Using a learner corpus to develop an EFL grammar teaching curriculum

Susana Murcia & Mick O’Donnell
Universidad Autónoma de Madrid

Funded by Ministerio de Ciencia e Innovación (FFI2009-14436/FILO)
1. Aim of the Paper

Using a learner corpus to tell us more about the learning process of Spanish learners of English:

- which grammatical structures are critical to Spanish EFL learners at each level of proficiency
- how much attention should be given to each structure.

(Parallel work done in the *English Profile* project)
Error analysis is one way to explore the grammatical competence of students at each level (e.g. Dagneaux et al 1998).

![Graph showing errors per 1000 words vs. Prof. Level]
Introduction

- **Error analysis** is one way to explore the grammatical competence of students at each level (e.g. Dagneaux et al 1998).

- However, some students make few errors, because they avoid structures they are not sure about.

- More adventurous students take risks and thus make more errors.
Error rate vs. Proficiency score

- X-axis: Proficiency
- Y-axis: Errors per 1000 words

The graph shows a downward trend indicating a decrease in error rate as proficiency increases.
Error rate vs. Proficiency score

Conservative students make fewer errors than their peers
Error rate vs. Proficiency score

Adventurous students make more errors than their peers
• We thus take a two-pronged approach:

Automatic **syntactic tagging** of corpus to see what structures students are attempting;

Manual **error analysis** to see what they do wrong.

Only both together give the full picture.
2. The TREACLE Project
2. The Project

The TREACLE Project

- Project: TREACLE
  - Teaching Resource Extraction from an Annotated Corpus of Learner English

- A cooperation between:
  Universidad Autónoma de Madrid and Universitat Politècnica de Valencia

- Funded by Ministerio de Ciencia e Innovación (FFI2009-14436/FILO)

- Runs: January 2010 – December 2012

Official Title: “Developing an annotated corpus of learner English for pedagogical application”
• Use learner English corpora to **profile** the grammatical skills of Spanish university learners at each proficiency level (A1, A2, B1, etc.)

• Use these profiles to **redesign the teaching curriculum**: determining which grammatical features need to be taught/reinforced, in what order, and with what degree of emphasis.

• Provide a **web-based language learning system** which dynamically adapts to the student.
The project uses two corpora:

- The WriCLE corpus (UAM) - *Written Corpus of Learner English*. 521 essays of ~1000 words each, written by Spanish learners of English at University level (about 500,000 words) (Rollinson and Mendikoetxea 2008)

- The UPV Learner Corpus (UPV) containing 150,000 words of shorter texts by ESP students. (Andreu et al 2010)

- Quick Oxford Placement test (UCLES, 2001) given at same time, to measure proficiency

- Other metadata: gender, academic year, degree, parent languages, time abroad, resources used in writing, etc.
3. Profiling students’ grammatical skills
Methodology

- Annotation using *UAM Corpus Tool* (O’Donnell, 2008)
  - Manual annotation of errors, based on coding scheme devised by our research team
  - Automatic annotation of syntactic structures, using Stanford Parser adapted to UAM Corpus Tool
3.2. Profiling Proficiency Levels

**Error Annotation**

- Each text annotated using a scheme which is organised hierarchically and contains 113 errors at the most delicate level.

- Errors are related to a typical grammar teaching curriculum (placing errors into the units to which they apply, e.g., NP-error includes errors in determiner usage, etc.)
3.2. Profiling Proficiency Levels

Error coding process

The legalisation of Marijuana has become a very polemic topic in our present society because it concerns a lot of different opinions and because the decision of legalizing it or not will directly affect a whole society. So, legalizing Marijuana is a social debate whose outcome couldn’t possibly please everyone. In this regard, I will try to discuss the different perspectives and opinions regarding the legalisation of soft-drugs such as Marijuana.

The general opinion of people who against the legalisation argue that if the government legalize this type of drugs the crime rates will increase and many youths will be able to obtain the drugs without any problem at all. They also believe that drugs are very dangerous substances that can cause enormous problems to the

1. Select text containing error.

2. Provide the corrected text here.

3. Assign features to current segment here.

UAM CorpusTool [http://www.wagsoft.com/CorpusTool](http://www.wagsoft.com/CorpusTool) (Free)
Error Coding progress

Words Coded

Errors Coded

234 texts coded
4. Applicability of the results

**Error Annotation: Global Results**

- By examining the types of errors made by students, we can determine how much teaching time to spend on each area.
4. Applicability of the results

Error Annotation: Results by proficiency

- By examining the types of errors made at each proficiency level, we can adapt teaching to each group’s needs.
## Most common Lexical Errors

<table>
<thead>
<tr>
<th></th>
<th>Spelling errors</th>
<th>Transfer errors</th>
<th>Wordchoice errors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spelling errors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inmigration</td>
<td>76 8.00%</td>
<td>actually 10 3.44%</td>
<td>persons 43 3.17%</td>
</tr>
<tr>
<td>inmigrants</td>
<td>64 6.74%</td>
<td>optative 5 1.72%</td>
<td>other 23 1.80%</td>
</tr>
<tr>
<td>live</td>
<td>20 2.11%</td>
<td>inmigrants 5 1.72%</td>
<td>work 17 1.25%</td>
</tr>
<tr>
<td>inmigrant</td>
<td>15 1.58%</td>
<td>supposes 5 1.72%</td>
<td>works 17 1.25%</td>
</tr>
<tr>
<td>religión</td>
<td>14 1.47%</td>
<td>fomenting 5 1.72%</td>
<td>do 13 0.96%</td>
</tr>
<tr>
<td>ilegal</td>
<td>11 1.16%</td>
<td>course 4 1.37%</td>
<td>make 13 0.96%</td>
</tr>
<tr>
<td>whit</td>
<td>11 1.16%</td>
<td>cannon 4 1.37%</td>
<td>economical 12 0.88%</td>
</tr>
<tr>
<td>wich</td>
<td>10 1.05%</td>
<td>important 4 1.37%</td>
<td>win 11 0.81%</td>
</tr>
<tr>
<td>gobernment</td>
<td>9 0.95%</td>
<td>sanity 3 1.03%</td>
<td>job 10 0.74%</td>
</tr>
<tr>
<td>lifes</td>
<td>9 0.95%</td>
<td>asignature 3 1.03%</td>
<td>have 9 0.66%</td>
</tr>
<tr>
<td>an</td>
<td>9 0.95%</td>
<td>poblacion 3 1.03%</td>
<td>take 8 0.59%</td>
</tr>
</tbody>
</table>
4. Applicability of the results

**Error Annotation: Results for Grammar**

- For all students, more attention needed on NPs and PPs!
- As students progress, more attention needed on clause structure issues.
3.1. Profiling Proficiency Levels

**Syntactic Analysis**

- UAM CorpusTool produces *automatic syntactic analysis* of the sentences in the text (embeds Stanford parser)
- We can then explore what grammatical structures each student uses in their essays.
- We can explore how often grammatical structures are used at each proficiency level.
- We can thus construct “grammatical profiles”: the degree to which each proficiency level uses each kind of structure.
- From these we can see when it is best to teach particular structures.
The new points system for driving offences will be established in Spain before summer.

With this new system, the driving licence will consist of a number of points that can be deducted for various violations.

I personally agree with the establishment of this new law, as I feel that it will contribute to a safer road environment.
<table>
<thead>
<tr>
<th>TENSE</th>
<th>FINITENESS</th>
<th>VERB-TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple-present</td>
<td>simple-finite</td>
<td>intranstive-verb</td>
</tr>
<tr>
<td>present-perfect</td>
<td>finite-with-connector</td>
<td>monotransitive-verb</td>
</tr>
<tr>
<td>present-progressive</td>
<td>relative-clause</td>
<td>ditransitive-verb</td>
</tr>
<tr>
<td>simple-past</td>
<td>that-clause</td>
<td>ergative-verb</td>
</tr>
<tr>
<td>past-progressive</td>
<td>wh-nominal-clause</td>
<td>relational-verb</td>
</tr>
<tr>
<td>past-progressive</td>
<td>infinitive-clause</td>
<td>verbal-verb</td>
</tr>
<tr>
<td>simple-modal</td>
<td>pres-participle-clause</td>
<td>mental-verb</td>
</tr>
<tr>
<td>modal-perfect</td>
<td>past-participle-clause</td>
<td></td>
</tr>
<tr>
<td>modal-progressive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MODALITY</th>
<th>DO-INSERTION</th>
<th>POLARITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>nonmodal-clause</td>
<td>do-inserted</td>
<td>positive-polarity</td>
</tr>
<tr>
<td>true-modal-clause</td>
<td>no-do-inserted</td>
<td>negative-polarity</td>
</tr>
<tr>
<td>future-clause</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PROCESS TYPE</th>
<th>VOICE</th>
<th>MOOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>material-clause</td>
<td>active-clause</td>
<td>declarative-clause</td>
</tr>
<tr>
<td>verbal-clause</td>
<td>passive-clause</td>
<td>imperative-clause</td>
</tr>
<tr>
<td>mental-clause</td>
<td></td>
<td>interrogative-clause</td>
</tr>
<tr>
<td>relational-clause</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Syntactic Analysis: Extracting Profiles

3.1. Profiling Proficiency Levels

After parsing:

- 30,000 sentences
- 100,000 clauses
- 175,000 NPs
- 700,000 words

- .... But what do we do with it all?
3.1. Profiling Proficiency Levels

**Syntactic Analysis: Extracting Profiles**

**Simple Frequency Approach**

- Some researchers contrast the learner’s degree of usage of a syntactic feature with the degree of usage of natives.
- Where students under-use the feature, more emphasis is needed in teaching.
- Over-usage also needs to be corrected (perhaps by teaching alternative lexico-grammatical strategies, or teaching appropriate contexts of use).

![Graph showing Use of Passive by Proficiency levels A2, B1, B2, C1, C2]
3.1. Profiling Proficiency Levels

Syntactic Analysis: Extracting Profiles

‘Onset of Use’ approach

• Our belief is that a first concern should be with whether a leaner is capable of producing a structure at all.

• We thus look at each text individually, to see if the structure is present or not.

• We then measure the percentage of texts which do not use the feature at all:

![Graph showing the percentage of texts which do not use passive voice by proficiency level.]

Texts which don’t use passive (%)
‘Onset of Use’ approach: another example
Use of Present-participle clauses: “He likes going to the zoo”
3.1. Profiling Proficiency Levels

Syntactic Analysis: Extracting Profiles

‘Onset of Use’ approach: another example

- Use of Past-participle clauses:
  - The man driven by hunger
  - Burnt by the sun, he marched on

% of Texts with no past participle clauses

Obviously, this structure is acquired much later, and thus should be taught later.
3.1. Profiling Proficiency Levels

Syntactic Analysis: Extracting Profiles

- By analysing the degree of non-usage of each grammatical feature at each proficiency level, we can determine when the feature is most critical to the group as a whole.
  - **When** the early adopters have started to use it
  - **Before** the cautious have started to use it
  - Exactly where in this range a structure is best taught needs to be decided.
  - Some flexibility good, to fit into a structured grammar teaching environment.
• So, far, only applied to a range of clause structures

• We need to explore the full range of structures taught in grammar courses (e.g., noun phrases, cohesion, reference, etc.)

• Also need to merge results from error analysis with the syntactic results.
5. Conclusions
5. Conclusions

- Our two pronged approach gives a full picture of what students need depending on the proficiency
6. Future Directions

- Syntactic features and error tags currently not directly relatable
- We need to provide a means of relating them
- List of **“1,000 concepts a learner needs to learn in order to use a language like a native”**, e.g.

  237 “much” cannot be used in positive statements
  - ❌ *I have much water*

  238 “much” can be used in negative statements
  - ✅ *I don’t have much money*

  239 “much” can be used in questions etc.
  - ✅ *Do you have much money?*
5. Future Directions

- The set of concepts present in each sentence can be recognised and assigned to the student as (perhaps partially) acquired.
- Error tags can be re-expressed as failures to comply with one of these concepts.
  - E.g., “The drugs are a problem in the society”
  - As error: `determiner-present-not-required`
  - As concept breached: `abstract-noun-does-not-take-determiner`
- Given a student text, syntactically parsed and error tagged, we can derive a student model, set of English concepts acquired or not.

Student Writing Correction System

Update estimates of student’s acquired concepts

Update estimates of concept difficulty

Tailored quiz for the student’s current needs

Recommended Reading List generated

Online Reference Website

Adaptive Quiz System

Student Models

Language Model
Thank you for your attention!

- **Treacle Web page:**
  http://www.uam.es/treacle

- **UAM CorpusTool (Free) Macosx, Windows**
  http://www.wagsoft.com/CorpusTool