading the sentence as 'I *am* really like wrestling,' whose parse is similar to the standard parse for 'I am really like my mother.' The parse involved dropping the copula verb *be*, an error common for some learners in this population but inconsistent with Middle-level performance. The mal-rule that handles dropped copulae (−`MV22>`) participates in a Knowledge Unit which, according to the SLALOM model, is in the ZPD for a Low-level learner, but is Acquired at the Middle level. Therefore, the parses involving the dropped copula were themselves 'dropped' for involving a mal-rule reflecting an error we would not expect from this learner. The parses receiving high scores from the parse selection mechanism were far more consistent with the Middle-level performance profile represented by that stereotype layer in SLALOM. These are shown in Figure 3. Either of these two parses would be acceptable to a human judge. Almost every rule used in these two parses is consistent with Middle-level performance; only the use of a gerundive verb as an NP or as a VP complement is above the expected performance of this user, resulting in a negative score on those nodes. The topmost tree in the figure obtained a slightly higher score. Even though each tree had one node that received a score of −1, the topmost tree had more nodes that were rated +1 and thus it had a higher overall score. We believe that this does not reflect negatively on the system's selection process. A parse tree containing more nodes typically reflects a more complex structure; however, if all of these additional nodes score positively, it means that this more complex structure is entirely consistent with what can be expected of our user. If it were not, some node(s) would receive negative scores, and the larger parse tree has no scoring advantage. Therefore, if a slightly larger/more complex tree ends up being preferred, and we are choosing between two structures which are both within the user's mastery, we are simply giving some small preference to the more complex of otherwise equally valid analyses.

The simple example presented above shows us that the mechanism for scoring a parse based on a stereotype works in a manner which is consistent with its design; it rewards nodes which involve rules we expect to be in I_i for this learner, while penalizing those it does not believe to be in I_i. In order to contrast the

[11] This sentence was chosen from a sample graded by an expert as belonging to a Middle stereotype level. It was chosen for this example because of its relatively small number of parses and its easily understood parse trees.

[12] Although *wrestling* is a verb form, the gerund is used as an NP in this interpretation and therefore does not function as a verb in this sentence.

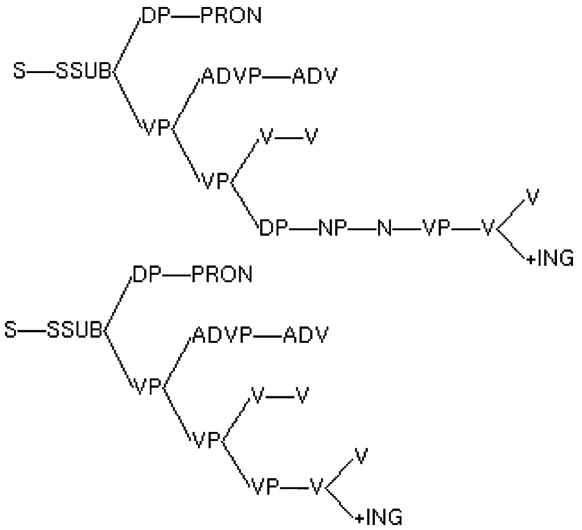

Figure 3. Acceptable parses of ‘I really like wrestling’ for a Middle-level user.

Table IV. Parse tree scores for ‘I really like wrestling,’ all stereotypes

Tree	Low	Middle	High	Notes
0	0.75	0.75	**1.0**	GOOD
1	**1.0**	0.5	0.5	Dropped copula and determiner
2	0.75	0.5	0.75	Dropped copula
3	0.78	0.56	0.778	Dropped copula
4	0.78	0.56	0.778	Dropped copula
5	0.82	**0.82**	**1.0**	GOOD

mechanism’s good performance on this example using the ‘correct’ stereotype for the level of the sentence, we also ran the same sentence with a Low stereotype and with a High stereotype. The results of these analyses were quite different and the resulting scores are shown in Table IV.

The expectations generated by the Low stereotype did not penalize the dropped copulae and, in fact, rewarded one parse (#1) which involved both the dropped copula and a dropped determiner. In this parse, *like* was treated as a noun without its required determiner.[13] The expectations of the High stereotype paralleled those of the Middle stereotype, except in that the gerundive use of ‘wrestling’ as a noun was considered more likely at this level, raising almost all of the scores. This also resulted in more than one parse receiving a maximal score under the High stereotype. As discussed in the Conclusion, the acquisition status of a KU is important, but it is only one aspect necessary for parse disambiguation. Later in Section 4, we

[13] In this instance, ‘[a] like wrestling’ was parsed as would be ‘a horse jumping.’

discuss how in our future work we intend to investigate the incorporation of frequency data in a probabilistic paradigm; this will help make interpretations such as '[a] like wrestling' to appear less likely for any user.

In sum, however, we see the reflections of the different stereotype expectations working as we had hoped. Given a Low level stereotype, 'I like wrestling' is assumed to contain errors because a Low level user is unlikely to have acquired this use of gerunds. At the higher levels, the gerund is more likely (and the dropped copula is less likely), resulting in an error-free interpretation.

To further illustrate the success of stereotype-based parse selection, we ran 14 additional sentences[14] under each of the three stereotypes. In no case did the three stereotypes give identical scores to the same parses; in a majority (11/14, or 76%), the Medium and High stereotypes scored the selected parse equally, and in one case Low and Medium scored a selected parse the same (reflecting the fact that many structures take more than one 'step' of the user's Interlanguage development to be acquired, and thus are not differentiated across layers or stereotype levels), but the Low stereotype never agreed with the High stereotype. The sentences used in this evaluation can be seen in the Appendix A at the end of this article.

2.2. UPDATING THE STEREOTYPE ASSIGNMENT

What if the stereotype the system has recorded for a user is wrong? Recall that we expect our user's language proficiency to be dynamic as learning progresses, and that eventually the stereotype recorded for a user at step i in his or her language acquisition will no longer be appropriate when the user is at some later step $> i$. In our next task, we sought to illustrate how the system may recognize the inappropriateness of a stereotype for a given learner and update that stereotype assignment over time.[15]

We chose to create a new user for this task, again with the stereotype level Middle. We wished to design a situation in which our learner was *previously* a Middle-level English user, but has now progressed to more advanced proficiency. In this situation, we wanted to update the stereotype selection to High. We selected a batch of 20 sentences from samples in our corpus. Fifteen of these came from samples which had been scored by expert judges[16] as representing a High proficiency level, and 5 came from samples which received a Low or Middle rating, but which contained some High-level syntactic structures. Our objective was to assemble

[14] The 14 sentences were randomly selected from those sentences in our corpus that can successfully parse under the current grammar. Of these sentences, 10 contained no grammatical error, while 4 contained errors.

[15] Recall that there is no bias toward upward revision of the stereotype. Although we chose to illustrate a learner's upward transition in this example, the ability of the user modeling component to adjust the stereotype is the same whether it is being revised higher or lower.

[16] Our judges were a panel of four instructors trained in the evaluation of writing by second language learners. The judging is discussed in (Michaud and McCoy, 2004).

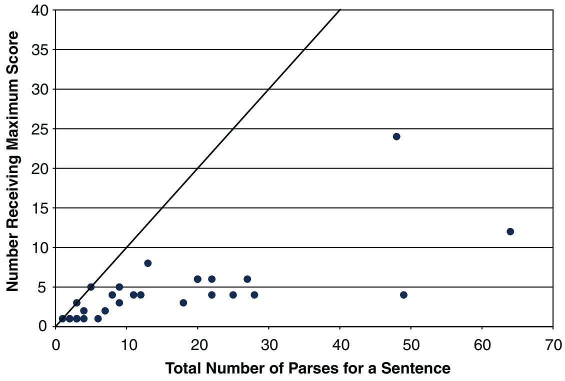

Figure 4. How many parses receive the same maximum score. The diagonal line indicates the point at which all of the parses for a sentence score equally.

sentences that clearly exhibit structures expected primarily of a High-level learner, the new stage to which our user had progressed.[17]

In our first step, we fed these 20 sentences into the ICICLE analyzer (with the SLALOM-based stereotype set to Middle) together in a batch as if they were a user-written essay. This way, the parser would analyze them each in turn, but would not send data back to the user model to update it until the analysis of all 20 had finished. We sought to see how successful the suggestions of the Middle-level stereotype would be for these High-level sentences.

We compared ICICLE's parse selections against the 'optimal' choices of a human judge.[18] The system gave the maximum score to an optimal parse 12 of 20 times (60%), despite having the 'wrong' stereotype (Middle rather than High) on which to base these decisions.

We report the 'maximum score' rate here because there are cases in which the system assigns equal scores to several parse trees. Figure 4 illustrates the relationship between the total number of parses for a sentence (x axis) *versus* the number of those parses all receiving the maximum score (y axis) for 28 of the sentences from the first and second evaluation sets.[19] It should be noted that in almost

[17] These sentences were extracted from the corpus through a search for specific types of error and specific levels of competence, and were screened to ensure that ICICLE was capable of parsing them with an appropriate interpretation among those parses obtained. They can be found in the Appendix A at the end of this article.

[18] In some cases, more than one parse is considered optimal because of inconsequential syntactic differences which allow multiple parses to be acceptable. In all cases, however, the judge's choice was non-controversial; we would expect any native speaker judge to have selected the same choice.

[19] This includes the 15 sentences from the first set (using data from the analysis as a 'Middle' user), plus 13 of the 20 sentences from the second set. Of the original 20, we can only currently collect data on 14 because of technical difficulties with the ICICLE front-end application. One of those 14 was left out of the figure because its number of parses – 256 – reflects a bug in our grammar and makes it infeasible to collect data on how many receive the maximum score.

all cases, the number of maximally scoring parse trees is small. We are currently working on techniques to further reduce these numbers, as we discuss later in this paper. In the meantime, we consider situations where the maximal score has been given to the optimal parse to be those in which the system has recognized the validity of that parse.

After looking at the success rate for selecting the optimal parse, we then inspected the resultant MOGUL markings to see if the system had learned something about the user in just this one sample of 20 sentences. In fact, eight Knowledge Units had now been observed sufficiently so as to now bear MOGUL tags,[20] all marking Acquired structures. Furthermore, those eight were now a closer match to a High stereotype than to a Middle one; the stereotype level of the user was therefore no longer set to Middle, but had been changed to High.

To further test the system's ability to adapt to a learning user over time, we next started with another new user (again set to 'Middle') and iteratively entered these sentences one at a time, requesting an analysis (and subsequent user model update) after each sentence. The question we sought to answer here was whether the cumulative evidence provided by the earlier sentences would enable the system to make more appropriate decisions about the sentences at the end than it had in the batch run, and if the total number of correct choices would be higher, given the access to the evidence provided in the first several sentences. Despite the fact that 20 sentences provide relatively little evidence, particularly when the number of 'hits' on a KU must exceed our Sufficient Data threshold before it receives a mark in MOGUL, the number of maximum-scoring optimal parses rose to 15/20 (75%). These results are compared with those of the batch run on the left in Figure 5.[21]

We also examined the results obtained for the last five sentences in particular. In the original batch run, only in one of these five had the optimal parse been maximally scored by the system. In the results of this iterative run, however, we observed the following:

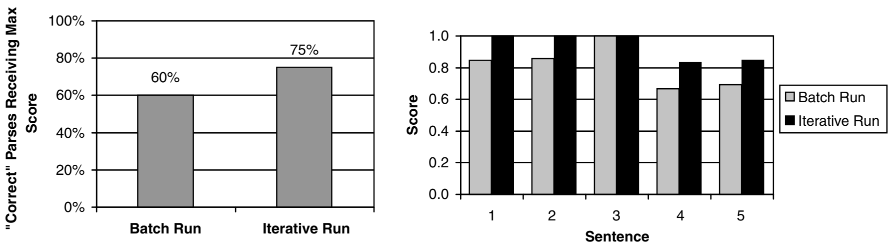

Figure 5. Comparing the batch and iterative runs.

[20] Recall that the Sufficient Data threshold is currently set to 10 rule hits.

[21] Note again that as illustrated in Figure 3, the correct parse may be one of several receiving the maximum score.

- Before reaching the final five sentences, the stereotype had already been revised to High, so that the decisions the system made on these were already affected significantly by the input from the earlier sentences.
- For the four sentences for which the optimal parse tree's score was originally lower than the maximum, the score increased.
- The number of maximum-scoring optimal parses rose from 1/5 to 3/5.

The differences in the two sets of scores are shown for all five sentences on the right in Figure 5.

2.3. OVERRIDING THE STEREOTYPE

The previous two sections have illustrated how ICICLE will prefer a parse that is consistent with stereotype expectations in the absence of other information, how the system can recognize when a user deviates significantly from that stereotype (and can update the stereotype categorization accordingly), and how it can adjust its parse selection decisions on the basis of collected evidence. In a final evaluation task, we explored the integration of the two facets of the user model; specifically, we tested ICICLE's ability to integrate the stereotype expectations which SLALOM provides with the specific history of parses that it records in the MOGUL model.

For this task, we chose to create a High-level user who is appropriately classified in that stereotype but in whose language mastery there exist certain fossilized structures that are executed with error in a fashion 'atypical' of the High-level stereotype.[22] We sought to illustrate that if the system has built up a performance history for this user illustrating these differences, it will select parses more appropriate to the user's actual interlanguage I_i even if that deviates from stereotype expectations.

We chose from the corpus 35 example sentences to represent our learner with fossilized errors. The majority of these sentences exhibit Middle- and High-level structures executed correctly (as would be expected with the High-level stereotype), while half of them (17/35) contained Low-level errors, focusing on errors in noun pluralization, third person singular verbs, and determiner usage.[23] The sentences we used are included in the Appendix A.

As shown in the previous section, the system is capable of adjusting to an individual user and the parse selection process will reflect that. To simulate an accumulated performance history, we trained the MOGUL model by withholding one-fifth of the sentences (randomly selected) and allowing the system to parse the remaining four-fifths and record the rule hits from these parses. The purpose of this was to build a profile of our user reflecting his unique linguistic performance. We then

[22] Fossilization is actually a common occurrence in second language learning, reflecting some low-level linguistic elements which remain unacquired even after the learner has significantly progressed.
[23] Many of these combined High-level structures with these errors, illustrating how common this type of user really is.

inflated the rule hits by a factor of 10 to ensure that the user's simulated past performance exceeded our threshold for Sufficient Data on rule execution.

The remaining sentences were then fed to ICICLE in order to see if the system could accurately address them. ICICLE was, in fact, able to choose the correct parse for every one of the test sentences for which there was a result.[24] On inspection of the user model, one can see why this is the case: unlike in the MOGUL tags of a typical High-level learner, among the 66 KUs for which this user now had tags, infinitive verbs were marked Unacquired, while determiner usage hovered on the edge of being downgraded from Acquired to ZPD, and third person singular verbs were firmly marked as being in this user's ZPD. The majority of the errors that occurred in the randomly selected seven sentences of the test set reflected this hypothetical user's difficulties with third person singular verbs, and the user model's reflection of this problem as perceived in the past history had led the system to correctly recognize this error in the user's current production.

What we have shown is that ICICLE's capacity for accurate analysis reflects the performance history it has garnered from an individual user. This allows the system to recognize the atypical structures that a writer at a particular stereotype level may exhibit; a history of atypical constituents better enables the system to correctly parse sentences containing similar atypical constituents. The user modeling component then has the ability to recognize the validity of parses whose rule usage is consistent with what it knows about the *individual*, not just expectations based on a population, using an integration of the SLALOM stereotype information and the MOGUL individual performance data.

3. Comparison to Related Systems

Explicitly modeled errors in a parsing grammar have been used in other systems, such as the French Grammar Analyser (Barchan et al., 1986), which used its grammar to parse free-form text in French. However, the creators of FGA point out the complexity of maintaining a grammar and lexicon large enough for a broad range of parsing tasks, and note that their system was designed for application only in restricted translation tasks, where grammar and lexicon could be held to a manageable subset.

The difficulty of designing a grammar for a large and broadly defined domain is precisely why ICICLE requires a user model. While grammars can be expanded to handle increasingly broad ranges of syntactic structures, both correct and containing errors, this inevitably introduces into the parsing process a large degree of uncertainty with regard to which interpretation of the user's utterance should be selected. Our belief is that taking the individual and their level of acquisition into account is an important component. This section briefly discusses some of the

[24] Due to a system malfunction, one of the seven sentences did not result in any information about which parse was selected.

systems which are comparable to ICICLE and how they have addressed these and similar challenges.

3.1. BELLOC

The designers of BELLOC (Chanier et al., 1992) share with us the approach of viewing the utterances produced by users as representative of sets of internal rules modeling language. In their application for English-speaking learners of French, users interact in written French within the domain of negotiating an inheritance. When a user produces an error, BELLOC seeks to determine the underlying rule (what we would refer to as a 'mal-rule,' and Chanier et al. refer to as an 'applicable rule' or AR) in order to provide constructive and helpful feedback to the learner.

Like in ICICLE, their grammar is augmented by a set of typical 'learner's rules,' or error-containing rules from their population. They do not, however, take note of which of these rules (if any) a given user can be expected to use. If correct rules fail to parse a sentence, a theorem-prover attempts to select from the set of error-containing rules to achieve a parse. For each sentence it cannot parse with correct rules, the system narrows the candidate applicable rules down to a specific set which could explain the resultant utterance, and then explicitly queries the user on judgments of certain error-containing phrases to determine which one the user believes is acceptable French, using this selection to diagnose which underlying rule is causing the current problem. This extra negotiation with the user would clearly be overly intrusive if the system were interpreting an extended piece of text, particularly if the user produced several errors per utterance.

3.2. VERB GENERATION EXPERT

The Verb Generation Expert was implemented by (Tasso et al.) as part of the English Tutor (ET) system for teaching English verb tense choices to Italian-speaking students. Their system combines a collection of 'bug rules' (mal-rules) and bug-construction techniques which allow them to make certain systematic perturbations to correct rules at run-time. The resultant expanded set of error-containing rules is applied to student input in order to diagnose why a student committed an error.

A user model is maintained in this system, and student performance is marked both in terms of correct grammatical knowledge applied and explicit errors committed. Like our model, theirs waits for a certain threshold of collected data before observations are committed.

An interesting characteristic of the VGE approach is that when student-based information is incomplete, the system attempts to first use correct, expert-level knowledge to diagnose the sentence. If this fails, it falls back to the standard mal-rules and then the run-time-generated mal-rules. In other words, in the absence of other data, VGE first assumes that the user has acquired a structure, and then applies arbitrary mal-rules to the task. We regard this as problematic, when the evidence presented by the acquisition status of other structures is ignored; SLALOM's stereotypical inferential data prevents ICICLE from having to make such arbitrary choices.

3.3. HYPERTUTOR

The HyperTutor system (Schuster and Burckett-Picker, 1996) is a learning tool for Spanish-speaking English learners of reasonable proficiency. It interacts with the student through a series of translation tasks, presenting Spanish sentences which the student then translates into English. Like ICICLE, the system gives notification on correctness and explanations of the error(s) found, if any.

Although the authors characterize the HyperTutor user model as modeling the learner's interlanguage, their approach is significantly different from ours. The essential nature of their model is a store of the *language learning strategies* the system has observed the student using, the idea being that a learner applies these strategies to build the interlanguage and to provoke transitions. Possible strategies include applying the L1 grammar to the L2 or over-generalizing an L2 construct to a larger set of instances than is appropriate. The HyperTutor consists therefore not of an actual profile of the interlanguage I_i, but rather of what possible reasons may lie behind an incorrect rule existing in I_i. While this supplies useful information to a tutorial component, it would fall far short of enabling the system to handle interpretation tasks outside of the proscribed domain of phrase translation.

3.4. GERMAN TUTOR

German Tutor (Heift and McFetridge, 1999), a CALL system for English-speaking learners of German, accepts single sentences, parses them, and provides the user with feedback on the most salient error found. Its student modeling architecture is very similar to MOGUL; it utilizes a database containing all of the grammatical 'constraints' the parser can recognize as met or broken (analogous to our KUs), each holding a score from 0 to 30 representing the user's knowledge on this constraint. A score from 0 to 9 represents expert knowledge, 10–20 is intermediate, and 21–30 is novice. As the system records user performance over time, these scores are incremented with each observed failure to execute a constraint correctly, and decremented with each success.

The details of this user model, however, are lost in the parse selection process, where the student's proficiency scores on each constraint in the user model are averaged, yielding a general proficiency score for the student. The possible parses are then ordered from simple to difficult, and the student's general proficiency level is used as an index into that list. This technique fails to take advantage of the information about the student's individual strengths and weaknesses that was stored in the model.

German Tutor also fails to take advantage of the possibility of becoming more accurate across multiple interactions with the user; the score for each constraint is initialized to 15 at the beginning of a session with the user, and is not stored from one session to the next. Because of this initialization, there is no distinction between structures which have not yet been observed in this user's performance and structures whose scores are decremented as often as they are incremented (remaining in the same range as their initialization).

3.5. MR. COLLINS

The CALL system Mr. Collins[25] (COLLaboratively constructed, INSpectable student model) (Bull, 1994; Bull et al., 1995a,b) is perhaps ICICLE's closest kin in the field of CALL, although its domain is restricted to the acquisition of Portuguese clitic pronoun usage, and its interaction with the user is largely drill-and-test, missing-constituent (fill-in-the-gap) questions. Mr. Collins' instructional objective is also largely concerned with the learning strategies being employed by the learner.

Relevant to this work, however, is the fact that Mr. Collins models the dynamic user through a *sequence* of student models illustrating the acquisition order in the domain of Portuguese clitic pronoun usage, spanning a spectrum from the novice to the expert (a native Portuguese speaker). This sequence was derived from an empirical analysis of the acquisition process of actual users, in their case a class of 47 Portuguese learners over five weeks of learning. Their results strongly proved an order of acquisition for the structures involved in their domain; however, this domain was constrained to merely clitic pronoun usage, and their examined population constrained to a group of individuals in the same educational environment. For this reason, the consistency of their results is not as surprising.

Like in ICICLE's user model, Mr. Collins combines this acquisition information with direct information about the user. As in VGE, when attempting to parse a sentence, an 'expert model' is consulted first – in order to try to parse the sentence as grammatical – and if this fails, the current model of user's misconceptions is attempted. If this fails as well, the system goes further back in the user's path along the acquisition sequence to attempt a parse with errors from earlier in the history. Subsequent to these attempts, the last place consulted is the realm of grammatical transfers from other languages which the student has learned.

3.6. SUMMARY OF RELATED SYSTEMS

The majority of the systems discussed here have far more narrow goals than ICICLE, primarily addressing very specific aspects of the L2 within strictly defined exercises such as translation. This strongly affects their user modeling requirements and objectives. BELLOC operates within a restricted domain, HyperTutor only parses strict translations of sentences it provides, and VGE and Mr. Collins both instruct only upon specific syntactic categories and not the broader grammar of

[25] 'Mr. Collins' actually refers to the user modeling component of a larger system, but the name is also used for the entire system for simplicity.

the language itself. Of the systems reviewed, only German Tutor accepts free-form text as ICICLE does. It is a very ambitious undertaking and it is clear from the discussions above that ICICLE has some growing yet to do before it can consider this task truly conquered.

Perhaps because ICICLE is such an ambitious project, our user modeling effort is far more precise than those others that have been discussed. BELLOC does not track the user at all; it has to chose between competing interpretations through directly querying the student, a time-consuming and intrusive task when there are many user utterances and/or many errors per utterance. Although HyperTutor shares ICICLE's goal of interpreting user actions through a view of the interlanguage, it is clear that Hypertutor's portrayal of a collection of learner strategies gives a very coarse-grained view of what learner rules may be in that interlanguage. It proposes to offer theories about which types of hypotheses may exist there without giving specific information about which hypotheses *are* there. HyperTutor also does not address the possibility of ambiguous errors (where more than one parse tree, and therefore more than one 'cause' of the error, could account for a user's utterance), which is a key component to the issues we face regularly with the ICICLE system.

German Tutor's user modeling technique is similar to the MOGUL facet of our user model, but in their application they lose the precision represented by tracking user performance on each individual structure when they translate it into a global competence level. This approach completely ignores the usefulness of knowing that an individual may exhibit competence or preference for specific structures while he or she struggles on others, and it relies heavily on the ability to 'rank' potential parses from most complicated to least. It also fails to engage any kind of stereotypic inference structure as is embodied in SLALOM.

Finally, while Mr. Collins' approach is remarkably similar to ours, it captures such a tiny piece of the second language acquisition question – only pronouns, and a subset of pronoun usage at that – that it does not face many of the complexities that we have addressed in our user modeling component for ICICLE.

It is therefore possible to conclude that the ICICLE approach to user modeling through the MOGUL and SLALOM facets, while incorporating some aspects of user modeling and user interpretation which have been seen before, is nonetheless novel both in the precision of concept and application of its user modeling component and in the ambitious scope of the entire story of syntactic acquisition which is embodied within the model.

4. Conclusions

The evaluative runs discussed in this paper illustrate that the ICICLE parse selection mechanism scores and selects trees appropriately given a profile of expected user performance, and that the adaptive nature of the model allows it to shift to adapt to differences in user behavior. They also clearly illuminate paths toward future improvement. Because there were several instances where the system gave the maximum score to many trees, the need for even more intelligent scoring is clear. While parse node scoring on the basis of rule membership in I_i is helpful for the selection of appropriate parse trees, taken alone it does not discriminate strongly enough; in some cases, the number of trees obtaining the highest score is fairly large. In fact, rule membership in I_i is only *part* of what signifies *the most likely tree*; other factors must be taken into account.

One piece of information that is being ignored in the current implement of ICICLE is information of typical frequency of rule use – that which is captured in standard probabilistic approaches to parsing. At first glance one may advocate developing a different probabilistic grammar for each stereotype (Low, Middle, and High) and then selecting the parse with highest probability given the stereotype grammar. There are two issues with that approach that we hope to address in future research:

1. We would need a huge corpus for each stereotype level.
2. How would the dynamic nature of the user changing over time be taken into account?

We feel that our MOGUL and SLALOM models illustrated in this paper will be integral to overcoming both of these issues because they will provide us with the ability to (dynamically) smooth our probabilistic grammar based on both the stereotypical and individual user performance. Consider, for example, how we might proceed in generating a standard probabilistic grammar for a Low-level learner. First, we would induce probabilities in the standard way (Jurafsky and Martin, 2000; Manning and Schultz, 2002) using our relatively small corpus of collected writing samples judged to be at the Low level. Of course, many of the rules in the grammar would be given a probability of 0. A standard approach to dealing with 0 probabilities which are due to small corpus size is to smooth the probabilities by giving a very small probability to all of these ‘unseen’ rules. This would not accomplish our goals, however, because many of these ‘unseen’ rules may be actually very appropriate for a specific user, even if they did not occur in our small corpus.

One way in which our approach overcomes this problem is that our SLALOM model has bundled rules into KUs (and has placed different KUs on different layers). Intuitively, if the training of the grammar has given a high probability to a rule falling into a specific KU (indicating that KU is likely Acquired), then in smoothing we would give all other good English rules associated with that KU a relatively high smoothed probability while giving all mal-rules associated with that KU a very low smoothed probability. Furthermore, in the absence of probability evidence from the corpus in the status of a KU, the smoothed probabilities can be inferred from the SLALOM layers and MOGUL markings in the user model. Given these smoothed values, a standard probabilistic parse could proceed. Working out the details of this concept is a focus of our future research.

While an investigation into a full probabilistic approach is planned, one of the most useful improvements we could make on the parse selection process would be to obtain 'weights' for the syntactic categories of items in our lexicon. Many of the syntactic analyses of user utterances appear *bizarre* to the human eye because of the low likelihood of certain syntactic categories being assigned to a lexical entry. An example would be *black boards* parsing as `V NP` because of an entry in the lexicon stating that *black* could be a transitive verb.[26] Again, this is a situation where in some cases very unlikely entries occur, and yet we do not wish to completely exclude them in the interest of maintaining as broad a coverage as possible. In this case, it would be advantageous to weight the likelihood of specific syntactic categories. Since lexical items are the leaf nodes of the parse tree, they, like the rules that form their ancestor nodes, could have probabilities assigned to them which weight more likely choices with heavier scores.

One possible source for this information is WordNet's Familiarity scores, which return a rating from *very rare* to *familiar* for the assignment of each syntactic category to a given lexical item based on the number of senses (polysemy count) that variant of the lexical item has. This is seen as an approximation for how *common* that variant of the word is. As an approximation, it may still have the strength to give some more discrimination power to our parse selection process.

In conclusion, the evaluation discussed in this paper shows that the design of the ICICLE user model has already found success in melding stereotypical and individual user information, and creating a dynamic form which poses a novel approach to the challenge of ambiguity in the natural language task.

5. Acknowledgments

This work has been supported by NSF Grants #GER-9354869 and #IIS-9978021. We would like to thank the ICICLE group at the University of Delaware, in particular: Christopher Pennington, Ezra Kissel, David Derman, H. Gregory Silber, and Michael Bloodgood for their work on the ICICLE implementation. We would also like to thank the reviewers of this paper and the guest editors of this issue, Sandra Carberry and Ingrid Zukerman, for their helpful comments that led us to greatly improve and strengthen this article.

Appendix A: The Sentences

The following is the complete list of 15 sentences used in the first evaluative task from Section 2.1[27]:

[26] Other examples from these evaluations include *like*, *stays*, and *must* as nouns, *so* as a pronoun, or *game* as an adjective.

[27] Any spelling errors are from the author of the essay being quoted. At this time, our system assumes misspelled words are nouns.

I like wrestling.
Because they all are like me.
We tourist to different place in NTID.
It was about my grandmother who passed away.
I knew what NTID is.
It is RULES!
I hate troublemakers because they disobey the rules at N.T.I.D. and R.I.T.
The football players should gain weight.
I dont like the war.
The cateria is just ok.
He was sorry to steal for a little thing.
I have many friends here.
The football uniform is not heavy.
I have other reasons about it.
He is an Italian.

The following are the 20 sentences which were used in the second evaluative task from Section 2.2 to illustrate the ability of the model to dynamically adjust to a user who is significantly different from the stereotypic assumptions:

There are no trees around this campus.
I do enjoy these things.
She really learned a lot.
They wear something to protect their teeth.
I can share my thoughts about NTID.
There are four black boards in each classroom.
There would also be more fun and memories.
The social life is really great.
NTID offers many different majors.
They also wear pads for shoulders and legs.
I like NTID in many ways.
It seems to be in the perfect place.
She love go to mall.*
I like live in dorm and have lots of new friends.*
Sometimes the game stays tied up.
Both sports have rules to follow.
Candy and Jan did some painting.
NTID seems have no problems around here.*
Everyone must have partner to be successful.*
I like to go to NITD in country.*

*These did not receive High-level scores from our judges, but were chosen because they were compatible with High-level performance.

The following 35 sentences were selected for the third evaluation, discussed in Section 2.3. They were used to represent a High-level learner with certain fossilized structures in his/her acquisition status.

Containing higher-level correct structures:
Learning new experiences is fun.†
I really love the strangers here too.
It was about my grandmother who passed away.
It seems like everybody knows each other.
The feeling was touching.†
They also wear pads for shoulders and legs.
She really learned a lot.
Students are on their own.
There are alot of things I like about NTID.
They should not be abolished at all.
They're there for you when you need someone to talk with.
The football players should gain weight.
It has beautiful landscapes with many trees.
It seems to be in the perfect place.
Candy and Jan did some painting.
All of us had some beer and late lunch.
Gallaudet should turn them away.†
Gallaudet University is great for deaf students.

Sentences containing low-level errors:
They told me that Gallaudet would meet my need.
Greek organization will have publicity.†
So two of us called cab and brought us home.
Aids is deadly disease.
There is no cure for aids and student dies from aids.†
I was little boy.
NTID is really a nice place and good college.
I waited in line for registration paper.
My old friend was interpreter.
I was attracted to the speaker by better interpreter.
The building of NTID and RIT are so beautiful and modern.
NTID should be one of best college.
Gallaudet need more students to enroll.†
The team have about twelve players.
She really like track.
NTID have more activities and sports.†
I am the new person who are in NTID.

†Appeared in the randomly selected test set.

References

Allen, J.: 1995, *Natural Language Understanding*. California: Benjamin/Cummings, Second edition.

Antworth, E. L.: 1990, *PC-KIMMO: A Two-level Processor for Morphological Analysis*, No. 16 in Occasional Publications in Academic Computing. Dallas, TX: Summer Institute of Linguistics.

Barchan, J., Woodmansee, B. and Yazdani, M.: 1986, A PROLOG-based tool for French grammar analysis. *Instructional Science* **15**, 21–48.

Bull, S.: 1994, Student modelling for second language acquisition. *Computers & Education* **23**(1/2), 13–20.

Bull, S., Brna, P. and Pain, H.: 1995a, Extending the scope of the student model. *User Modeling and User-Adapted Interaction* **5**(1), 45–65.

Bull, S., Pain, H. and Brna, P.: 1995b, Mr. Collins: a collaboratively constructed, inspectable student model for intelligent computer assisted language learning. *Instructional Science* **23**(1–3), 65–87.

Cawsey, A.: 1990, Generating explanatory discourse. In: R. Dale, C. Mellish, and M. Zock (eds.): *Current Research in Natural Language Generation*, pp. 75–101.

Cawsey, A.: 1993, *Explanation and Interaction: The Computer Generation of Explanatory Dialogues*. Cambridge, MA: MIT Press.

Chanier, T., Pengelly, M., Twidale, M. and Self, J.: 1992, Conceptual modelling in error analysis in computer-assisted language learning systems. In: M. L. Swartz and M. Yazdani (eds.): *Intelligent Tutoring Systems for Second-Language Learning*, Vol. F80 of *NATO ASI Series*. Berlin, Heidelberg: Springer-Verlag, pp. 125–150.

Chin, D. N.: 1986, User Modeling in UC, the UNIX Consultant. In: *Proceedings of the CHI'86 Conference*. Boston, MA, pp. 13–17.

Chin, D. N.: 1989, KNOME: modeling what the User knows in UC. In: A. Kobsa and W. Wahlster (eds.): *User Models in Dialog Systems*. Berlin, Heidelberg: Springer-verlag.

Grishman, R., Macleod, C. and Meyers A.: 1994, Comlex syntax: Building a Computational lexicon. In: *Proceedings of the 15th International Conference on Computational Linguistics*. Kyoto, Japan.

Heift, T. and McFetridge, P.: 1999, Exploiting the Student Model to Empasize Language Teaching in Natural Language Processing. In: M. B. Olsen (ed.): *Proceedings of Computer-Mediated Language Assessment and Evaluation in Natural Language Processing, an ACL-IALL Symposium*. College Park, MD, pp. 55–61.

Jurafsky, D. and Martin, J. H.: 2000, *Speech and Natural Language Processing*. Upper Saddle River, NJ: Prentice-Hall.

Manning, C. D. and Schultz, H.: 2002, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Michaud, L. N.: 2002, Modeling User Interlanguage in a Second Language Tutoring System for Deaf Users of American Sign Language. Ph.D. thesis, Dept. of Computer and Information Sciences, University of Delaware. Tech. Report #2002-08.

Michaud, L. N. and McCoy, K. F.: 2001, Error Profiling: Toward a Model of English Acquisition for Deaf Learners. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, pp. 386–393.

Michaud, L. N. and McCoy, K. F.: 2003, Evaluating a Model to Disambiguate Natural Language. In: P. Brusilovsky, A. Corbett, and F. de Rosis (eds.): *Proceedings of the 9th International Conference on User Modeling*, Vol. 2702 of *Lecture Notes in Artificial Intelligence*. Johnstown, PA, pp. 96–105.

Michaud, L. N. and McCoy, K. F.: 2004, Empirical Derivation of a Sequence of User Stereotypes. *User Modeling and User-Adapted Interaction* **14**(4), 317–350.

Michaud, L. N., McCoy, K. F. and Stark, L. A.: 2001, Modeling the Acquisition of English: an Intelligent CALL Approach. In: M. Bauer, P. J. Gmytrasiewicz, and J. Vassileva (eds.): *Proceedings of the 8th International Conference on User Modeling*, Vol. 2109 of *Lecture Notes in Artificial Intelligence*. Sonthofen, Germany, pp. 14–23.

Power, D. J. and Quigley, S. P.: 1973, Deaf children's acquisition of the passive voice. *Journal of Speech and Hearing Research* **16**(1), 5–11.

Quigley, S. P., Power, D. J. and Steinkamp, M. W.: 1977, The language structure of deaf children. *The Volta Review* **79**(2), 73–84.

Schneider, D. and McCoy, K. F.: 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*, Vol. 2. Université de Montréal, Montréal, Québec, Canada, pp. 1198–1204.

Schuster, E. and Burckett-Picker, J.: 1996, Interlanguage Errors Becoming the Target Language Through Student Modeling. In: *Proceedings of the Fifth International Conference on User Modeling*. Kailua-Kona, Hawai'i, pp. 99–103.

Selinker, L.: 1972, Interlanguage. *International Review of Applied Linguistics* **10**(3), 209–231.

Steinkamp, M. W. and Quigley, S. P.: 1977, Assessing deaf children's written language. *The Volta Review* **79**(1), 10–18.

Suri, L. Z. and McCoy, K. F.: 1993, A methodology for developing an error taxonomy for a computer assisted language learning tool for second language learners. Technical Report TR-93-16, Department of Computer and Information Sciences, University of Delaware.

Tasso, C., Fum, D. and Giangrandi, P. The use of explanation-based learning for modeling student behavior in foreign language tutoring. In: *Intelligent Tutoring Systems for Foreign Language Learning: The Bridge to International Communication*, pp. 151–170.

Tversky, A.: 1977, Features of Similarity. *Psychological Review* **84**(4), 327–352.

Vygotsky, L. S.: 1986, *Thought and Language*. Cambridge, Massachusetts: The MIT Press. Translation revised and edited by Alex Kozulin; originally published in 1934.

Wilbur, R. B.: 1977, An explanation of deaf children's difficulty with certain syntactic structures of English. *The Volta Review* **79**(2), 85–92.

Author's vitae

Dr. Lisa N. Michaud

Department of Mathematics and Computer Science, Wheaton College, Norton, MA 02766, USA. Lisa N. Michaud is an Assistant Professor of Computer Science at Wheaton College in Norton, Massachusetts. She earned her B.A. in Computer Science and English Literature from Williams College, and she completed an M.S. and a Ph.D. in Computer and Information Sciences at the University of Delaware. Her research interests center on the application of User Modeling to tutoring environments, especially those involving the acquisition of linguistic skills. In addition to continuing her collaboration with the ICICLE project, she is working on implementing user modeling in the King Alfred system, an application used by Michael Drout of Wheaton's English department to teach the syntax of Anglo-Saxon English to undergraduates.

Dr. Kathleen F. McCoy

Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA. Kathleen F. McCoy is a Professor of Computer and Information Sciences and the Director of the Center for Applied Science and Engineering in Rehabilitation at the University of Delaware. She received her B.S. in Computer and Information Sciences from the University of Delaware, and her M.S. and her Ph.D. in Computer and Information Science from the University of Pennsylvania. Her research interests include Natural Language Processing (Computational Linguistics) and its subfields Natural Language Generation and Discourse, User Modeling, and Intelligent Tutoring Systems, with emphasis on applications for people with disabilities.

Rashida Z. Davis

Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA. Rashida Z. Davis is a Ph.D. student in the Department of Computer and Information Sciences at the University of Delaware. She graduated with a B.A. in Mathematics from the University of Rochester and received her M.S. in Computer and Information Sciences at the University of Delaware. Her research interests include Intelligent Tutoring Systems and Human-Computer Interaction, and she is currently working under Dr. McCoy on the ICICLE project.