Investigating the Promise of Learner Corpora: Methodological Issues

Nick Pendar
Iowa State University

Carol Chapelle
Iowa State University, carolc@iastate.edu

Follow this and additional works at: http://lib.dr.iastate.edu/engl_pubs

Part of the Bilingual, Multilingual, and Multicultural Education Commons, Curriculum and Instruction Commons, Educational Methods Commons, and the Other Computer Sciences Commons

The complete bibliographic information for this item can be found at http://lib.dr.iastate.edu/engl_pubs/45. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.
Investigating the Promise of Learner Corpora: Methodological Issues

Abstract
Researchers working with learner corpora promise quantitative results that would be of greater practical value in areas such as CALL than those from small-scale and qualitative studies. However, learner corpus research has not yet had an impact on practices in teaching and assessment. Significant methodological issues need to be examined if results from learner corpus research are going to provide convincing results about language development. This study explored the use of the International Corpus of Learner English (ICLE) focusing on methodological issues such as identification of variation in learners’ levels and statistical analysis of large numbers of predictors consisting of lexical and quantitative text features. Results show promise for the lexical and quantitative variables and machine learning statistical procedures investigated in the study. They also suggest the need for a larger corpus with more systematically sampled subcorpora from across language groups and a clear classification of the texts in terms of levels of L2 development based on objective criteria.

Keywords
Learner Corpora, Automatic Classification, Lexical Analysis of Learner Language

Disciplines
Bilingual, Multilingual, and Multicultural Education | Curriculum and Instruction | Educational Methods | Other Computer Sciences

Comments
**Investigating the Promise of Learner Corpora:**

**Methodological Issues**

**Nick Pendar**

**Carol A. Chapelle**

*Iowa State University*

**ABSTRACT**

Researchers working with learner corpora promise quantitative results that would be of greater practical value in areas such as CALL than those from small-scale and qualitative studies. However, learner corpus research has not yet had an impact on practices in teaching and assessment. Significant methodological issues need to be examined if results from learner corpus research are going to provide convincing results about language development. This study explored the use of the International Corpus of Learner English (ICLE) focusing on methodological issues such as identification of variation in learners’ levels and statistical analysis of large numbers of predictors consisting of lexical and quantitative text features. Results show promise for the lexical and quantitative variables and machine learning statistical procedures investigated in the study. They also suggest the need for a larger corpus with more systematically sampled subcorpora from across language groups and a clear classification of the texts in terms of levels of L2 development based on objective criteria.

**KEYWORDS**

Learner Corpora, Automatic Classification, Lexical Analysis of Learner Language

**INTRODUCTION**

One strand of second language acquisition (SLA) research that is potentially useful for instruction and assessment investigates observable changes in learners’ performance. Results from such research would be extremely useful as a basis for tutorial CALL and formative assessments that could be used in conjunction with classroom practices or by autonomous learners. Skehan (1998) presents a general framework for investigating observable language development through increases in learners’ linguistic fluency, accuracy, and complexity. These dimensions are a useful way of conceptualizing important aspects of language development because they can be studied in observable features of learners’ language performance, such as essays, which can be analyzed by computer. Some research has attempted to identify appropriate features of learners’ language performance that serve as indicators of increasing levels of fluency, accuracy, and complexity.

Such research was the subject of a volume by Wolfe-Quintero, Inagaki, and Kim (1998), who were interested in this kind of research for practical purposes. They reviewed empirical studies that attempted to identify relevant features of written language, including over 100 measures of language development, as indicators of variation in fluency, accuracy, and both grammatical and lexical complexity. They found a large variety of ways of operationalizing these variables, and in fact this variation, along with the variation in the way proficiency levels of the learners were operationalized across studies, made synthesis of results difficult to
interpret. Despite the need for much additional research, Wolfe-Quintero, Inagaki, and Kim argue for the practical value in pursuing this line of research.

Once more is known about the best developmental measures, the potential applications include program placement, test validation, end-of-course assessment, trait analysis of holistic ratings, identification of developmental levels in research studies, and measuring the global effect of instructional treatments. (p. 126)

Ten years later it would be difficult to find second language learning and assessment applications in CALL or elsewhere that are based on this type of research. Instead, except for some research projects, the detailed analysis of learners’ written and spoken language being used in large-scale real applications appears in L1 proficiency testing (e.g., Powers, Burstein, Chodorow, Fowles, & Kukich, 2001). In proficiency testing, where such methods are used operationally for scoring the writing of L1 English writers, success has been achieved largely through training an automatic rating procedure to award the same holistic scores as human scorers do. These methods have been refined to do what they are supposed to, but they have not provided notable insight into substantive features that help to identify different levels of language and that might be useful for assessing L2 performance. Other uses and a variety of approaches to automatic scoring for assessment and instruction are the subject of current research (Shermis & Burstein, 2003), but little if any of this work attempts to provide an understanding of development that we might build upon in second language studies.

In L2 assessment, automatic scoring of second language writing has been exploratory, but Chapelle and Douglas (2006) point out that "a principled approach to scoring constructed linguistic responses must rely on a theory of the construct that the test is intended to measure" (p. 56). For example, features need to be chosen to reflect fluency, accuracy, and complexity because these are theorized dimensions of proficiency. To ultimately assess the extent to which automatic linguistic analysis accomplishes construct-relevant scoring “appropriate evaluation methods—not limited to correlations—need to be employed” (p. 57). In other words, research methods in this area need to extend beyond demonstrating correlations with human raters which appear empirically, but cannot be explained theoretically, and therefore contribute little to our understanding of language development. Knowledge about the development of identifiable linguistic features in writing at stages of progressive complexity, accuracy, and fluency is critical to meeting both of these goals.

Despite the need for results about observable indicators of writing development, many applied linguists appear to view results from SLA research as too tentative to make clear recommendations for practice. As a consequence, learning and assessment materials designed on the basis of research findings about linguistic development are very difficult to find. Researchers who gather learner corpora promise that this situation may change through the use of large collections of learner language data stored electronically, and searched automatically, which will provide solid quantitative results about learner language. As Granger (2002) put it,

one of the reasons why the samples of learner data used in SLA studies have traditionally been rather small is that until quite recently data collection and analysis required tremendous time and effort on the part of the researcher. Now, however, technological progress has made it perfectly possible to collect learner data in large quantities, store it on the computer and analyze it automatically or semi-automatically using currently available linguistic software. (p. 7)
On the surface, it seems clear that a large data set of learner language would be useful to produce solid findings, but it is less clear how such data should be analyzed in a way that draws upon knowledge from SLA and contributes to practices such as assessment and learning. In the volume edited by Granger, Hung and Petch-Tyson (2002), for example, only two of the papers attempt an analysis of learners’ language in order to reveal its characteristics at different levels. Each focuses on a particular aspect of the learners’ language: development of the modal system (Housen, 2002) and small words as indicators of fluency (Hasselgren, 2002). Both of these studies provide results that are relevant to language development through a combination of qualitative and quantitative analysis of learner language collected specifically for these purposes. Can results about language development be found using a larger corpus, automatic analysis, and quantitative methods? In order to tackle this question, we needed a large corpus, procedures for automatic identification of relevant linguistic features, and statistical methods for assessing the significance of feature differences across levels within the corpus. These were the three methodological issues that we explored through the use of an existing corpus of learner language, the International Corpus of Learner English (ICLE).

The commercially available ICLE was developed for researchers to investigate learner language. This corpus consists of a large collection of argumentative essays written by advanced English learners in Europe. This corpus is of particular interest to us because we work in higher education, where variation in advanced ESL learners' writing is critical to assessment and teaching. As in many universities in the United States and Canada, students with high TOEFL scores are admitted to the university. They are then tested to determine which students need additional ESL classes (i.e., identify the relevant variation in their language ability and place them in classes on the basis of this variation). These classes then aim to help students sufficiently develop their writing in order to succeed in their university classes. This scenario takes place across universities throughout the English-speaking world. We therefore need a better understanding of the specific features which reliably signal writing at the low and high ends of the variation that appears in advanced writing. The purpose of this study is to explore the use of the ICLE for potential investigation of variation in advanced learner language. If we can find any quantitative differences among the learners’ language represented in the corpus, the corpus would be worth additional exploration to identify other features.

**RESEARCH QUESTIONS**

Use of learner corpora for identifying reliable indicators of learners’ level entails a number of methodological issues. This study focused on three questions that needed to be answered in order to use the ICLE for investigating learners’ language. First, is the ICLE large enough to find features that reliably predict levels? The question of size for learner corpora is a complex issue. Corpus linguistics does not provide operationally useful guidelines on this question. Instead, the rule in corpus linguistics is “the bigger the better.” However, because a learner corpus is compiled one essay at a time and written by students under particular conditions, a more precise answer is needed. The time and effort that goes into defining a relevant sample and collecting each piece for the corpus dictate that “big enough” needs to be defined. Big enough, of course, depends upon the purpose for which the corpus is used, and we therefore addressed the question of whether or not the ICLE would be big enough for identification of reliable indicators of level.

Second, do automatically identifiable features signal levels? Automatically identifiable features were of particular interest for this study because of the need to conduct an expedient analysis of the corpus at this exploratory stage. In the long term, however, automatic analy-
sis is also of interest. With respect to relevant features, previous research suggests three categories of variables that can be used as independent variables: (a) quantitative measures of mean sentence length, number of sentences, length in words, mean words per sentence, lexical density, and type-token ratio, (b) vocabulary choice in terms of word frequency and simplicity, and (c) syntactic complexity and accuracy. We explored the predictive power of the first two groups of variables for identifying variation in levels. The challenge of working with the ICLE corpus, however, was the identification of levels to serve as a dependent variable. The corpus contained information about length of English study in school, at university, and through exposure in English-speaking environments. It seems clear that the longer the study the higher level the essay should be, but which of these values, or combination of these values, would serve as a valid indicator that could be used as a dependent variable? No evaluation of the individual essays was included in the corpus, and therefore we used data about amount of previous English study that was included with the corpus.

Third, how can statistical analysis be conducted? The basic problem of prediction can be conceptualized statistically as multiple regression analysis: the analysis should reveal significant predictors of variation in a dependent variable within a group of cases. The independent variables (predictors) are linguistic characteristics of the essays such as vocabulary levels (based on frequencies), and the dependent variable is a holistic indicator of proficiency level of the writer. In multiple regression analysis, however, the assumptions that must be made about number of variables and normality of distributions do not hold for these types of data, particularly for an exploratory study in which a large number of predictors are investigated. Use of multiple features for predicting categories is a problem for which analytical tools have been developed by researchers in machine learning. We explored the use of one of these tools, a statistical method called “decision trees.”

METHOD

The methodology for this study drew upon an existing corpus, linguistic features recognized in prior research, and statistical analysis developed for machine learning. We purchased the ICLE corpus from Centre for English Corpus Linguistics at Université Catholique de Louvain in spring 2005. We reviewed a number of studies that had identified features of advanced level ESL writing that varied with level to find some features that could be found in texts through automatic search procedures. We performed a series of decision tree analyses to get an overall error score for the quality of these features for predicting levels. Error is the opposite of reliability, so we calculated reliability as one minus the error score.

Data

At the time of this study, the ICLE contained approximately 2,000,000 words of English written by “advanced” learners of English defined as university students of English in the third and fourth years of their studies. The corpus was comprised of 3,640 text files each of which was accompanied by the following information:

1. type (mostly argumentative and literary essays),
2. length (in words),
3. conditions (timed or not timed),
4. ref tools (whether reference tools were allowed or not),
5. exam (whether the text was written as part of an exam or not),
6. age (writer’s age),
7. sex (writer’s sex),
8. country (country where the text was written),
9. language (writer’s native language),
10. home lang. 1-3 (other languages spoken at home),
11. schooleng (number of years studying English at grade school, 0-12),
12. unieng (number of years studying English at the university, 0-10),
13. monthseng (number of months spent in an English speaking country, 0-192), and
14. otherlang 1-3 (other languages spoken by the writer).

Unfortunately, no rating of the essays was included in the accompanying information. Such a holistic indicator of variation in the essays was critical for the analysis of linguistic features as predictors. We therefore began by having raters attempt to make distinctions among a subset of the essays but found that the interrater agreement was unacceptably low and that there were many essays for which the raters could not decide on a rating. We then returned to the information included with each essay in the corpus to assign a level designation to the writers by separately examining each of the three length-of-study variables (schooleng, unieng, and monthseng).

We settled on a way of combining the school and university length of study data. Even though the corpus was comprised of “advanced” writing, the writers varied in the amount of exposure to and study of English, and the corpus planners saw this variation as best encoded through the length-of-study variables. For example, all other things being equal, a person with 2 years of exposure to English at grade school and 4 years at the university is likely to be less proficient in English than someone with 5 years of exposure to English at grade school and 10 years at the university, even if they are both considered advanced learners of English in terms of their year of university study in English. One would also expect someone who has spent 16 years in an English-speaking country to be more proficient than someone who had not spent any time in an Anglophone environment. Based on this type of reasoning, we combined the three pieces of information about prior English exposure and study according to the following scheme.

We selected essays with monthseng ≤ 0.75 (i.e., 3 weeks or less spent in an English-speaking country) and no other languages spoken by writers (in two of four decision tree models that we built, see the next section). We chose 0.75 months as the cutoff point because that was the median in the whole corpus: half of the writers had spent 3 weeks or less in an English-speaking country and the other half had spent between 3 weeks and 16 years. Because of the unpredictable effect of exposure to English outside the classroom, we excluded the top half. This left us with 292 essays (203,947 words) for writers who spoke no other language than English and their native language, and 1,921 essays (1,338,795 words) for writers who might also know other languages besides their native language and English. Within this group, the length of classroom study variables (schooleng and unieng) were expected to accurately reflect the learners’ levels.

We therefore assigned one of three levels to the remaining essays based on the time spent by the writers studying English. The three levels L (low), M (mid), and H (high) were assigned according to the following scheme:

1. L for schooleng ≤ 7 and unieng ≤ 3
2. H for schooleng > 7 and unieng > 3
3. M otherwise
The numbers 7 and 3 are respectively the medians of *schooleng* and *unieng* in the whole corpus. We used these values for the three levels as the dependent variable in each analysis. We conducted the analyses separately for the group of writers with a second language other than English and those with only English as their second language. The size of each of these two groups at the three levels is shown in Figure 1, and the number of essays and words by group and level are shown in Table 1.

**Figure 1**
Number of Essays per Level per Group

**Table 1**
Number of Words per Level per Group

<table>
<thead>
<tr>
<th>Group</th>
<th>With other 2nd languages</th>
<th>Without other 2nd languages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Essays</td>
<td>No. of words</td>
</tr>
<tr>
<td>Low</td>
<td>755</td>
<td>674,148</td>
</tr>
<tr>
<td>Mid</td>
<td>970</td>
<td>534,964</td>
</tr>
<tr>
<td>High</td>
<td>196</td>
<td>129,683</td>
</tr>
<tr>
<td>Total</td>
<td>1,921</td>
<td>1,338,795</td>
</tr>
</tbody>
</table>
Lexical Features

We focused on the lexical properties of texts that would be indicative of varying levels of proficient writing and that would be easily identified using automatic techniques. The lexical analysis searched for specific words in the essays that should serve to indicate levels. Rather than assume that a single word list would be best, we used three sets of word lists in order to see which lists or sublists would function as good predictors of level.

The three sets of word lists each contain sublists of words that one would expect to find in greater quantities in low-level versus high-level writing.

1. HQ/UWL
   Based on the semantic categories of Hinkel (2003) and Quirk et al. (1985) (HQ) plus University Word List (UWL) levels 1-11 (Xue & Nation, 1984), this list included the following sublists:
   a. Expecting/Tentative Verbs (e.g., attempt, desire, expect)
   b. Private Verbs (e.g., accept, deem, understand)
   c. Public Verbs (e.g., acknowledge, admit, ask, remark)
   d. Semantically Light Items (e.g., be, do, have, it, there)
   e. Vague Nouns (e.g., all, anything, people, stuff, world)
   f. UWL 1-11 (Xue & Nation, 1984)
   g. Off HQ/UWL (items not found in the above subcategories)

2. GSL/AWL
   This list included the General Service List (GSL) (West, 1953) plus the Academic Word List (AWL) levels 1-10 (Coxhead, 2000)
   a. GSL 1000 (first 1,000 words in GSL)
   b. GSL 2000 (second 1,000 words in GSL)
   c. AWL1-10 (Academic Word List levels 1-10)
   d. Off GSL/AWL (items not found in the above subcategories)

3. BNC
   This list included the 14,000 most frequent words in the British National Corpus (BNC) based on a 10,000,000 word sample (BNC 1-14, BNC PN, BNC EX)
   a. BNC1-14 (each level consists of 1,000 word types; levels in the descending order of frequency)
   b. BNC PN (a list of proper nouns based on BNC)
   c. BNC EX (a list of exclamations, disfluencies, etc. based on BNC)
   d. Off BNC (items not found in the above subcategories)

For each essay, we calculated percentages of words in each subcategory of these lists along with type/token ratios and essay lengths. Based on the three word lists, comparisons of the distributions for the high, mid, and low level groups are illustrated in Figures 2-4 for the essays whose writers do not speak any language other than their native language and English.
Figure 2
Lexical Patterns in Groups (HQ/UWL)

Figure 3
Lexical Patterns in Groups (GSL/AWL)
Figures 2–4 do not present coherent lexical patterns in the three levels. Hinkel (2003) found that English learners have in their writing a significantly higher percentage of expecting/tentative verbs, public verbs, semantically light items like it and there (as used in clefts and existentials), and vague nouns than native speakers of English. One would expect the same trend to hold between lower level learners and higher level learners of English. Figure 2 shows that this only holds for public verbs and vague nouns, but the differences among levels are not statistically significant. The error bars on the top and bottom of the means represent 95% confidence intervals, and, if the error bars around two means overlap, the difference between those two means is not statistically significant. On the other hand, if the different levels of the UWL represent levels of difficulty of the words or reflect their frequency of use, then one would expect to see higher percentages of lower level UWL words in the lower groups and higher percentages of higher level UWL words in the higher groups. We see this trend in UWL4 but then the opposite in UWL5 and UWL8. We do however see that the L group has a significantly lower percentage of off-list words than the M group. The wide margin of error for the H group can be attributed to its small sample size (n = 27). The fact that the L group has a higher percentage of off-list words tells us that the word list is small and does not contain many of the words that could potentially help to distinguish the levels among these advanced learners of English or that the writers in the L group have a higher percentage of misspelled words.

The situation is almost the same with GSL/AWL word lists with a slight trend evident in the results from the GLS1000 word list analysis. The L group has a higher percentage of these words than the M and H groups, which is expected, but the reverse is true of GLS2000. As regards AWL, the results are again confusing. Since the AWL levels are intended to correspond to the difficulty of the words in them, lower level writers should have lower percentages of the more difficult words (i.e., the higher levels of AWL), and higher levels of less difficult words (i.e., the lower levels of AWL). This trend appears in AWL1 results; the L group has a
higher percentage of AWL1 words than M and H, respectively, even though the differences are not statistically significant. But the same pattern is also observed in AWL6, which is supposed to contain more difficult words. Again, the off-list words seem to provide a clue. Lower-level writers have a lower percentage of off-list words.

The BNC word lists do not provide a very clear picture either. We see a good pattern in BNC1 and BNC2, the first and second 1,000 most frequent words in a 10,000,000 word sample of the British National Corpus. The lower level writers have systematically higher percentages of BNC1 than the higher level writers, but the reverse situation is true with BNC2. Some promising patterns are also observed in BNC3 and BNC4, but they are contradicted by BNC5 and especially by BNC6. We see other conflicting results in higher BNC groups as well.

These observations may suggest that frequency-based word lists are not good predictors of level after all, at least at higher language competency levels. Also the fact that off-list words show a pattern in HQ/UWL and GSL/AWL suggests that maybe these lists are not comprehensive enough to capture the differences at higher levels.

Figures 5-6 show the means and boxplots of the essays’ type-token ratios and lengths. Figure 5 shows that the L group has a significantly lower type/token ratio than the M group. This is expected, but we do not see this in the H group. This again may be the result of the small sample size. On the other hand, we see that the L group, on average, has a much higher essay length than the M and H groups, which is unexpected. This is only because the distributions of essay lengths are highly left skewed and that is why the mean essay length is not a good indicator of level (see Figure 6).

Figure 5
Type/Token Ratios and Essay Lengths (Means)
In view of the large number of predictors which relate to the levels in various complex ways, an analysis is needed to combine predictors in a way that maximizes prediction based on the information each provides. As mentioned above, in principle, this is the issue that multiple regression analysis is intended to address; however, these data do not meet the assumptions of multiple regression analysis. We, therefore, used a statistical method called decision trees which was developed to test the predictiveness of multiple features.

**Decision Trees**

Decision tree modeling is a widely used technique in machine learning for automatically learning to classify data based on a series of parameters (for details, see chapter 3 of Mitchell, 1997). To run a decision tree model, the researcher divides the data into two sections, called training data and test data. A decision tree algorithm is run on the set of training data, for which a classification is known, to construct a tree containing the features that are useful for prediction. The nodes of this tree represent tests based on the values observed in the training data, and each branch in the tree corresponds to a binary decision about the classification. The paths in the tree lead to the leaves which correspond to any of the known classifications for the data. The training data is in the form of vectors of values. The last item in the vector is the classification of that data item. For example, a decision tree that models a person’s decision to play tennis based on weather conditions would start with observations about that person’s behavior for a few weeks resulting in a set of training data in the following form:

(Outlook = Sunny, Temperature = Hot, Humidity = High, Wind = Strong, PlayTennis = No)
The above vector contains attributes, such as outlook and temperature, and their values, such as sunny or hot. Once we have a reasonable amount of data, we can run a decision tree learning algorithm on the data to model that person’s behavior. Such a decision tree might look like the one presented in Figure 7. If the outlook is sunny and the humidity is high or if the outlook has rain and the wind is strong, then the person will not play tennis.

Figure 7
A Sample Decision Tree for Playing Tennis Based on Weather Conditions (from Mitchell, 1997, p. 53)

Decision tree learning is best suited for problems with the following properties (Mitchell 1997, p. 54):

1. Instances are represented by attribute-value pairs.
2. The target function has discrete output values (although more sophisticated algorithms allow for target functions with real-valued outputs).
3. Disjunctive descriptions may be required.
4. The training data may contain errors.
5. The training data may contain missing attribute values.

Our problem suits this scenario quite well, which makes decision tree learning a good candidate approach, because

1. We can represent each essay as a set of attribute-value pairs. The attributes are the word sublists such as GSL1000 and AWL1 and the values are the percentages of these items observed in the essays.
2. Our target function has discrete output values, that is, L, M, H.
3. Disjunctive descriptions may be required. It may be the case that more than one possible set of values for the attributes are good predictors for one category.
4. The training data may contain errors. There may be errors in the apparently self-reported number of years studying English in the database, which may have resulted in erroneous classification into the L, M or H groups.
5. The training data may contain missing attribute values. In the case of data that includes writers with other second languages than English, there are cases where information about writers’ other languages is unavailable.
The nodes in the tree are selected based on the statistical notion of information gain, which is based on the concept of entropy from information theory (Mitchell, 1997, pp. 55-60). This means that the higher nodes are the most informative ones; that is, they are able to classify the data with more accuracy than the lower nodes. The lower the position of the node in the tree, the less informative that node is. Therefore, once a decision tree is built, we can see which variables play a more important role in classifying our data.

**Modeling the Data**

We used the C4.5 decision tree program (Quinlan, 1993) to model our data. We built several models with different parameter settings. Here we are going to report only four of the models based on the following sets of data:

1. **NOL**: Essays by writers who knew no other languages than their L1 and English (292 essays; 203,947 words)
   - NOL-NL1: Native language not used as an attribute in training the model
   - NOL-WL1: Native language used as an attribute in training the model

2. **WOL**: Essays by writers who may have known other languages as well (1,921 essays; 1,338,795 words)
   - WOL-NL1: Native language not used as an attribute in training the model
   - WOL-WL1: Native language used as an attribute in training the model

Recall that in these models we excluded the essays by writers who had spent over 3 weeks in an English-speaking country.

We built decision trees by using three fourths of the data as the training data and the remaining one fourth as the test data. We were building a model to predict the level of an essay based on its lexical profile (the word lists used) as well as its length and type/token ratio. These added up to 50 attributes to build the model upon. The resulting decision tree would give us an indication of which attributes are the best predictors and how reliable they are. Figures 8 and 9 and Table 2 summarize the results of the decision tree models.
These results tell us that we were able to capture between 94.0% and 97.3% of the variability in the training data using the 50 attributes that we used. The system did not perform nearly as well on the test data indicating that the models were not very reliable, that is, they were specific to the one set of essays rather than generalizable to additional data. Notice that the system’s accuracy improves considerably when we use native language (the WL1 models). This could either be because a person’s native language influences that person’s English use or be the result of overfitting the model to the training data, which is quite likely. Overfitting occurs when a learned model is attuned to the peculiarities of the training data preventing it from making generalizable findings. However, since the models that incorporate the native language into the attributes also perform better on the test data, it makes one wonder how much of this improvement is because of overfitting and how much of it is a genuine pattern. In other words, it may very well be that the use of features of writing to predict level of performance requires information about the writer’s first language.
Best Predictors
Looking at the decision trees, we can now identify what the best predictors of levels are for our sample. C4.5 can report the most useful classification rules it has induced and their reliability. For example, the following rule turns up as the most useful for the WOL-NL1 model with a 90.6% reliability:

\[
\begin{align*}
\text{TypeTokenRatio} & \leq 0.56 \\
\text{BNC3} & > 1.47 \\
\text{BNC5} & > 0.47 \\
\text{BNC10} & \leq 0.36 \\
\text{BNC13} & \leq 0.29 \\
\text{BNC14} & \leq 0 \\
\text{VagueNouns} & \leq 1.6 \\
\text{UWL1} & \leq 1.48 \\
\text{UWL2} & > 0.38 \\
\text{AWL2} & \leq 1.9 \\
\text{AWL3} & > 0.55 \\
\text{OffGSLAWL} & \leq 12.27 \\
\rightarrow \text{class H}
\end{align*}
\]

This rule says that if a text’s type/token ratio is less than or equal to 0.56 and the percentage of BNC3 in that text is greater than 1.47, and so on, then that text has a 90.6% chance of belonging to the H group. When we look at the most useful rules induced by C4.5, we notice that the best predictors of essay level for our sample in rules with over 80% reliability are as follows (in alphabetical order): AWL2, AWL3, BNC2, BNC3, BNC5, BNC6, BNC10, BNC11, BNC13, BNC14, expecting/tentative verbs, GSL 2000, Off GSL/AWL, type/token ratio, UWL1, UWL2, UWL8, UWL10, and vague nouns. Notice that the BNC sublists are the most frequent attributes in this list, and that is because they simply cover a wider range of lexical items than the other word lists. This and the fact that Off GSL/AWL also shows up as a reliable predictor suggest that more comprehensive word lists or simply taking into account all words in the corpus will result in better effectiveness for our classifier and in turn the recognition of more useful and reliable patterns.

CONCLUSION
The analysis in this exploratory study incorporated a large number of variables and complex decisions about partitioning of the data. As a consequence, much could be written about various alternatives and how they affected results. Rather than attempting to interpret these various scenarios, we return to the three original questions to pinpoint the implications of this study for future quantitative research on learner language.

First, we attempted to discover whether or not the ICLE is large enough to find linguistic features that vary reliably across levels. We assessed reliability through the machine-learning technique of creating a prediction model from a training set of data and testing it on the rest of the data. In so doing, we used just a fraction of the ICLE corpus for each part of the analysis. Moreover, before partitioning the data into training and test sets, we had already divided the corpus two times. First, in attempting to identify the best indicators of level to use as the dependent variable, we had already omitted the essays whose writers had spent more than 3 weeks in an English-speaking country. Second, in attempting to group essays in a manner that would allow for generalizability of results to a particular type of group, we
conducted two separate analyses—one using essays from writers who spoke only English and their native language and a second from those who spoke additional languages as well. Thus, the ICLE that we began with contained approximately 2,000,000 words, but the data sets for which we were actually assessing reliability of prediction were much smaller. The assumption of researchers working with learner corpora has been that large quantities of data would result in reliable results, but, if that assumption about reliability is actually to be tested in the way that we did it, the definition of large needs to be reconsidered in order to create both a training set and a test set of data for a group to which the research results should generalize. In addition, having a large dataset of learner language without a clear and reliable indication of the competency levels of those learners makes studies on the variability of learner language based on that dataset difficult.

Using the subsets of the ICLE corpus, we found that the reliability for the test data was not sufficient. One reason could be that we are only looking at lexical patterns in the essays. Using lexical patterns is advantageous because, as we saw, they can capture a lot of the variability in the data and they are readily accessible without the need for a lot of sophisticated algorithms. However, language development is reflected in more than words, and it may be that incorporating syntactic information and other aspects of text would help. Another reason for the poor performance on the test data could be the questionable validity of the dependent variable. It seems that the amount of time studying English should be relevant as a dependent variable in this analysis, but a rating of the specific essays used in the corpus would have produced a measure that was closer to the actual language used in the analysis. Unfortunately, the subsample that we attempted to rate resulted in ratings with low reliability, and therefore we were unable to add this piece of information to the existing corpus. Thus, the essays themselves create a corpus that is very homogeneous—much more so than the advanced ESL writers that we work with at the university level.

Second, we wanted to see how automatically identifiable lexical features could be used to signal levels. We therefore included levels of vocabulary from a variety of word lists as well as some specific lexical items that have been identified in studies of ESL writing. Overall the best predictors based on lexical profiles seem to be BNC and AWL/UWL word lists. BNC proved more useful than GSL and HK because it simply contains more lexical items and is based on frequencies in naturally occurring text. The AWL and UWL word lists were also useful because they contain words that are likely to be used in higher level academic writings, thus making them good predictors for more advanced learners.

The third question addressed the issue of an appropriate statistical methodology for selecting good predictors from a large number of prospective ones and estimating the reliability of the analysis. Machine-learning techniques such as decision trees hold promise for analysis of multiple features as predictors of levels.

These are three basic questions about investigation of variation in advanced ESL learners’ writing which need to be addressed. Answers to these questions would help applied linguists move closer to analysis of learner corpora in a manner that is useful for development of learning and assessment tools. We used the only large corpus of advanced learner language that was commercially available, automatically identifiable features, and a statistical approach developed for this purpose. Results showed some promise for the use of the lexical features. The statistical procedure provided results that were interpretable and provided an overall answer to the question of how reliable the predication was. The corpus, however, proved to be very difficult to use for this purpose because of the lack of information it contained about the evaluation of the essays. Although we worked with the information available about learners’ level of development, it is not clear that the level estimates were valid. Future research investigating reliable differences in learner English needs to start with the arduous task of gather-
ing a learner corpus. In such a corpus, a valid indicator of the level of the essays is needed for each sample of learner language, and therefore the issue of corpus collection intersects with language assessment in a critical way.

REFERENCES


AUTHORS’ BIODATA

Nick Pendar is Assistant Professor of computational linguistics and natural language processing at Iowa State University. He teaches courses in theoretical and computational linguistics, and his research interests include constraint-based theories of natural language, intelligent tutoring systems, automatic essay scoring, and text categorization.

Carol Chapelle is Professor of applied linguistics. She is also Chair of the linguistics program at Iowa State University and Past President of the American Association for Applied Linguistics. Her research explores issues at the intersection of computer technology and applied linguistics. She has authored several books on computer applications and second language acquisition, teaching, and assessment.

AUTHORS’ ADDRESSES

Nick Pendar
Applied Linguistics & Technology Program
Department of English
355 Ross Hall
Iowa State University
Ames, Iowa 50011
Phone: 515 294 3368
Fax: 515 294 6814
Email: pendar@iastate.edu

Carol A. Chapelle
Applied Linguistics & Technology Program
Department of English
339 Ross Hall
Iowa State University
Ames, Iowa 50011
Phone: 515 294 7274
Fax: 515 294 6814
Email: carolc@iastate.edu