Exploring the proficiency of English learners
The TREACLE project

Mick O'Donnell, Susana Murcia, Rebeca García,
Clara Molina, Paul Rollinson
Department of English Studies
Universidad Autónoma de Madrid
michael.odonnell@uam.es

Penny MacDonald, Keith Stuart, Maria Boquera
Department of Applied Linguistics
Universitat Politècnica de València

Mick O'Donnell, Susana Murcia, Rebeca García, Clara Molina, Paul Rollinson, Penny MacDonald, Keith Stuart, Maria Boquera. (2009) "Exploring the proficiency of English learners: The TREACLE Project". Proceedings of the Fifth Corpus Linguistics, Liverpool.

Abstract

To organise the teaching of English as a Foreign Language (EFL), it is important to have a clear picture of the grammatical competence of the learners at each level of proficiency. The TREACLE project has been set up to develop a methodology for producing grammatical profiles which detail the degree to which learners at each proficiency level have mastered the various grammatical features they need to know. The corpus-based methodology takes a two-pronged approach, using automatic grammatical analysis, to see what the students are getting correct at each level, and manual Error Analysis to see what the learners do wrong. This paper describes this methodology, providing details on the derivation and use of the automatic grammatical analysis, and on the error annotation process. Some notions of profiling will also be discussed.

1 Introduction

This paper describes the goals and progress of the TREACLE project, a project recently started within two Spanish universities. The main aim of the project is to profile the specific grammatical skills of Spanish university learners of English at various proficiency levels, and, on the basis of these profiles, develop proposals for re-designing curriculum and teaching materials particularly focused on the real needs of Spanish students at distinct proficiency levels.

To produce this profile of grammatical skills, we are analysing a large corpus of essays written in English by Spanish university students, using the state of the art in corpus software. We intend to analyse each proficiency level separately to see which grammatical structures the students (taken as a group) have mastered, which structures they are still developing, and which they have not yet attempted. We will use these profiles to make recommendations as to the topics taught at each proficiency level, the sequence of topics, and the emphasis given to particular topics.

Learner proficiency has been explored in various ways. Error Analysis (Corder 1967) has been used to this end (James 1998): studying the errors that learners make at different levels of proficiency. However, while errors do cast invaluable light onto the process of learning a language, we believe that for the proper measurement of student proficiency, this approach by itself is not complete: some students make few mistakes because they stick within the language forms that they are familiar with, while more proficient students might make more mistakes because they continually experiment with forms they are less familiar with. The assessment of a

student using Error Analysis by itself would over-reward the cautious student and penalise the experimental learner.

A better measure of learner proficiency thus needs not only to look at what the learner does wrong (error analysis), but also what they do right: the syntactic structures they use, and how frequently they use them.

Following this view, we are thus developing a methodology for studying the proficiency of English learners based on a two-pronged analysis:

- (Automatic) **syntactic analysis** of student texts to see what structures students are attempting;
- (Manual) **error analysis** of these texts to see what they do wrong.

The rest of this paper outlines our approach in detail. Firstly, Section 2 describes related work in pedagogical applications of learner corpora. Section 3 introduces the TREACLE project, and the learner corpora we are using to derive the profiles. Section 4 then outlines our process of analysis of the corpus, both in terms of syntactic analysis of the corpus and the error analysis, and discusses the derivation of grammatical profiles from the annotated corpus. Section 5 provides summary and conclusion to the chapter, and some future developments of the work.

2 Pedagogic Applications of Learner Corpora

To gain systematic knowledge of the problems faced by learners of a foreign language, researchers have recently turned to the study of learner corpora. There have been two main approaches to studying these corpora. Firstly, some have followed the Error Analysis (EA) path (James 1998; Corder 1967), studying the errors made by learners of a foreign language. By finding systematic explanations behind errors, the researchers hope to better understand the process of learning a language, and to distinguish errors which are part of the general developmental process from those which derive from linguistic traits of the mother tongue.

The second approach uses corpora to explore the "interlanguage" of learners, based on the hypothesis that the language produced by a learner has a grammar of its own, and thus the errors are systematic in relation to this interlanguage (Selinker 1972). They use *Contrastive Interlanguage Analysis* (CIA) (Granger 1998:12) to chart the difference between the interlanguage (IL) and the language being learned (L2), often attempting to explain these differences in terms of the influence of the mother tongue (L1). Such studies usually do not focus on errors, but rather on the syntactic structures or word choices used, comparing the frequency of use in learners compared to native producers (e.g., Biber and Reppen 1998 on complement clauses; Aijmer 2002 on modal words; Römer 2005 on progressive forms, etc.).

Both EA and CIA practitioners have applied their results towards pedagogic purposes. For instance, annotated corpora can be made available to students in the classroom to explore particular grammatical phenomena (e.g., displaying examples of causative constructions) to allow the students to discover grammar rules for themselves (cf. the 'corpus-aided discovery learning' of Bernardini (2002)). Studies of overuse/underuse of syntactic features can be presented to students, making them aware of how their own writing differs from native writing.

One interesting application involves using the corpus as a source for student exercises. Keith Stuart, a member of TREACLE, developed *TextWorks* (Stuart 2003), a system which helps teachers produce exercises from a learner corpus, including cloze, comprehension questions, jumbled sentences, matching exercise and multiple choice questions.

2.1 Towards Learner Profiling

Our particular interest is in using the learner corpus for curriculum design. There have been uses of *native* corpora to inform pedagogical design (Grabowski & Mindt 1995; Biber *et al.*

1994). However, the application of *learner* corpora to pedagogical design is much rarer. Work charting the use of syntactic structures in relation to proficiency levels includes that of Díez Bedmar (2007), who compares the use of the article system in upper secondary and lower tertiary learners of English. Granger (1999) explores verb tense errors in high proficiency learners, and concludes that this can lead to more targeted teaching of this area for this proficiency level. Note however that the existing work either explored single structures (or topics) over several proficiency points, or looked at a single proficiency point.

More ambitious studies target a wider range of syntactic structures at a number of distinct proficiency levels. However, most of these studies seem to get tied up on the construction of the corpus, and never reach the point of pedagogical application. For instance, the good intentions of Muehleisen (2006) are clear when she says: "The corpus is being created to better understand the state of students' writing as they enter SILS and as it develops through the course of their first few semesters. The corpus will be immediately useful for the SILS language program developers in creating course material for the writing classes" (p.119). However, the paper makes clear that this work had not been attempted at that point. Rankin (2010) also considers applications to the curriculum, proposing to compare the kinds of adverb errors found in student texts to those taught in the course, with the objective of including material for common errors where they are not already covered. However, this is currently just a proposal. As Meunier (2002) said, "the actual implementation of corpus research results in curriculum design is timid, if not absent." (p.123).

There are recent indications that interest is increasing in the application of learner corpora to curriculum design and learner profiling. 2007 saw the First International Conference on Corpus-Based Approaches to ELT, held in Castellón, Spain, with some attention to these issues (selected papers to be published in Campoy *et al.* 2010).

Of particular interest, English Profile[1] is a research group based in the U.K. that aims "to provide a detailed set of Reference Level Descriptions for English. Linked to the Common European Framework of Reference for Languages (CEFR), these will provide specific criteria for describing what a learner knows at a particular level of English".[2]

3 The TREACLE Project

The TREACLE project is a co-operation between the Universidad Autónoma de Madrid (UAM) and the Universitat Politècnica de València (UPV). TREACLE stands for ***T**eaching **R**esource **E**xtraction from an **A**nnotated **C**orpus of **L**earner **E**nglish*. The project has been informally in process since January 2009, and will receive funding from the Spanish Ministry of Education (research grant: FFI2009-14436/FILO) from January 2010 until the end of 2012.

This paper covers part of the project's goals: profiling learner proficiency to help inform English teaching curriculum design. The project however has further goals, to provide a web-based language learning system which dynamically adapts materials and exercises presented to the student by reference to the student's current performance within the system, and the proficiency profiles derived above. See Section 5.1 for more details.

The project is making use of two corpora written by learners of English at University level within Spain.

3.1 The WriCLE corpus (UAM)

The WriCLE corpus (Rollinson and Mendikoetxea, 2008; Mendikoetxea *et al.*, this volume) is a corpus of essays written by Spanish university students learning English at the UAM. The corpus was collected by Paul Rollinson during 2005-2008, and consists of 719 essays containing approximately 710,000 words. WriCLE stands for ***Wri**tten **C**orpus of **L**earner **E**nglish*.

The corpus was collected as part of the WOSLAC project (Research grant HUM2005-01728/FILO from the Spanish Ministry of Education: "The lexicon-syntax and discourse-syntax interfaces: Syntactic and pragmatic factors in the acquisition of L2 English and L2 Spanish". See Chocano *et al.* 2007 for details).

The corpus consists of essays submitted as class assignments within Academic Writing courses in the first and the third year of the English Studies degree. Paul Rollinson, the teacher, then normalised the submitted text in accordance with the process used in the ICLE corpus (Granger 2003): all personal data, titles, footnotes, endnotes, graphics, maps and bibliographies were stripped out, and quotations and references were replaced with <Q> and <R> respectively. 752 essays were collected, and 43 were eliminated where the learner's L1 was not Spanish. The essays are stored in electronic format and range from 500 words up to 2,000 words.

Various metadata were collected as well: A *Release forms/Essay Profile*: was provided by the learner for each essay, detailing the resources they used to write the essay. The form also includes a section where the student grants permission for the essay to be used for research purposes. They also provided a *Learner Profile*, detailing age, gender, language background, English language proficiency, etc. Additionally, students also took the *Oxford Quick Placement Test* (UCLES 2001) at a time close to the writing of the essays. The normalised text files, and metadata, are now available for free download for research purposes from: *http://www.uam.es/woslac/Wricle/.*

3.2 The MiLC corpus (UPV)

The development of the MiLC Corpus (Andreu et al 2010) has been carried out in two different stages. The project was started in 2004 by the members of the DIAAL Research Group and other collaborators at the UPV, with the aim of obtaining information concerning the interlanguage of university students attending language classes and the influence of their L1 in the learning process. Initially the corpus comprised a great variety of written work including formal and informal letters, summaries, curriculum vitae, essays, reports, translations, synchronous and asynchronous communication exchanges, business letters and so on, of the students learning English, Spanish and French as a foreign language, and also Catalan, as a first, second or foreign language. The student population at the UPV is approximately 30,000, and more than 1500 credits on the curriculum are assigned to the Department of Applied Linguistics. The degree courses on offer include Architecture, Fine Arts, Civil Engineering, Agronomy, Applied Computer Science, Industrial Engineering, Geodesy and Telecommunications. The students have to read a large amount of scientific and technical texts, produce written texts themselves which may be of a specific nature and related to their mainstream subjects, or involve general language output. Included in the multilingual corpus we have a 120,000 word sub-corpus involving a series of intercultural telematic simulations which has been analysed (MacDonald 2004) using the Error Tagging Method (Dagneaux *et al.* 1996) and Error Editor developed by the Centre for English Corpus Linguistics at the Université Catholique de Louvain in Belgium.

During the second stage of the corpus building, the bulk of the corpus has been taken from the students learning English as a foreign language who have been assigned to write a text on the topic of Immigration. In the same way as the WriCLE Corpus, these learners have also been graded according to the CEF, whilst the other variables, such as their age, gender, mother tongue and the second languages they study are included in the database, and all texts are used for the research project with the permission of the writers. This sub-corpus has been named the UPV Learner Corpus, and at present it stands at 150,000 words. For the on-going error analysis and syntactic analysis the *UAM Corpus Tool* is being used.

It is an important aspect of the project that we bring together two groups which have very different students, those within a dedicated English Studies program at the UAM, and those

studying English for Specific Purposes at the UPV. We hope to explore the differences between our students, for instance to see if the learning environment affects the grammatical profiles of the students.

4 Exploring Proficiency

In this section, we outline our methodology for deriving learner profiles from learner corpora.

4.1 Corpus Annotation with UAM CorpusTool

Our proficiency profiles are derived from a corpus analysis of the learner corpora. For both the automatic syntactic analysis and manual error analysis, we use *UAM CorpusTool* (O'Donnell 2008), which allows manual and automatic annotation of collections of text at multiple annotation layers. Figure 1 shows the main window of *UAM CorpusTool* with the TREACLE corpus open. The window shows that 5 annotation layers are defined:

- *Document*: where features relating to the document as a whole and its writer are recorded. Currently, we record proficiency level of the writer, their university year, gender and native language (for this study always Spanish), and the language of the document (in this study, always English).
- *Sentence*: Each text is automatically segmented into sentences, and this annotation layer records the start and end of the sentences. The interface allows the user to manually correct errors made by the software. We do not code features at the sentence level, the level is used to allow us to make cross-level queries regarding sentence units (e.g., *sentence containing nonfinite-clause* would find all sentences containing a non-finite-clause).
- *Error*: This layer is used to record the manual annotation of errors. See below for more details.
- *Grammar*: This layer is used to record the automatically generated grammar analysis of each sentence. For more details, see below.
- *STNDFParse*: To produce our grammatical parse, *UAM CorpusTool* first parses each sentence using the Stanford parser (see below). The Stanford parse trees are recorded as a layer on their own.

Below the list of layers, the window shows the files included in the corpus. For each file, there is a button for each of the layers, and pressing one of these buttons brings up the editing window for the file at that layer.

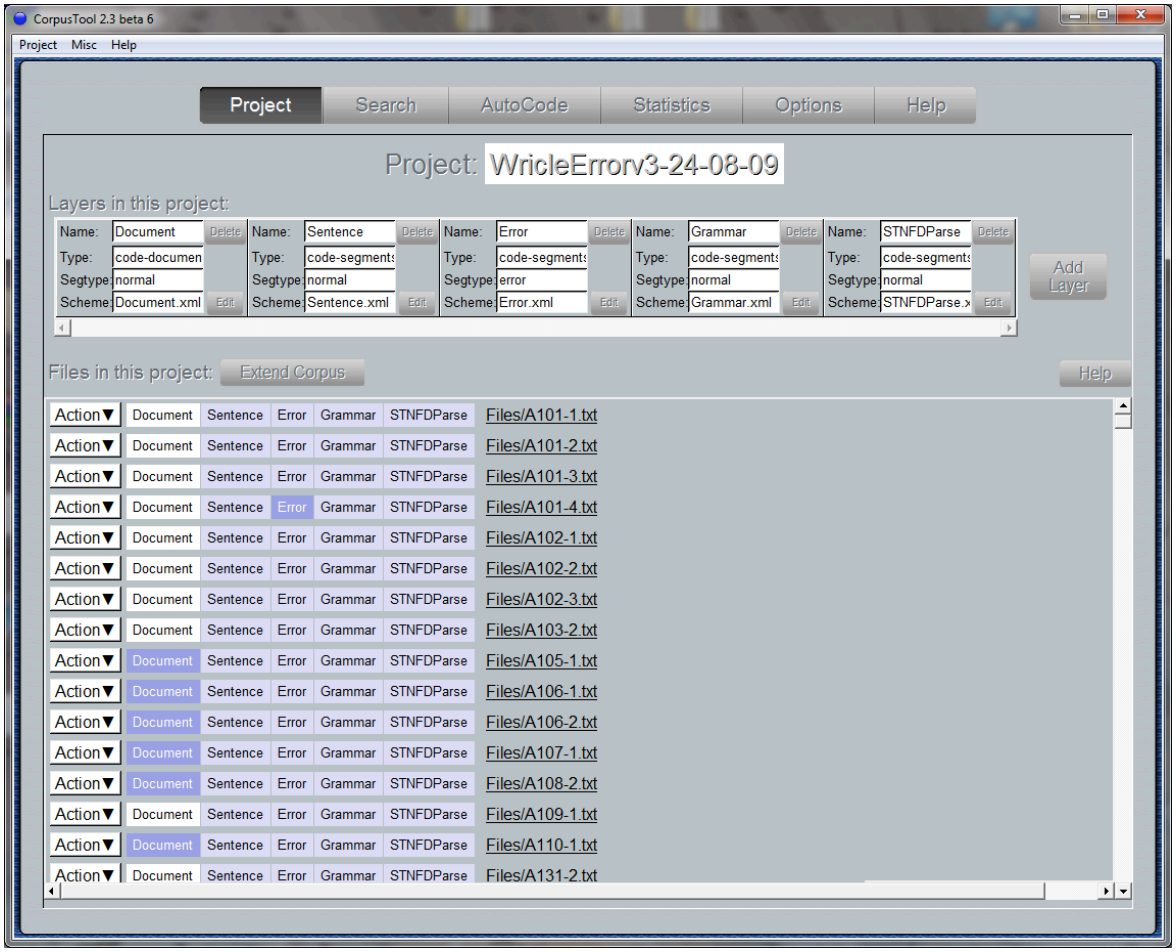


Figure 1: Main Window of UAM CorpusTool with the WriCLE corpus open

4.2 Automatic Grammar Annotation

The first step in profile construction involves the automatic grammatical analyses of each text. As each text is added to the corpus, *UAM CorpusTool* calls the Stanford Parser (Klein and Manning 2003) to produce a syntactic parse tree for each sentence in the text (sentences with 40 or more words are ignored as they take too long to parse, and the reliability of the parse is poor).

The Stanford Parser produces parse trees similar to that shown in Figure 2, with a context-free phrase structure grammar (PSG) representation. However, for EFL research, we believe that traditional grammar categories are more appropriate (Subj/Pred/Obj, active/passive, relative-clause, etc.), as these are the categories most often used in the EFL classroom.

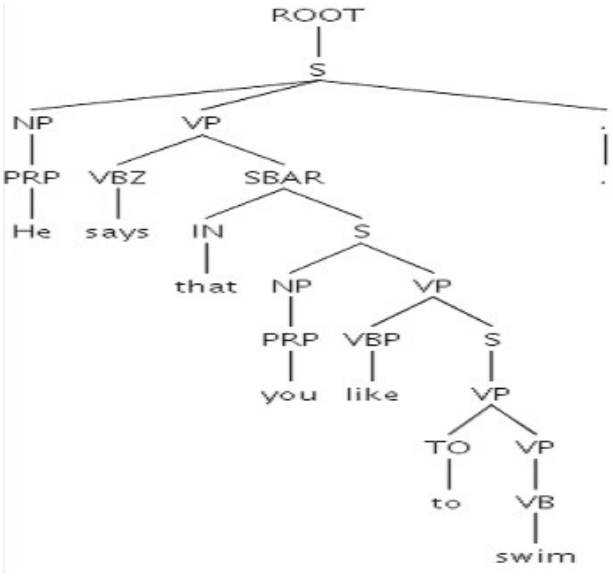


Figure 2: A parse tree produced by the Stanford Parser.

Because of this, we extended *UAM CorpusTool* to automatically transform the PSG analysis into a traditional grammar analysis, showing function structure at each level (Subject, Object, Adjunct, etc.) and assigning grammatical features to each unit (e.g., passive-clause, relative-clause, modal-clause, pronominal-phrase, etc.). This transformed analysis is available as a separate annotation layer, and can be modified by the user where necessary (due to either errors in the Stanford Parser or in the transformation). Figure 3 shows the annotation window for one student essay. The grammatical features of the selected element (the Subject of the first sentence, displayed in grey) are shown in the box in the lower part of the window.

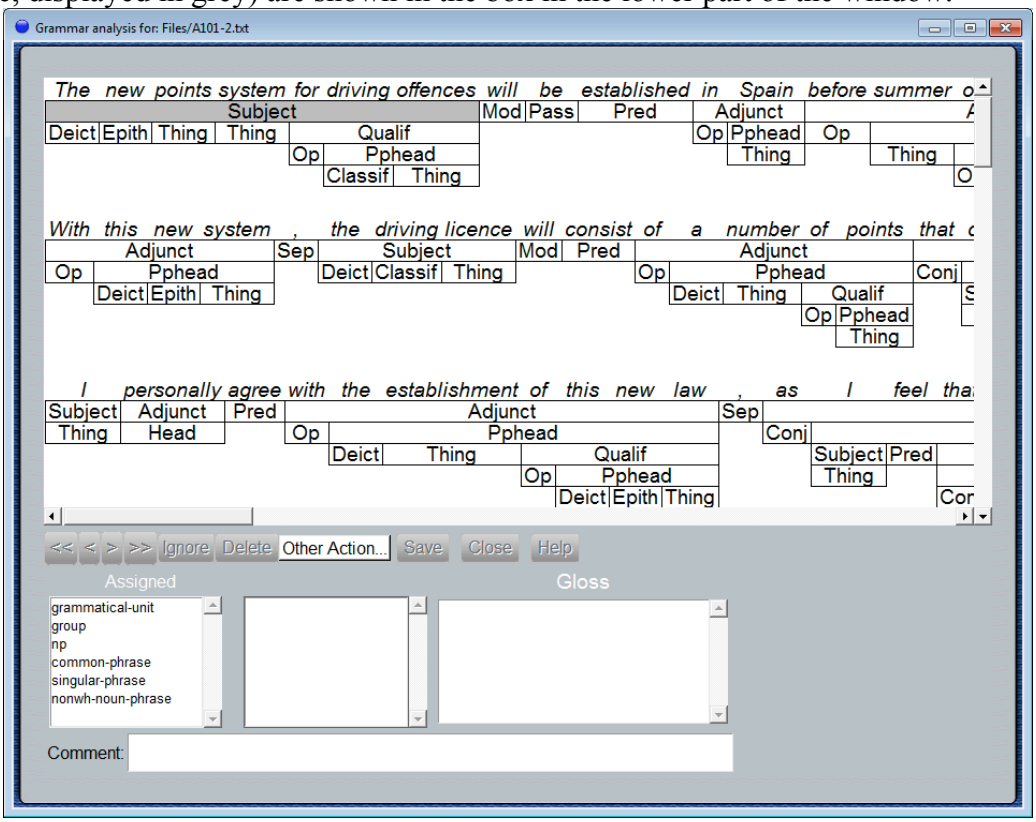


Figure 3: A window showing the part of the grammatical analysis of a text

Work in this phase of the project is yet incomplete, although we have produced the grammatical analysis of a 500,000 word sub-corpus containing 20,000 sentences. Figure 4

shows the features covered in the analysis of clauses, although the recognition of relative-clauses, fact-clauses (*He said that he was coming*) and linked-subjunctive-clauses (*He left because he was tired*) is not yet complete.

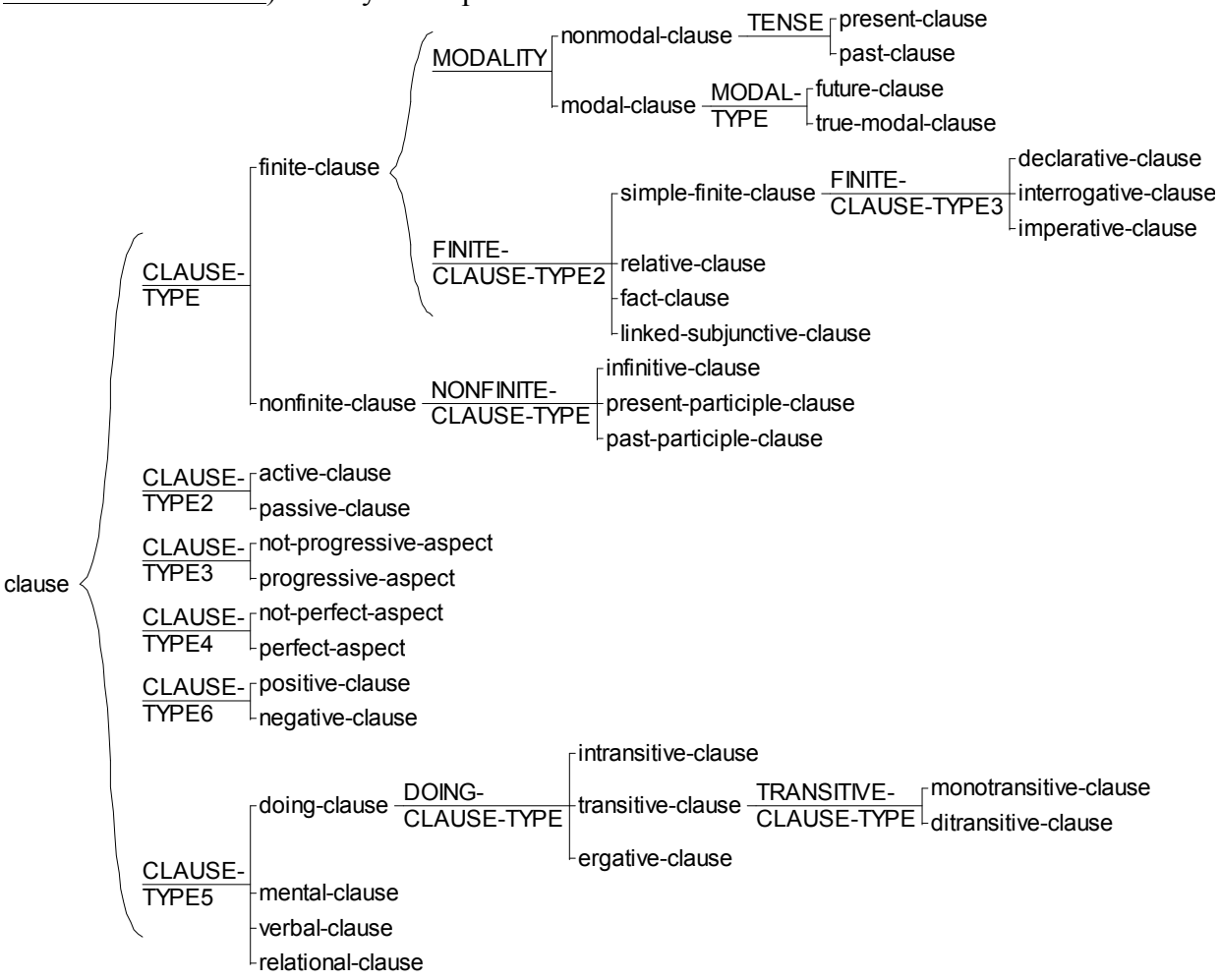


Figure 4: The clause features of the Grammar scheme

4.3 Error Annotation

The second step in analysing the learner corpus consists of manually annotating the errors in each text. The Error annotation window for a text appears as in Figure 5. The human coder reads through the text looking for errors, and when one is located, they select the text of the error (step 1 shown in the diagram). When text is selected, the Correction field at the bottom of the window changes to display the selected text (step 2 in the figure) and the coder should replace this with the corrected text.

Just above this are some boxes which allow the coder to assign error codes to the error. The system is provided with a hierarchically organised set of error codes (see below) and the user walks through the hierarchy to assign a code, e.g., selecting first “grammar-error”, then “np-error”, then “determiner-error” and then “determiner-choice-error”.

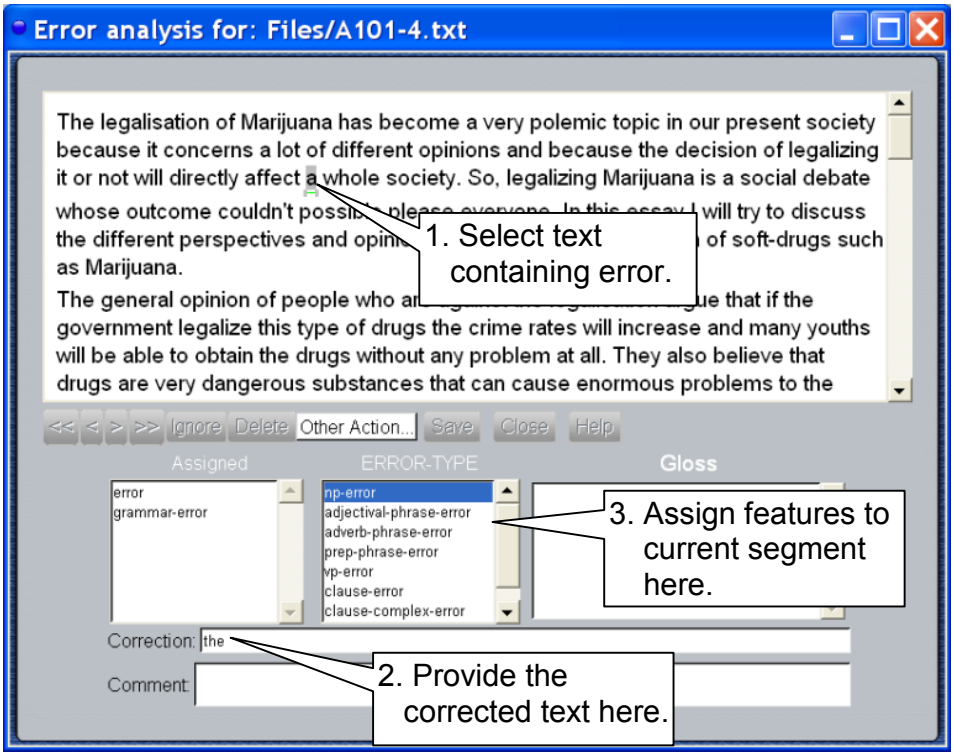


Figure 5: Error Coding a Text

This process of gradual refinement of error codes facilitates the coding process, because often the coder does not know what leaf of the error tree they should code, but can make a series of decisions on more general grounds (e.g., *Is it a grammatical or lexical error?*).
To facilitate the coder's job, the error scheme incorporates coding criteria ('glosses') with each feature in the scheme, which are displayed in the coding window. For instance, the glosses attached to the choice between lexical-error and grammatical error are:

- **lexical-error**: Errors relating to a single word, and not affecting other parts of the phrase or clause. This includes spelling errors and false friends, etc., but does not include cases where wrong inflections are used.
- **grammar-error**: Errors where some grammatical rule is broken (wrong class for slot, word order, agreement problem, missing but necessary element, present but unnecessary element, etc.)

UAM CorpusTool uses stand-off annotation to record its error annotation, meaning that the error annotation is not stored in the same file as the original text. Rather, the software records character offsets of the start and end of the error segment. These offsets, along with the features assigned and the corrected text, are stored in an XML format (see Figure 6). The use of stand-off annotation, as opposed to traditional embedded mark-up, means that the system can represent overlapping error segments without problem.

```
<document>
 <header>textfile>Files/A101-3.txt</textfile></header>
 <segments>
   <segment id="44" start="11" end="16"
     features="error;lexical-error;spelling-error"
     state="active" correction="Mayor" />
   <segment id="45" start="77" end="86"
     features="error;lexical-error;spelling-error"
     state="active" correction="vehicles" />
 ....
```

Figure 6: Part of the XML content of an Error layer file

The Error Scheme: The TREACLE error scheme has been designed from the start to integrate into a University level teaching. The main design principle has been to ensure that the error scheme should map cleanly onto the organisation of grammar topics which are taught within EFL courses. The main reason for this is that our goals are pedagogical, and we later want to be able to recover those errors which are relevant to each teaching topic, so as to inform our teaching of that topic.

To demonstrate our approach, we will focus on one particular part of the error scheme. Let's assume the student has written "this results". At the root of the error hierarchy, we distinguish between various types of error, including lexical errors, grammatical errors, pragmatic errors, etc. (see Figure 7). The "..." after a feature indicates that there are further choices under that feature. A Coding Criteria document (15 pages long) provides clear criteria for determining which of these categories is appropriate for a given error.

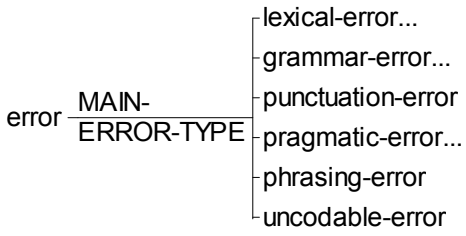


Figure 7: Main Error Types

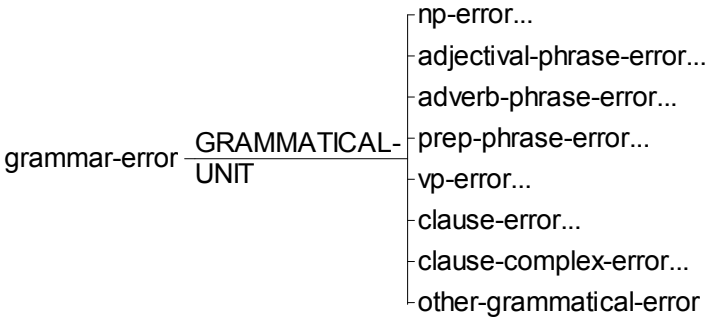


Figure 8: Basic grammatical error types

Assuming the error is grammatical, one next chooses the type of grammatical error (see Figure 8). While some error coding systems are more oriented to coding errors in terms of the part of speech of the word concerned, our approach is more focused on the grammatical phrase in which that word occurs. Thus, while teaching about adjectival phrases, we can find errors within adjectival phrases, whether they involve the adjective itself, or any adverbial premodifier of the adjective.

The principle we use to determine the syntactic unit of an error is as follows:

- If the error is in regards to the appropriateness of a segment for its slot, the unit of the error is the unit which contains the slot (e.g., an error in Deictic slot will be coded as an NP error).
- If the error is in regards to a disagreement between two slots (e.g., Deictic and Head of an NP disagree in number), then the unit of the error is the unit which contains both slots.

Continuing with our example, assume the error is within the selection of the determiner. We thus select *np-error*. This leads to the next level of delicacy, as shown in Figure 9. Our division of error codes within the NP reflects the fact that, in many courses, the teaching of the Noun Phrase is divided into topics: determiners, pre-modifiers, the Head, and post-modifiers.

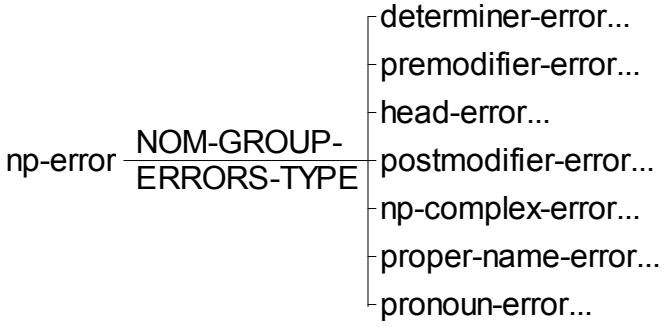


Figure 9: NP Error sub-classes

Selecting *determiner-error*, we are presented with the next set of choices, as shown in Figure 10. Note that the error codes shown here are not exhaustive, the coder can add new error codes as examples are encountered in the learner texts. The particular example we started with, "this results", would be coded as *determiner-agreement*.

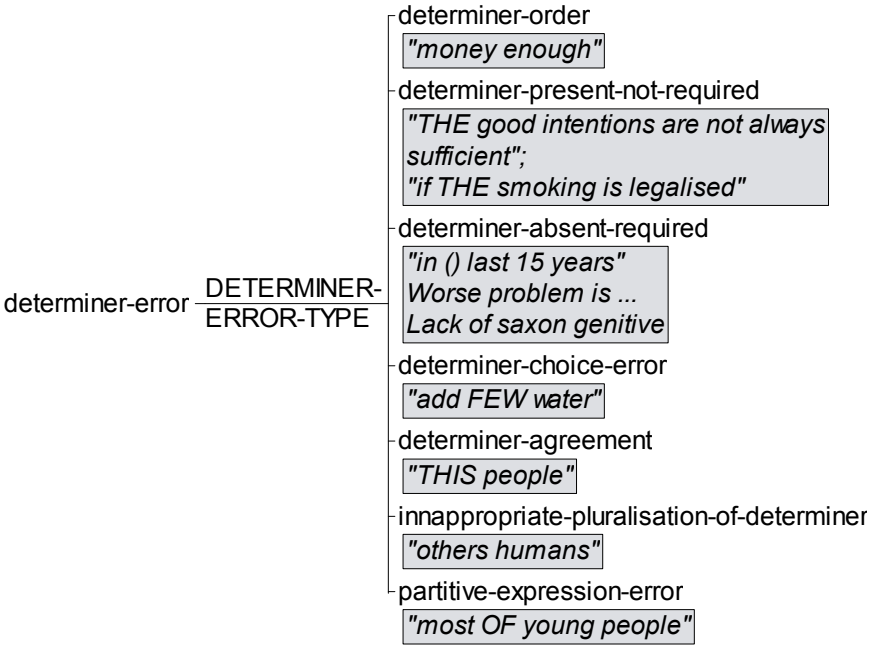


Figure 10: Subtypes of determiner errors

So far we have coded only a small sub-corpus of errors, 17 texts containing 12,500 words, with just over 1,000 errors. However, this trail established that the error scheme and annotation process is viable. We intend to code 10,000 errors by the end of 2011.

Even this small sample of errors starts to show viable results. Figure 11 shows a comparison of the percentage of errors in each of the main error categories between first and third year students. The graph suggests that the number of grammatical errors has decreased by third year, with punctuation errors increasing. Note however that this may be an artefact of coding: as the quality of the student writing increases, the coder can pay more attention to issues such as punctuation.

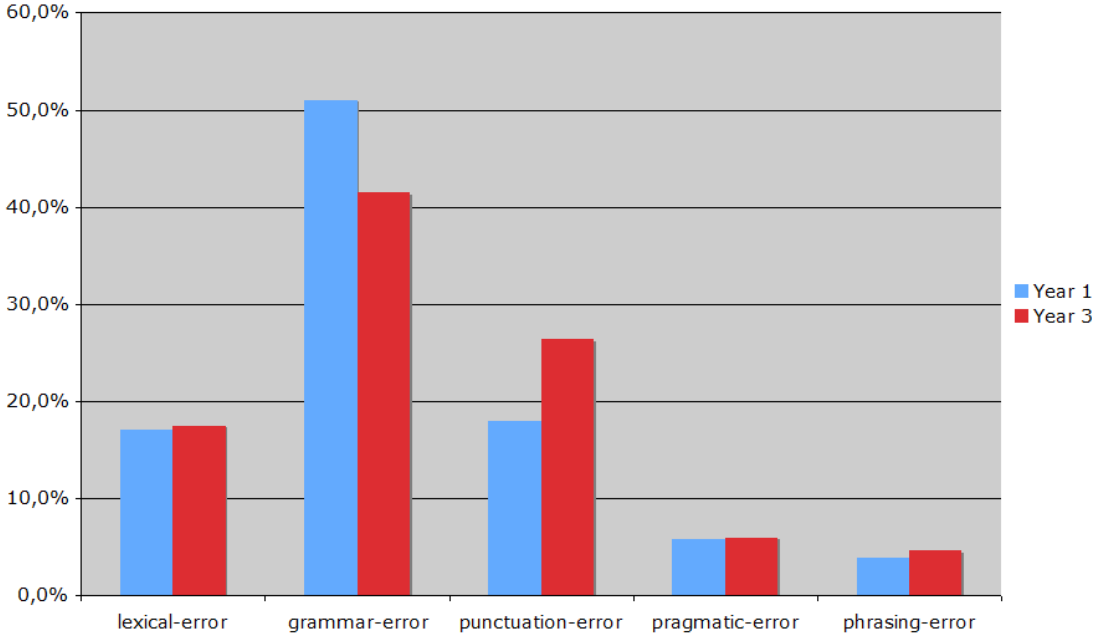


Figure 11: Change of error types between 1st and 3rd year students

One of our pedagogic applications can be seen via the results presented in Figure 12. By examining the types of errors made at each proficiency level, we can determine how much teaching time to spend on each area. This graph suggests more attention is required on NPs and VPs.

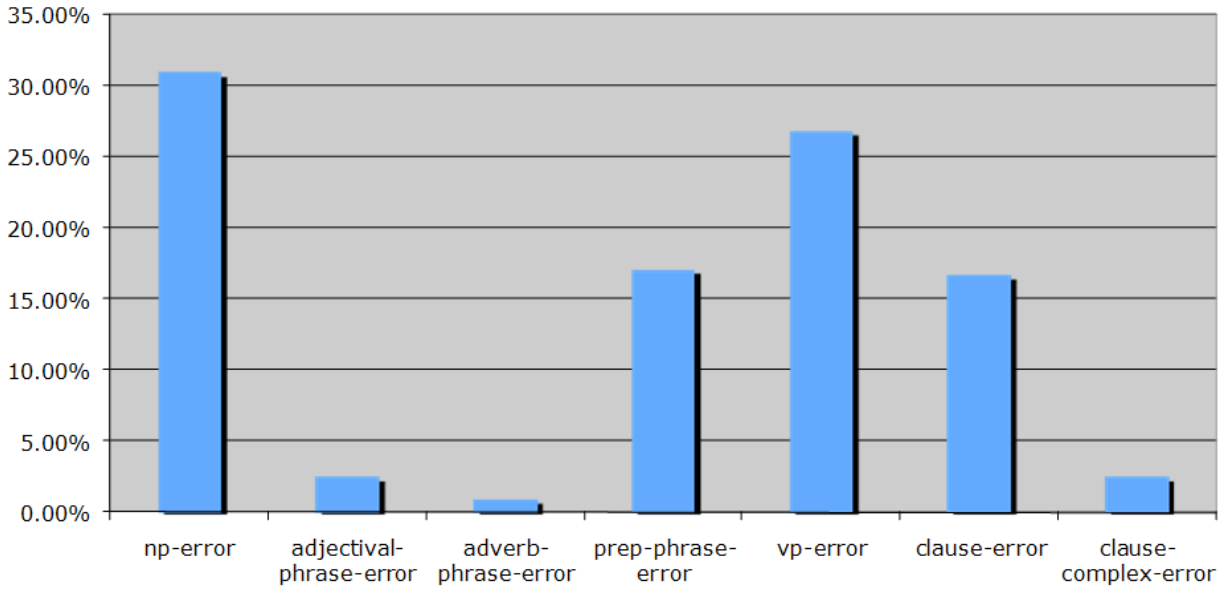
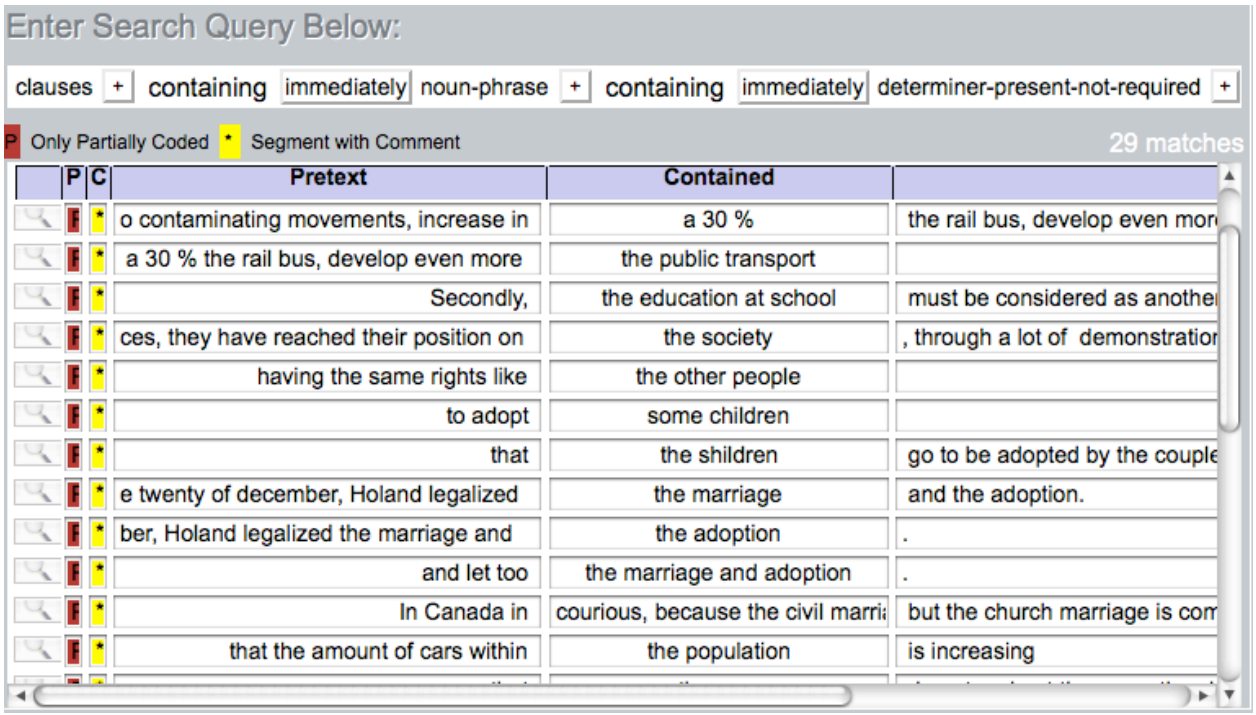


Figure 12: Distribution of subtypes of grammatical errors

A second application of this work is visible in Figure 13, which shows the results of a search operation, finding all NPs which contain an error of type: *determiner-present-not-required*. Such searches can be used to recover examples which can be used for the teaching of particular topics.



P	C	Pretext	Contained	
F	*	o contaminating movements, increase in	a 30 %	the rail bus, develop even mor
F	*	a 30 % the rail bus, develop even more	the public transport	
F	*	Secondly,	the education at school	must be considered as anothe
F	*	ces, they have reached their position on	the society	, through a lot of demonstratio
F	*	having the same rights like	the other people	
F	*	to adopt	some children	
F	*	that	the shildren	go to be adopted by the couple
F	*	e twenty of december, Holand legalized	the marriage	and the adoption.
F	*	ber, Holand legalized the marriage and	the adoption	.
F	*	and let too	the marriage and adoption	.
F	*	In Canada in	courious, because the civil marri	but the church marriage is com
F	*	that the amount of cars within	the population	is increasing

Figure 13: Corpus search to recover teaching examples

4.4 Assessing proficiency

The TREACLE project has not yet reached the point where we can begin assessing proficiency. We need to complete more manual error analysis (we will use a 100,000 word corpus as a starting point, to be annotated over the next 2 years), and extend the automatic grammar analysis to cover a range of other grammatical phenomena. However, in this section we present our current thinking of how we will proceed.

Firstly, it is common in analysing learner corpora to focus on issues of frequency of use of grammatical features, with learners sometimes under-using features, and sometimes over-using them. The frequency of usage of a feature is very register-dependent, so not to be trusted. For our purposes, we are not so concerned with how frequently the feature is used, but rather with whether the student has the competence to produce the structure at all, and if so, whether they produce the structure correctly. Adapting to native frequency of use is something that comes with practice. In relation to curriculum design, what is at issue is whether the student is competent with the feature, not how often they use it.

One point here is that some of the under-use/over-use studies look at usage within a group of learners, and are thus mixing learners who do not yet use the structure with those who do. We would rather measure the proficiency of a group in regards to a particular structure by counting the number of individuals who fall into the following groups:

A. Learners who don't use the structure,
B. Learners who use the structure but with errors,
C. Learners who use the structure without error.

Some flexibility is required here, as even a native writer may make mistakes with a structure, so we suppose that students who use a feature with 90% correctness should be put into group C, and students who use the structure very infrequently might be placed in group A.

Another possibility is to assess each student in terms of two indexes:

1. Degree of correct usage / degree of usage: basically, the proportion that they use the structure and get it correct, indicating their degree of competence.
2. Degree of usage / degree of usage by natives: this proportion measures under-usage or over-usage of the feature (although note our reservations on the use of this statistic).

We are also interested in performing some kind of cluster analysis on the students in each proficiency level, to see if they separate into distinct types, e.g., two students might score the same on the proficiency tests, but have totally different approaches to learning. By separating students into learner types, patterns might become clearer, and these learner types could be provided with different teaching materials.

As stated above, we have not yet reached the point in the project where we have resolved these issues.

4.5 Some early results

Below we show some early results from our study, using the 64,000 clauses automatically parsed within the WriCLE corpus. Firstly, Figure 14 shows the development of the use of passive clauses with increased proficiency. Note however that if we look at an individual's use of passives, we find that 6.5% of the B1 learners did not use a passive in their text (with average essay length over 800 words!) while above this level, only 1% of learners did not use the passive.

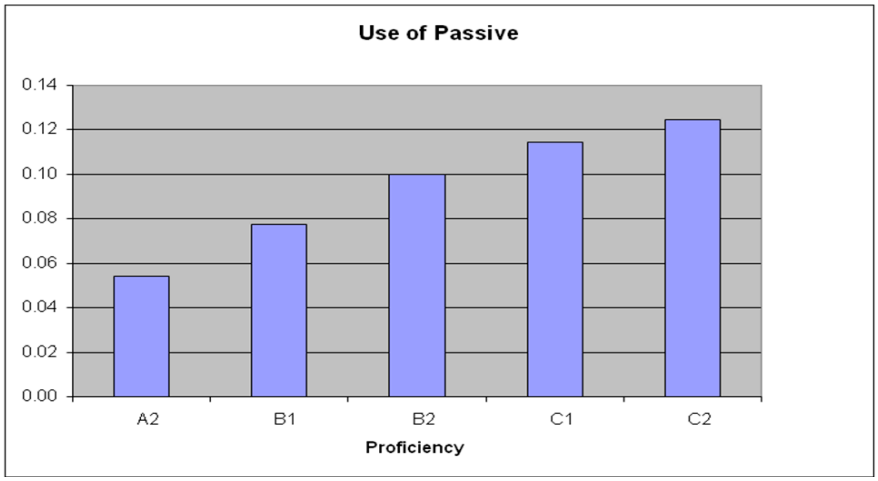


Figure 14: Use of passive clauses with increasing proficiency

Figure 15 shows how use of present-participle clauses increases with proficiency, when data is conglomerated over all users at that proficiency level. Perhaps more informative, we can see from Figure 16 the percentage of learners who do not use *present-participle* clauses at all in their text decreases rapidly with increasing proficiency. By C1, all learners are using the structure. However we need more data from the error analysis to see if they are using the structure correctly. Ignoring this factor, the data suggests that this structure would most constructively be taught at the B1 or B2 stage.

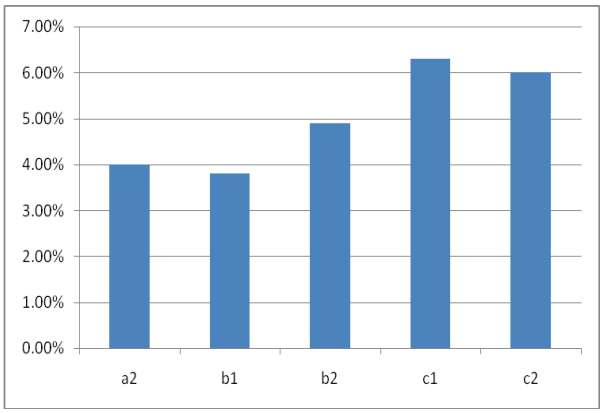


Figure 15: Percent of clauses which are *present-participle* clauses with increasing proficiency

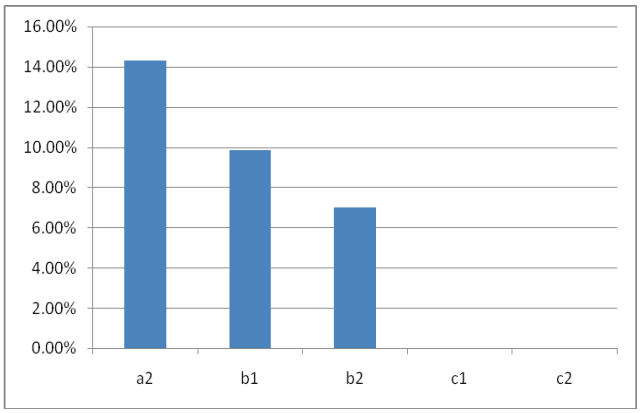


Figure 16: Percent of learners who do not use *present-participle* clauses

Similarly, Figures 17 and 18 show similar results for past-participle clauses. The data suggests that these clauses are acquired a bit later in the learning process.

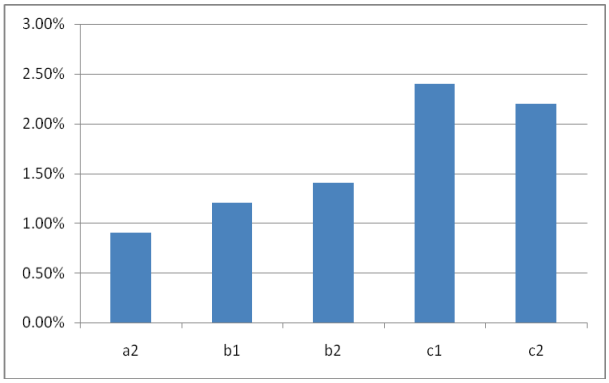


Figure 17: Percent of clauses which are *past-participle* clauses with increasing proficiency

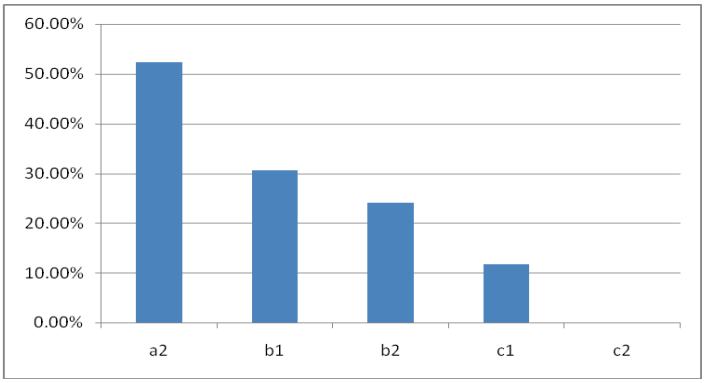


Figure 18: Percent of learners who do not use *past-participle* clauses

5 Conclusions

We have described the state of progress within the TREACLE project, which is developing a methodology for measuring the grammatical proficiency of learners of English with a range of grammatical structures. The project is still in its early stages, but already substantial results have been achieved, including:

- Development of software for error annotation (UAM CorpusTool);
- Development of an error annotation scheme linked to the pedagogic goals of the project;
- Verification of this error coding scheme on a small sub-corpus;
- Extension of UAM CorpusTool to derive a traditional grammar analysis from the PSG analysis provided by the Stanford Parser;
- Application of this tool to grammatically annotate a 500,000 word sub-corpus with detailed grammatical analysis.

Work in the immediate future will involve:

- Error annotation of a 200,000 word subset of the WriCLE corpus and equivalent in the MiLC corpus.
- Inter-coder reliability studies of the error coding;
- Extension of the automatic grammatical analysis to a range of other grammatical features;
- Application of the automatic grammar analysis to the remaining 200,000 words of the WriCLE corpus and the whole of the MiLC corpus.
- Further development of the method to assess student and group proficiency based on the automatic and manual annotation.

5.1 Long Term Extensions

Some longer terms developments that we intend to follow are discussed below.

Extension to other mother tongues and target languages: The proficiency profiles developed for a particular mother tongue are of course particular to that particular mother tongue. It is well established that the linguistic similarities and differences between a mother tongue and the target language influence the relative ease or difficulty of learning particular aspects of target language.

For this reason, the proficiency profiles developed with the TREACLE project for Spanish learners of English will be of little use for those in other language regions. However, the methodology developed within the project for constructing the profiles is readily applicable to the profiling of learners of English with whatever mother tongue.

For those interested in learners of language other than English, at the point of writing, the software we use only supports grammatical analysis of English. The software will later be extended to incorporate existing parsers of other languages, starting with Spanish (the *FreeLing* parser: Atserias *et al.* 2006) and German (the Stanford Parser). Our idea is to map the syntactic analyses produced by these parsers onto a grammatical analysis as close as possible to that which we use for English. However, this work is still for the future.

Extension to other teaching contexts: The methodology could also be applied to various teaching contexts. We have applied the methodology in the context of a dedicated English Studies program, but will soon apply the methodology to the MiLC corpus, which consists of ESP students. The application to levels outside of university (e.g., school) is also possible. One problem is that it is often difficult to get ESP and school students to produce long texts in

English. For our first study, all texts are over 600 words, while the MiLC corpus has texts of 300 words or less. With shorter texts, the non-occurrence of a particular grammatical structure could be explained by the text having fewer clauses.

Adaptive web-based education: A field that has recently emerged is called "adaptive web-based educational systems" (Brusilovsky and Peylo 2003). The goal in this field is the provision of courseware that intelligently keeps track of the abilities/knowledge of the user (the user model), and adapts the course material to this model. In this way, the courseware can adapt itself to the particular strengths and weaknesses of the student. We believe that this technology can have value in the Spanish university environment.

As part of the Spanish ministry funded project, we will construct an intelligently adaptive web-based learning system to complement traditional teaching. To better integrate with the work on our first aim, we intend this system to be driven by the learner profiles derived from our annotated corpus. A program provided with these profiles will have knowledge of the types of structures that a student at a particular proficiency level will have problems with, and can focus on these problems. Used in reverse, the program can observe the types of problems a student exhibits within the system, and use these observations to place the student's proficiency level. We can thus build a program which can provide exercises to a student matched to their current needs, and also track when they have progressed to a new level.

6 Notes

[1] http://www.englishprofile.org/
[2] http://www.englishprofile.org/index.php?option=com_content&view=article&id=11&Itemid=2

7 References

Aijmer, K. (2002). "Modality in advanced Swedish learners' written interlanguage". In Granger *et al.* 2002, pp. 55-76.

Andreu, M., A. Astor, M. Boquera, P. Macdonald, B. Montero and C. Pérez (2010). "Analysing EFL learner output in the MiLC project: An error it's*, but which tag?". In Campoy, M.C., B. Belles-Fortuno and M.L. Gea-Valor (eds) 2010. *Corpus-Based Approaches to English Language Teaching*. London: Continuum.

Atserias, J. B. Casas, E. Comelles, M. González, L. Padró and Muntsa Padró. (2006). "FreeLing 1.3: Syntactic and semantic services in an open-source NLP library". *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC 2006), ELRA.Genoa, Italy. May, 2006.

Bernardini, S. (2002). "Exploring new directions for discovery learning". In B. Kettemann and G. Marko (eds.) *Teaching and Learning by Doing Corpus Analysis*. Amsterdam and New York: Rodopi, pp. 165-182.

Biber, D., S. Conrad and R. Reppen (1994). "Corpus-Based Approaches and Issues in Applied Linguistics". Applied Linguistics 15: 169-189.

Biber, D. and R. Reppen (1998). "Comparing native and learner perspectives on English grammar: A study of complement clauses". In S. Granger (ed) *Learner English on Computer*. London: Addison Wesley Longman. pp. 145-158.

Brusilovsky P. and C. Peylo (2003). "Adaptive and intelligent Web-based educational systems". In P. Brusilovsky and C. Peylo (eds.), *International Journal of Artificial Intelligence in Education 13 (2-4), Special Issue on Adaptive and Intelligent Web-based Educational Systems*, pp. 156-169.

Campoy, M.C., B. Belles-Fortuno and M. L. Gea-Valor (eds) (2010). *Corpus-Based Approaches to English Language Teaching*. London: Continuum.

Chocano, G., R. Jiménez, C. Lozano, A. Mendikoetxea, S. Murcia, M. O'Donnell, P. Rollinson, I. Teomiro (2007). "An exploration into word order in learner corpora: The WOSLAC Project". *Proceedings of the Corpus Linguistics Conference 2007*, Matthew Davies, Paul Rayson, Susan Hunston, Pernilla Danielsson (eds.), University of Birmingham, UK.

Corder, S. P. (1967). "The Significance of Learners' Errors". *International Review of Applied Linguistics* 5:4: 161-170.

Council of Europe (2001). *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Dagneaux, E., S. Denness, S. Granger, S. And F. Meunier (1996). *Error Tagging Manual Version 1.1*. Centre for English Corpus Linguistics, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

MacDonald, P. (2004). *An Analysis of Interlanguage Errors in Synchronous /Asynchronous Intercultural Communication Exchanges*. . Ph.D. Thesis. Valencia: Universitat de València.

Díez Bedmar, M.B. (2007). "From Secondary School to University: the Use of the English Article System by Spanish Learners". Paper presented at the 1st Int. Conf. on Corpus-based approaches to ELT, Castellon, Spain, November 2007.

Grabowski, E. and D. Mindt (1995). "A corpus-based learning list of irregular verbs in English". *ICAME Journal* 19: 5-22.

Granger, S. (1993). "The International Corpus of Learner English", in J. Aarts, P. de Haan, and N. Oostdijk (eds.) *English language corpora: Design, analysis and exploitation*. Amsterdam: Rodopi, 57-69.

Granger, S. (1998). "The computer learner corpus: a versatile new source of data for SLA research". In S. Granger (ed) *Learner English on Computer*. London: Addison Wesley Longman. pp. 3-18.

Granger, S. (1999). "Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus". In H. Hasselgard, S. Oksefjell (eds.), *Out of Corpora- Studies in Honour of Stig Johannson*. Amsterdam: Rodopi, pp.191-202.

James, C. (1998). Errors in Language Learning and Use: Exploring error analysis. Harlow: Longman.

Klein, D., Manning, C. (2003). "Fast Exact Inference with a Factored Model for Natural Language Parsing", *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, Cambridge, MA: MIT Press, 3-10.

Mendikoetxea, A., M. O'Donnell and P. Rollinson (2009). "WriCLE: A learner corpus for Second Language Acquisition Research", Proceedings of the 5th Corpus Linguistics Conference, Liverpool, England, 20-23 July, 2009.

Meunier, F. (2002). "The pedagogical value of native and learner corpora in EFL grammar teaching". In Granger *et al.* 2000, pp. 119-142.

Muehleisen, V. (2006). "Introducing the SILS Learners' Corpus: A Tool for Writing Curriculum Development". Waseda Global Forum, No. 3. 119-125.

O'Donnell M. (2008). "Demonstration of the UAM CorpusTool for text and image annotation". *Proceedings of the ACL-08:HLT Demo Session*, Columbus, Ohio, June 2008. Association for Computational Linguistics. 13–16.

Rankin, T. (2010). "Incorporating learner corpora data in grammar teaching". In M.C. Campoy, B. Belles-Fortuno and M.L. Gea-Valor (eds) *Corpus-Based Approaches to English Language Teaching*. London: Continuum.

Rollinson, P. and A. Mendikoetxea (2008). *The WriCLE corpus.* http://web.uam.es/woslac/Wricle/
Römer, U. (2005). Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics. Amsterdam: Benjamins.
Selinker, L. (1972). "Interlanguage". *International Review of Applied Linguistics*, 10, 209-31.
Stuart, K. (2003). "A Software Application for Text Analysis". Proceedings of the International Conference on Engineering Education, July 21–25, 2003, Valencia, Spain.
UCLES (2001). *Quick Placement Test (Paper and pencil version)*. Oxford: Oxford University Press.