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

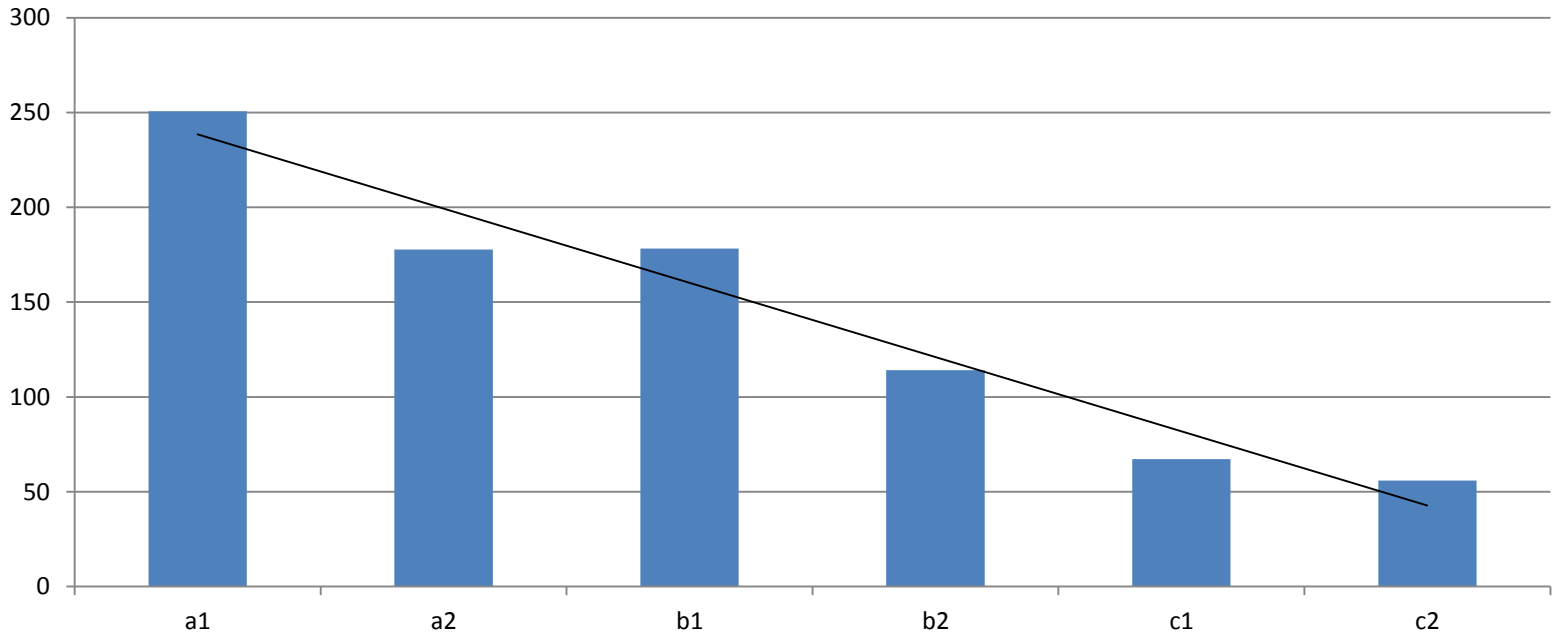
Using a learner corpus to tell us more about the learning process of Spanish learners of English:

- 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)

Introduction

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



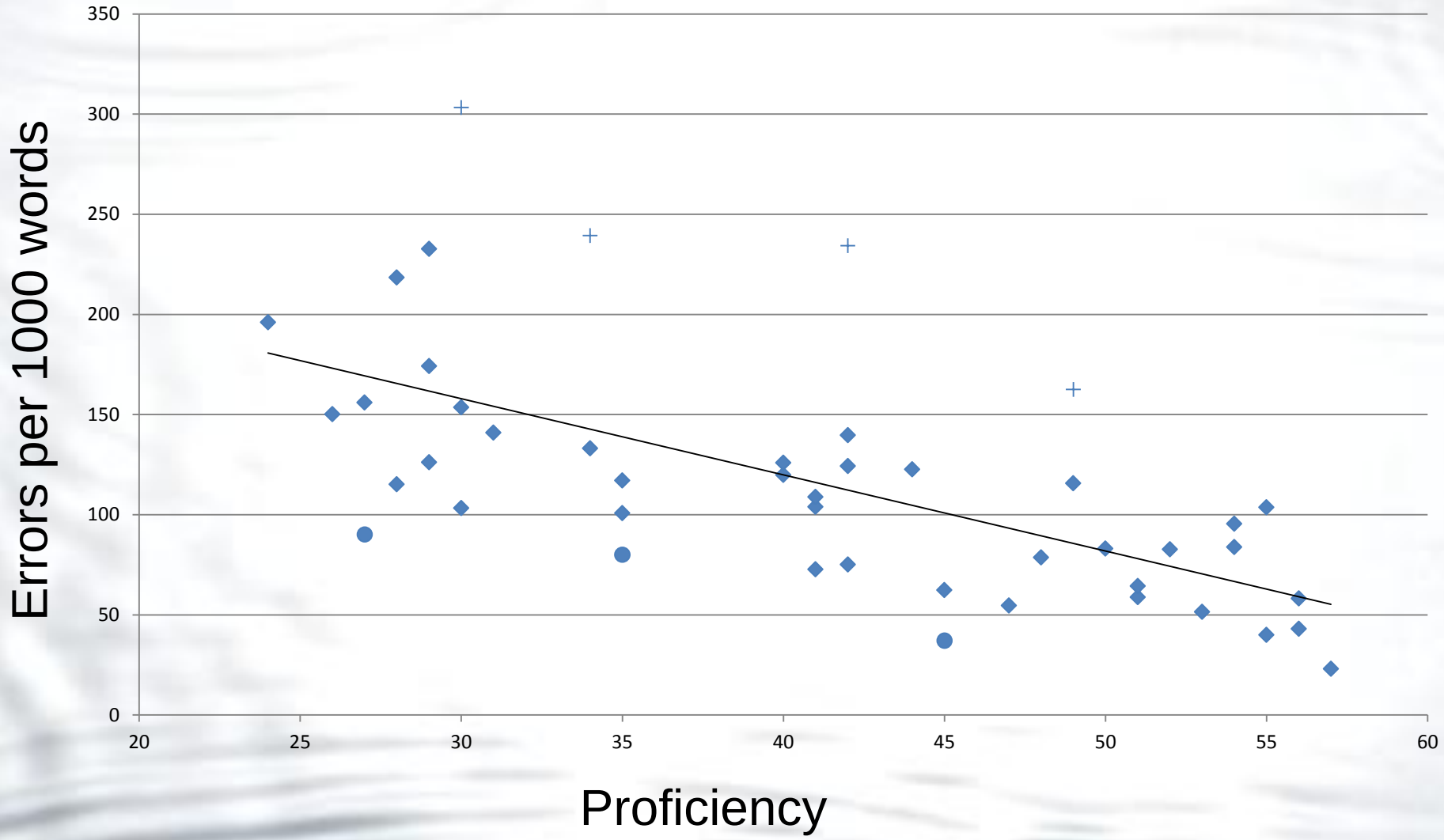
Errors per 1000 words vs. Prof. Level

Introduction

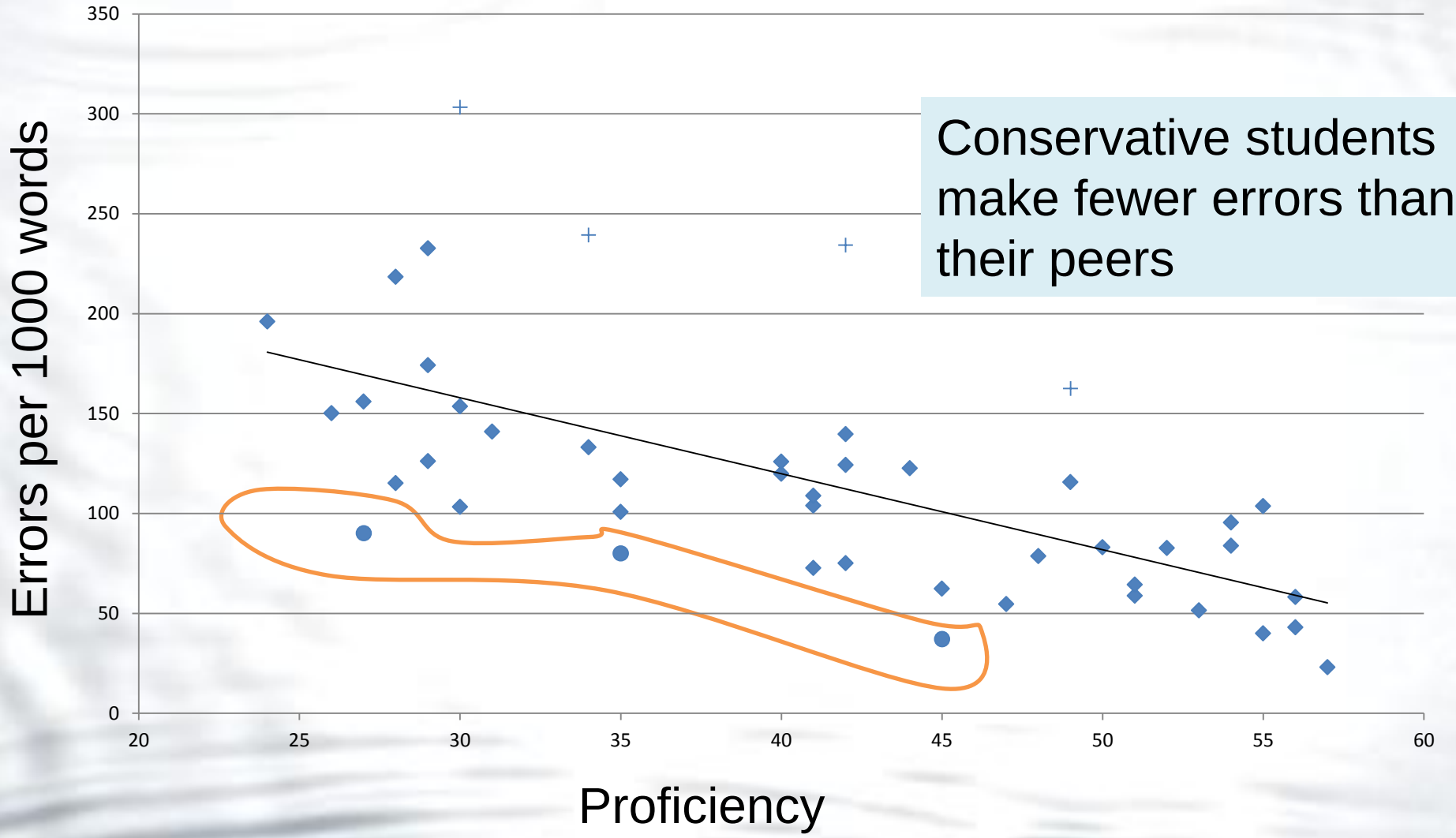
- Error analysis is one way to explore the grammatical competence of students at each level (e.g. Dagneaux et al 1998).

- 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.

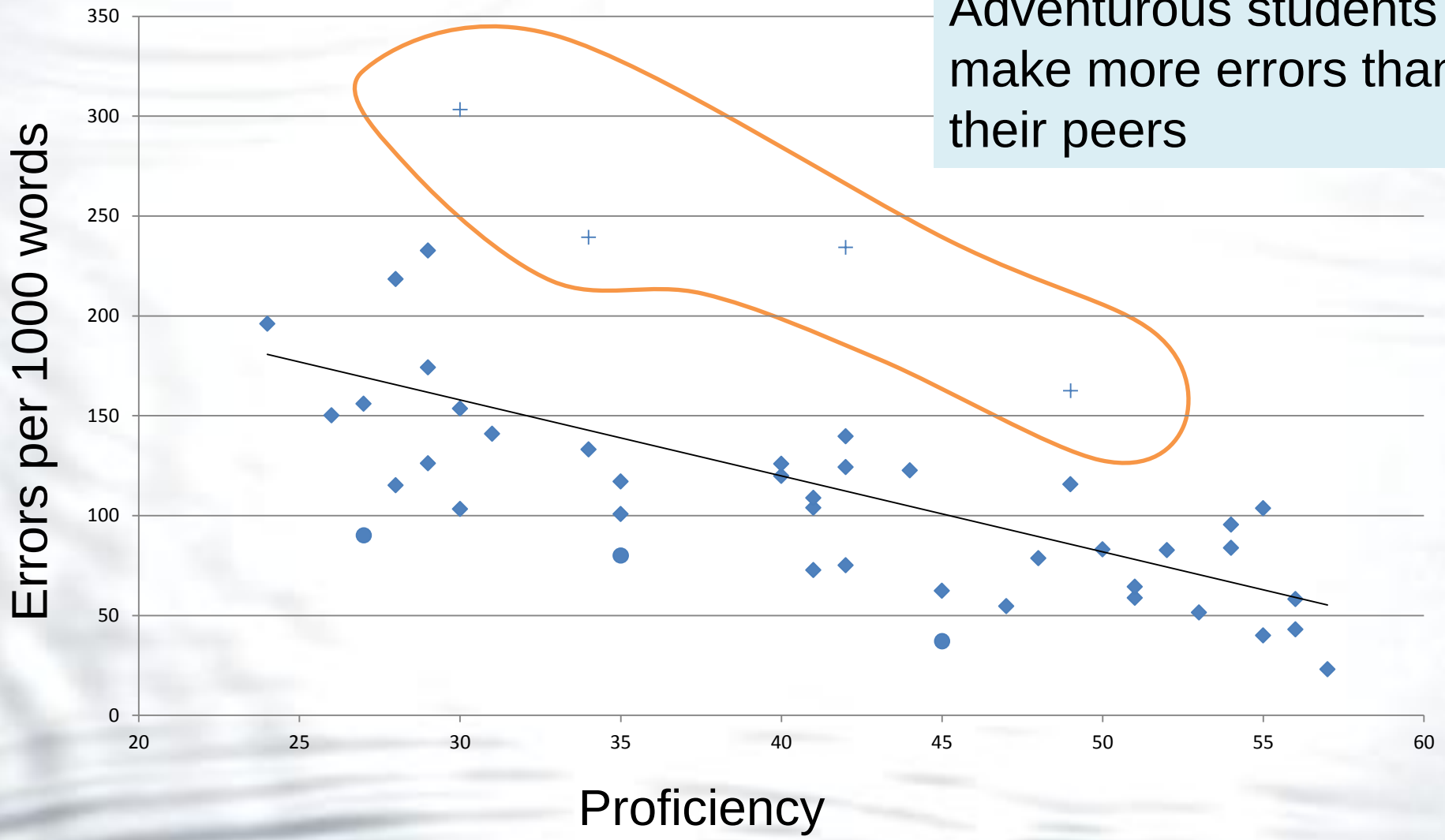
Error rate vs. Proficiency score



Error rate vs. Proficiency score

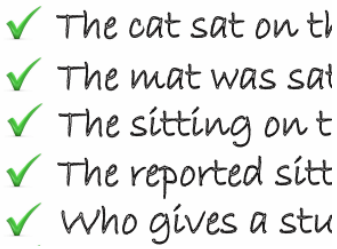


Error rate vs. Proficiency score

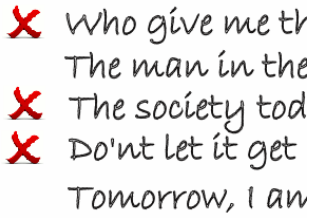


Introduction

- 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

The TREACLE Project

- Project: TREACLE

Teaching
Resource
Extraction from an
Annotated
Corpus of
Learner
English

Official Title: "Developing an annotated corpus of learner English for pedagogical application"

- A cooperation between:
  - Universidad Autónoma de Madrid and
  - Universitat Politécnica de Valencia
- Funded by Ministerio de Ciencia e Innovación (FFI2009-14436/FILO)
- Runs: January 2010 – December 2012

Goals of the project

- 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 Corpora

- 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.

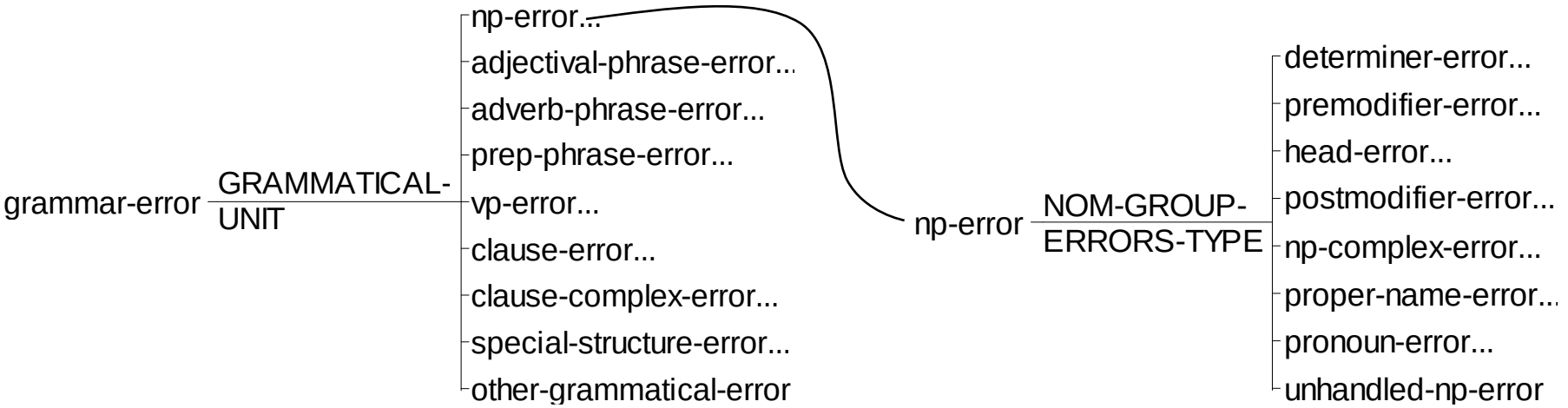
3. Profiling students' grammatical skills

Methodology

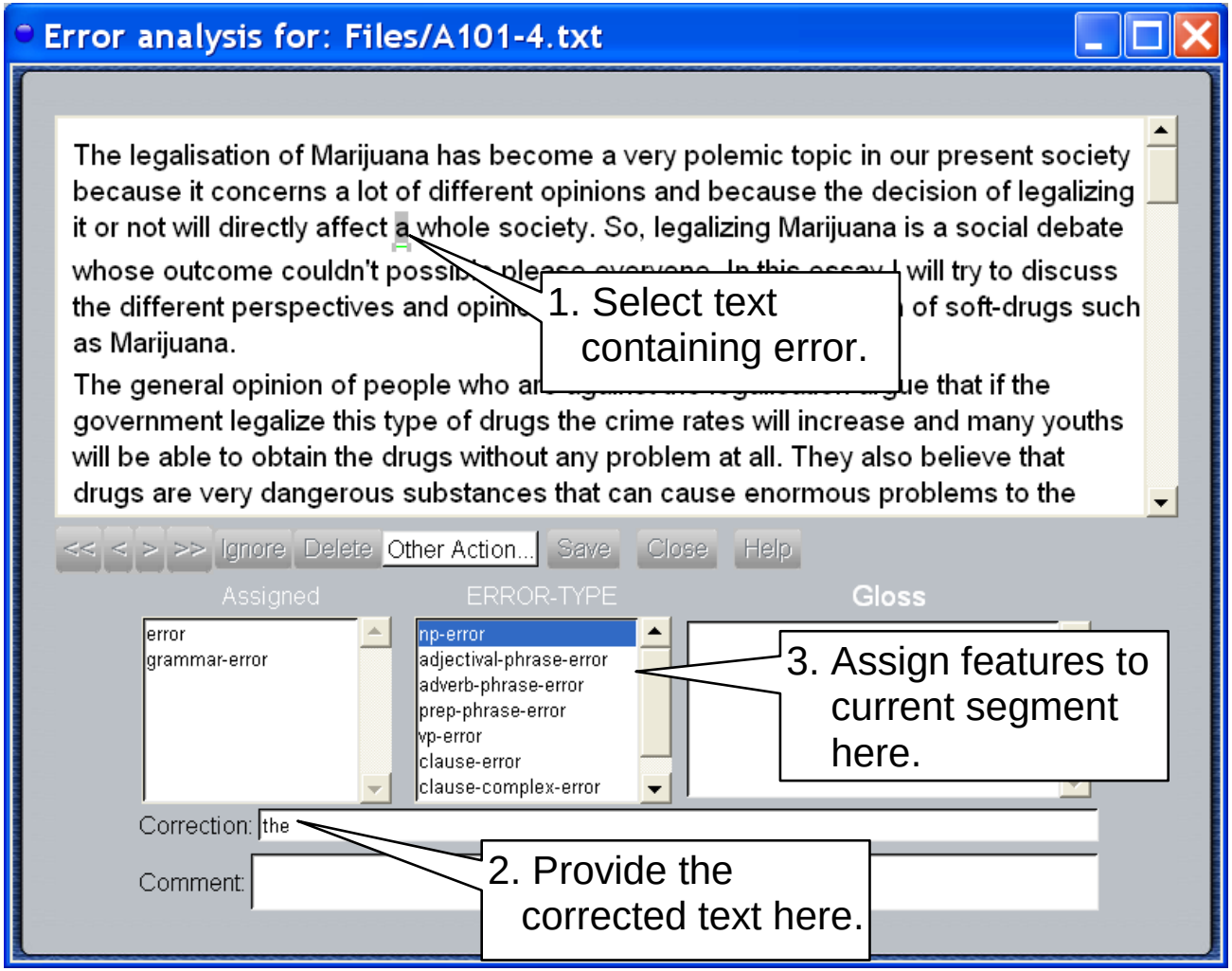
- Annotation using *UAM Corpus Tool* (O'Donnell, 2008)
  - Manual annotation of errors, based on coding scheme devised by our research team
  - Automatic annotation of syntactic structures, using Stanford Parser adapted to UAM Corpus Tool

Error Annotation

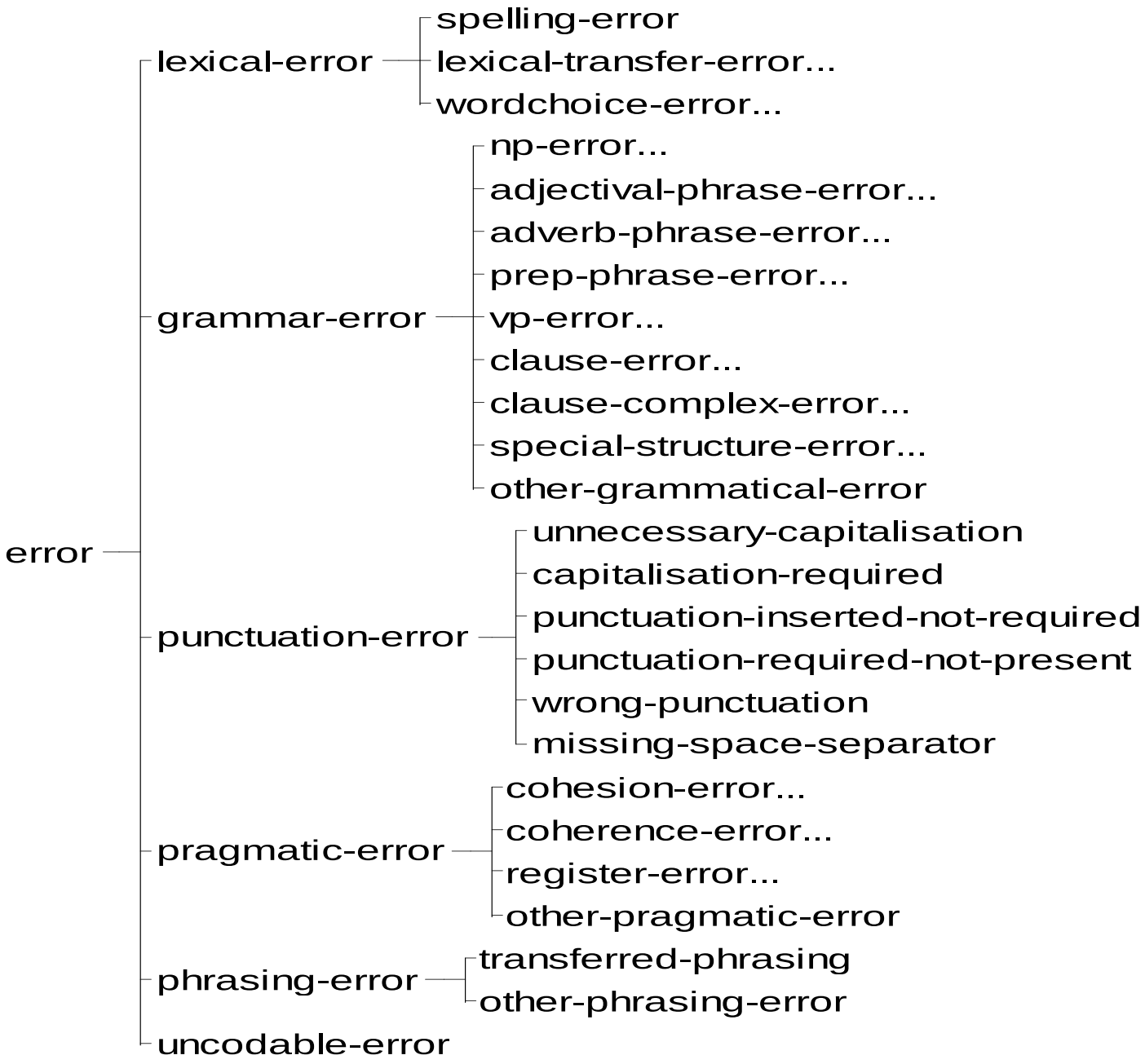
- Each text annotated using a scheme which is organised hierarchically and contains 113 errors at the most delicate level
- Errors are related to a typical grammar teaching curriculum (placing errors into the units to which they apply, e.g., NP-error includes errors in determiner usage, etc.)



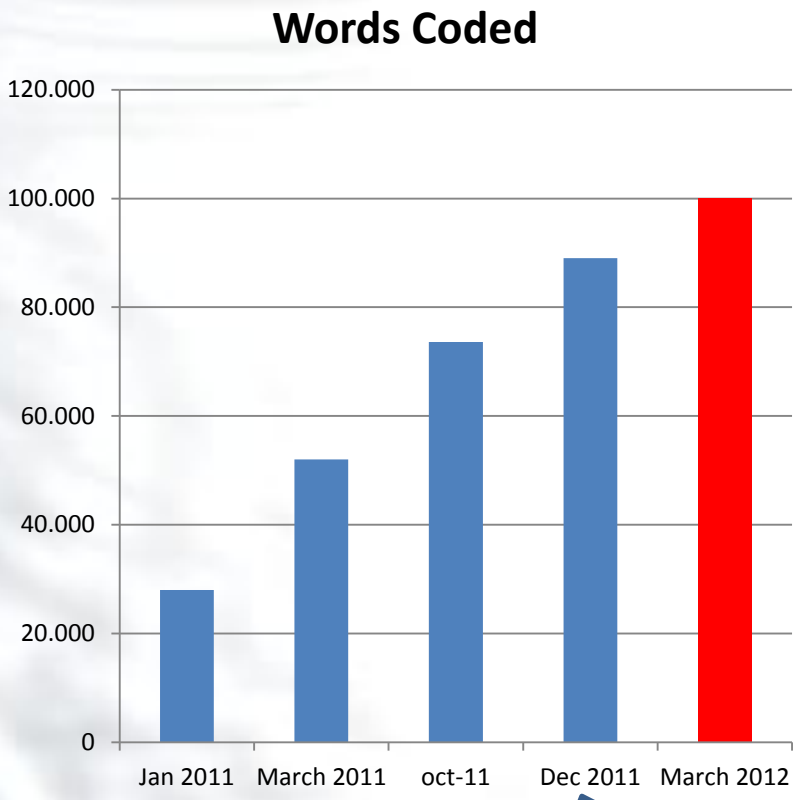
Error coding process

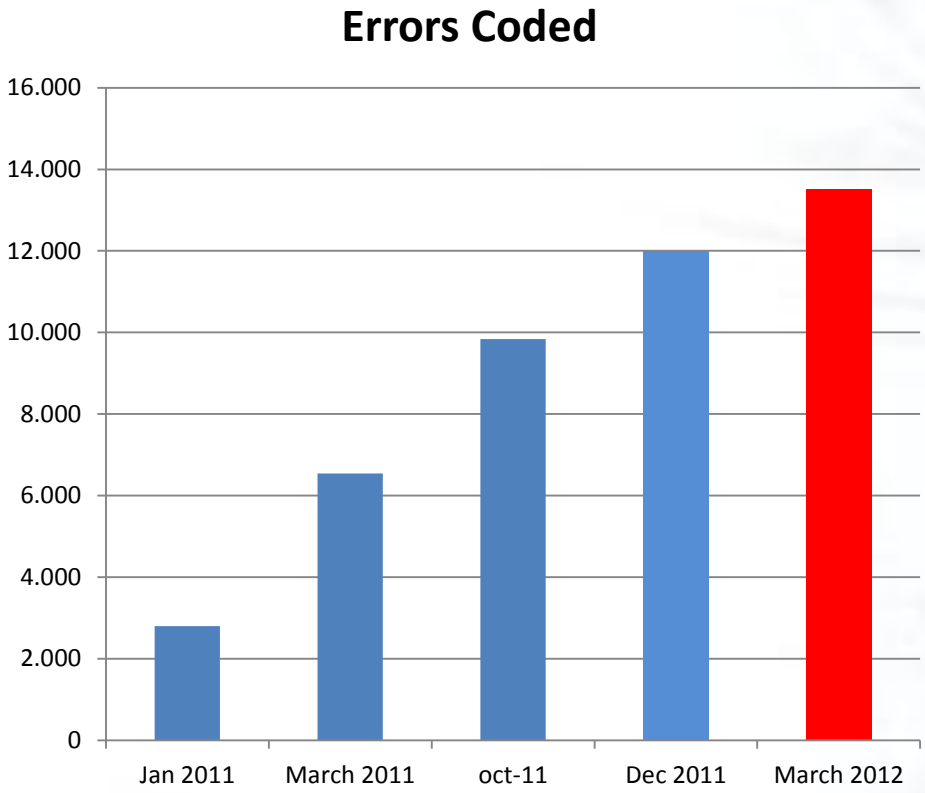


UAM CorpusTool http://www.wagsoft.com/CorpusTool (Free)


error
lexical-error
spelling-error
lexical-transfer-error...
wordchoice-error...
grammar-error
np-error...
adjectival-phrase-error...
adverb-phrase-error...
prep-phrase-error...
vp-error...
clause-error...
clause-complex-error...
special-structure-error...
other-grammatical-error
punctuation-error
unnecessary-capitalisation
capitalisation-required
punctuation-inserted-not-required
punctuation-required-not-present
wrong-punctuation
missing-space-separator
pragmatic-error
cohesion-error...
coherence-error...
register-error...
other-pragmatic-error
phrasing-error
transferred-phrasing
other-phrasing-error
uncodable-error

Error Coding progress

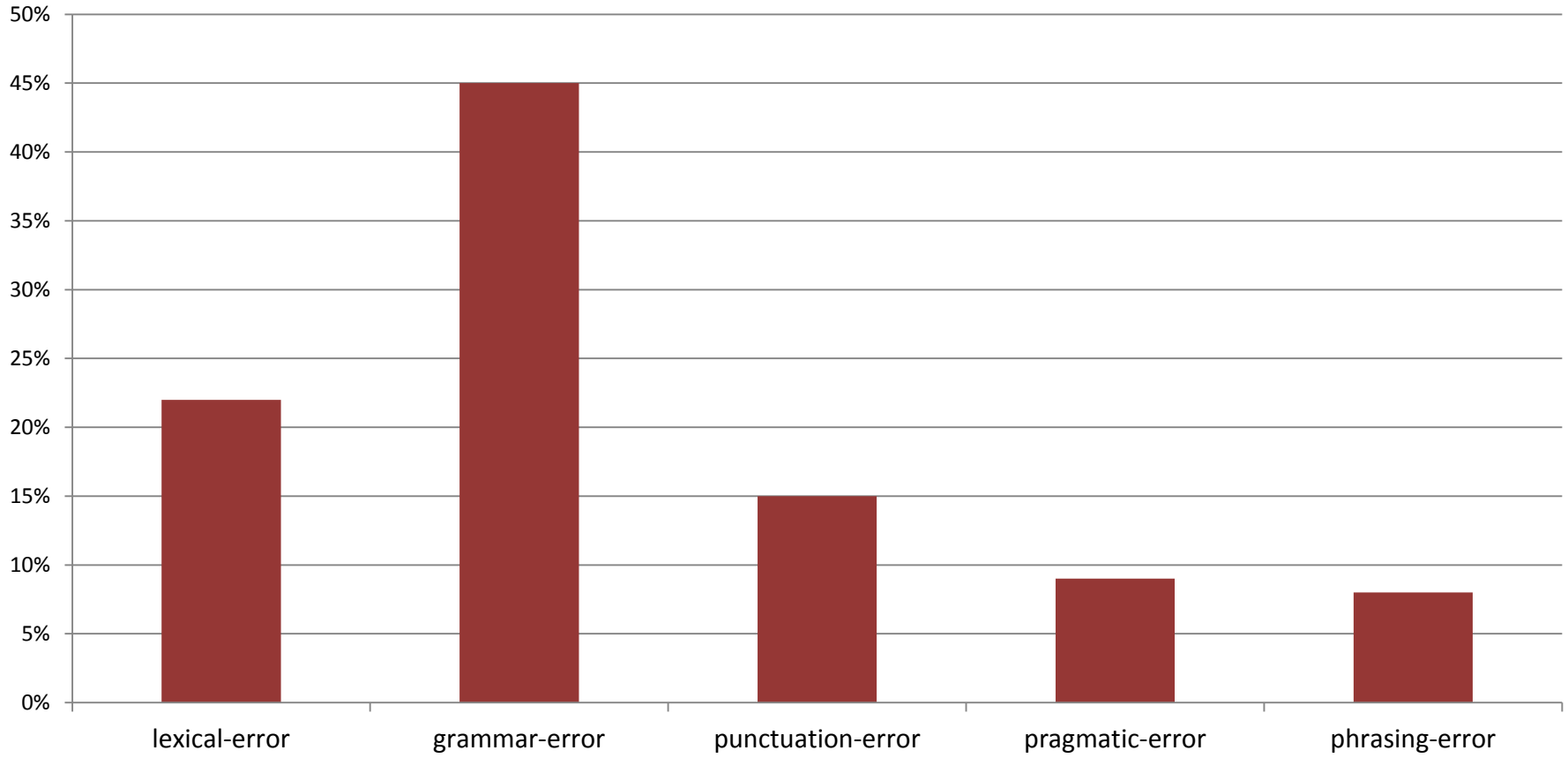




234 texts coded

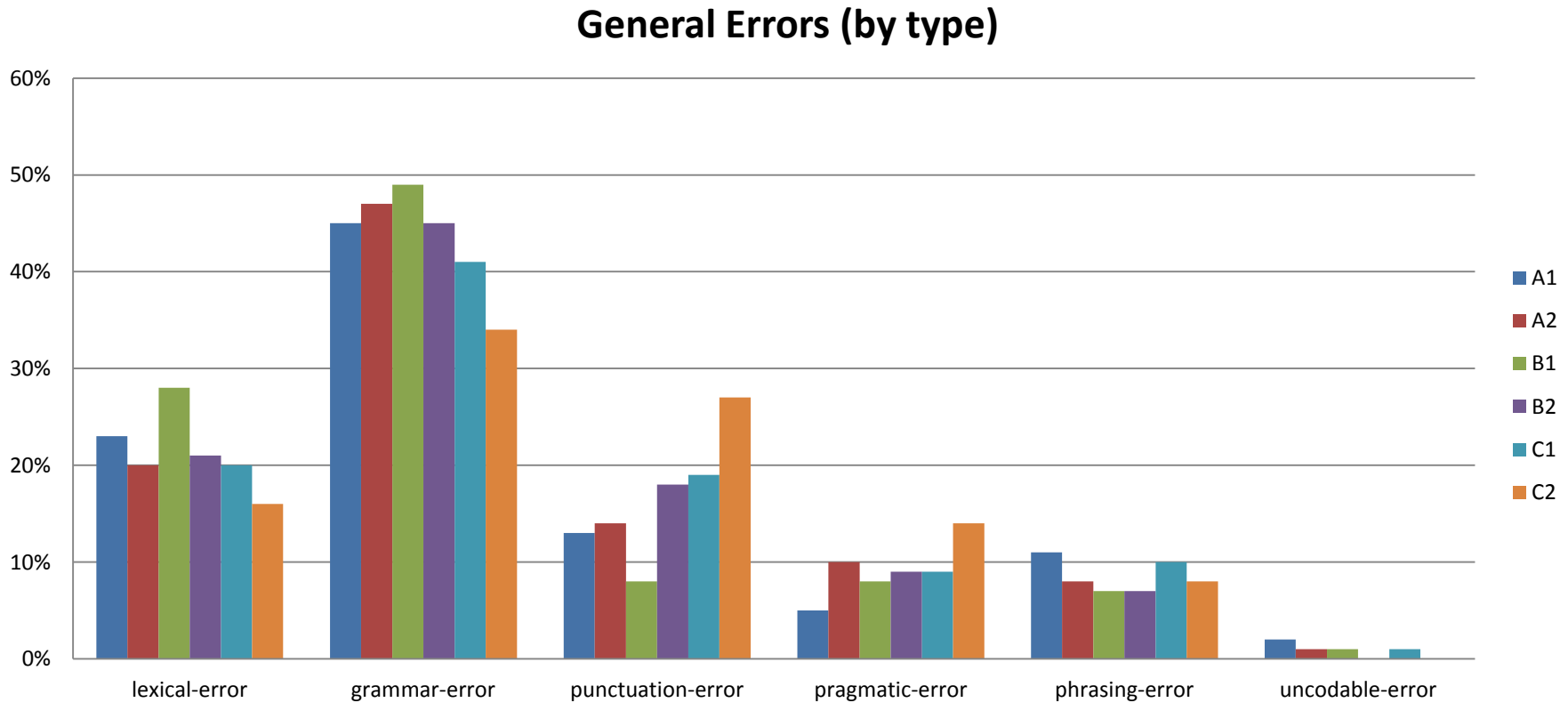
Error Annotation: Global Results

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



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

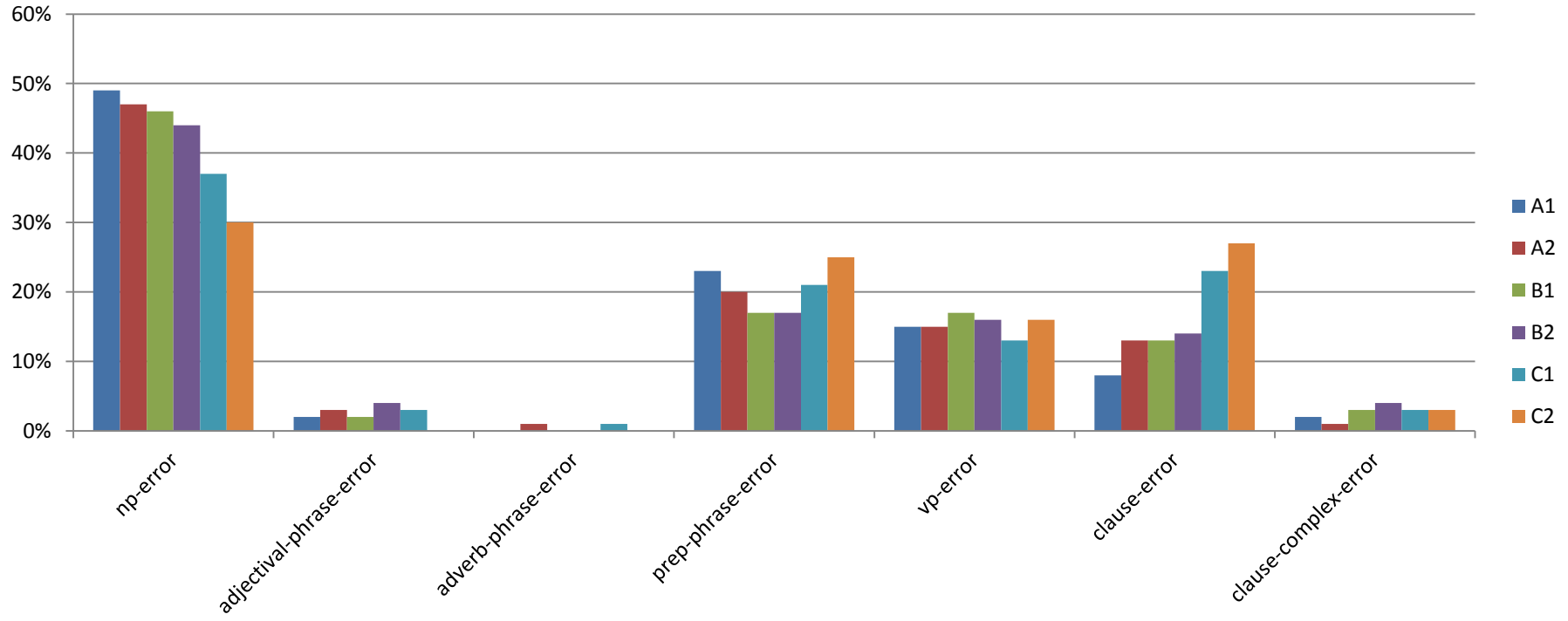
Spelling errors		
inmigration	76	8.00%
inmigrants	64	6.74%
live	20	2.11%
inmigrant	15	1.58%
religión	14	1.47%
ilegal	11	1.16%
whit	11	1.16%
wich	10	1.05%
gobernment	9	0.95%
lifes	9	0.95%
an	9	0.95%

Transfer errors		
actually	10	3.44%
optative	5	1.72%
inmigrants	5	1.72%
supposes	5	1.72%
fomenting	5	1.72%
course	4	1.37%
cannon	4	1.37%
important	4	1.37%
sanity	3	1.03%
asignature	3	1.03%
poblation	3	1.03%

Wordchoice errors		
persons	43	3.17%
other	23	1.80%
work	17	1.25%
works	17	1.25%
do	13	0.96%
make	13	0.96%
economical	12	0.88%
win	11	0.81%
job	10	0.74%
have	9	0.66%
take	8	0.59%

Error Annotation: Results for Grammar

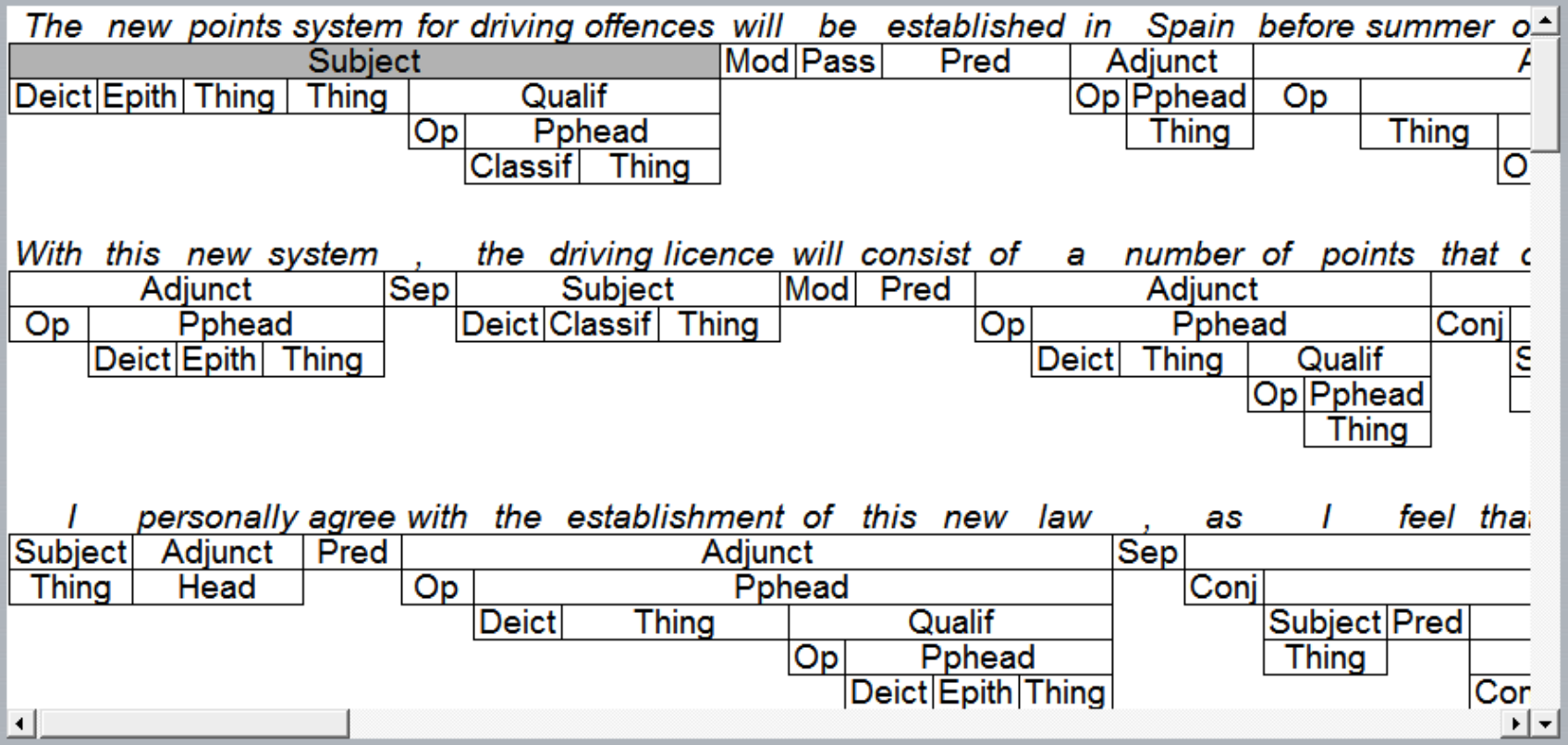
- For all students, more attention needed on NPs and PPs!
- As students progress, more attention needed on clause structure issues.



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.

Grammar analysis for: Files/A101-2.txt



<< < > >> Ignore Delete Other Action... Save Close Help

Assigned

- grammatical-unit
- group
- np
- common-phrase
- singular-phrase
- nonwh-noun-phrase

Gloss

Comment:

TENSE simple-present present-perfect present-progressive simple-past past-progressive past-progressive simple-modal modal-perfect modal-progressive	FINITENESS simple-finite finite-with-connector relative-clause that-clause wh-nominal-clause infinitive-clause pres-participle-clause past-participle-clause	VERB-TYPE intranstive-verb monotransitive-verb ditransitive-verb ergative-verb relational-verb verbal-verb mental-verb
MODALITY nonmodal-clause true-modal-clause future-clause	DO-INSERTION do-inserted no-do-inserted	POLARITY positive-polarity negative-polarity
PROCESS TYPE material-clause verbal-clause mental-clause relational-clause	VOICE active-clause passive-clause	MOOD declarative-clause imperative-clause interrogative-clause

Syntactic Analysis: Extracting Profiles

After parsing:

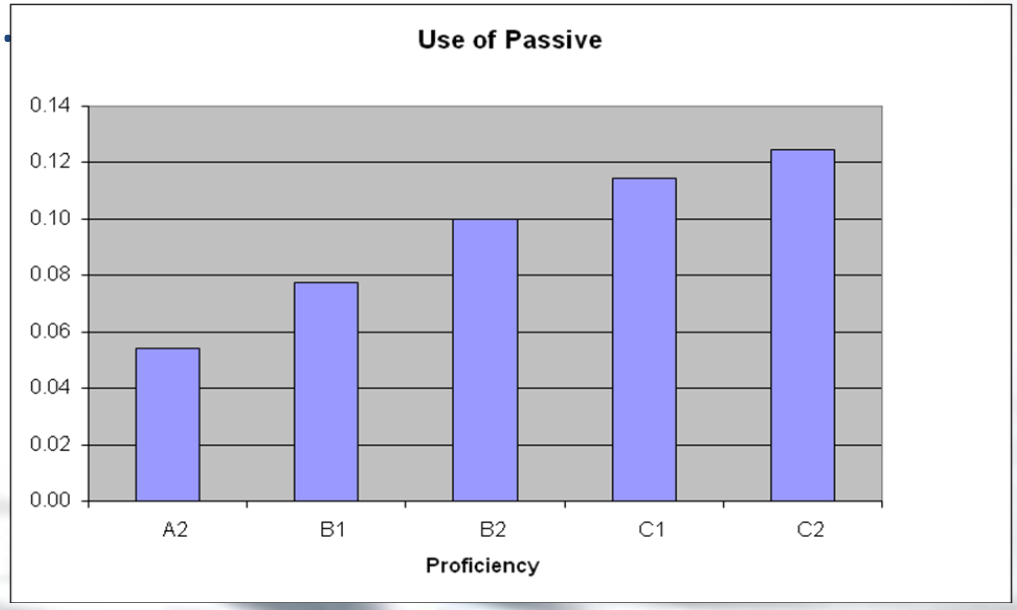
- **30,000 sentences**
- **100,000 clauses**
- **175,000 NPs**
- **700,000 words**

- **…. But what do we do with it all?**

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).

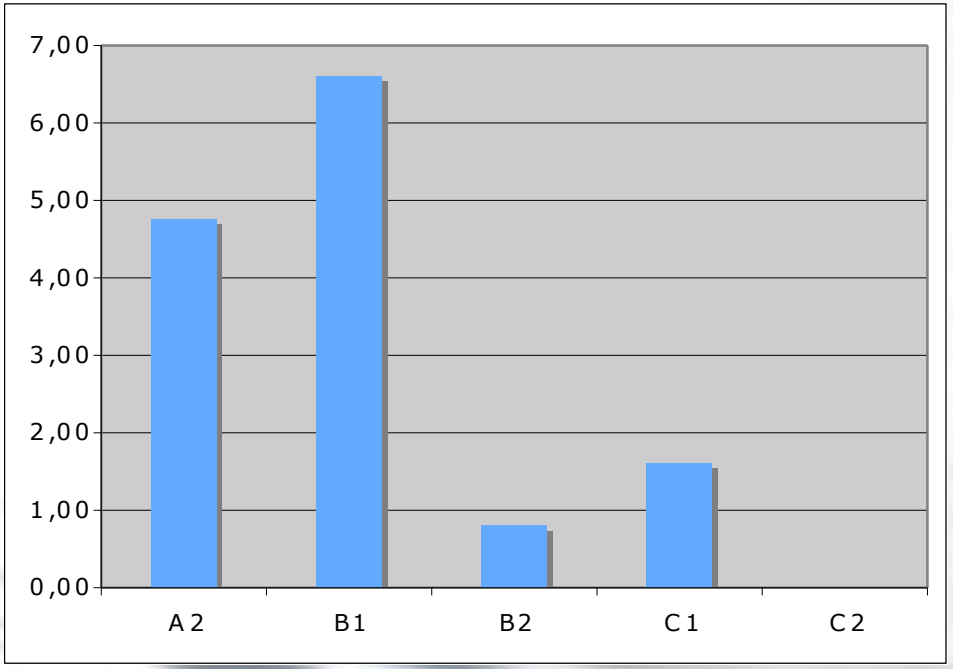


Syntactic Analysis: Extracting Profiles

'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**:

Texts which don't use passive (%)

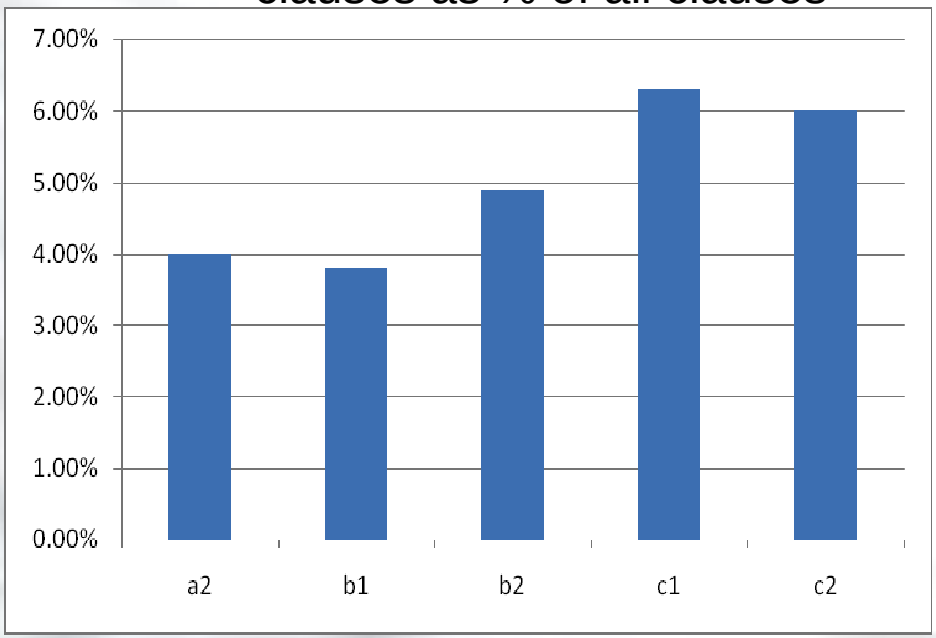


Syntactic Analysis: Extracting Profiles

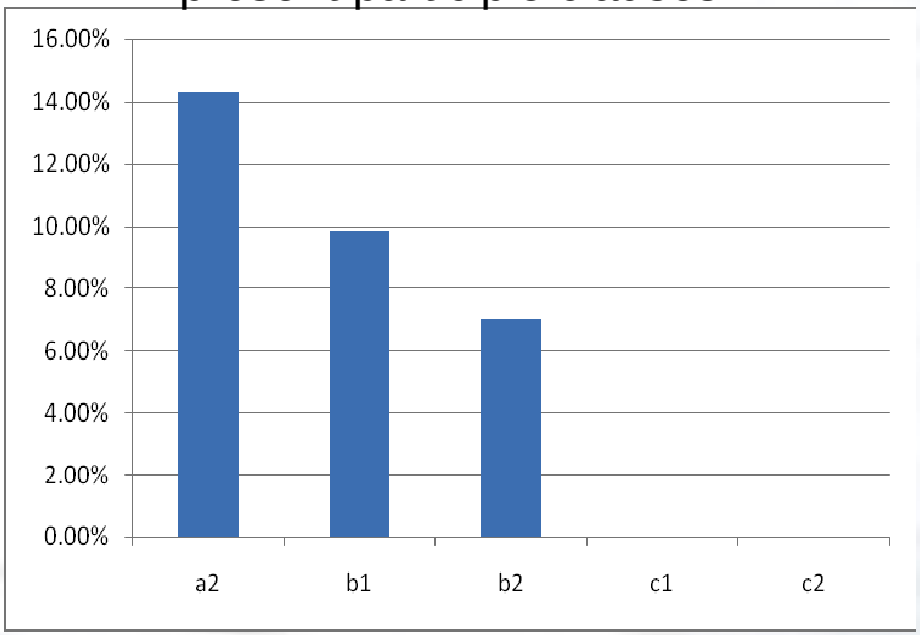
'Onset of Use' approach: another example

Use of **Present-participle clauses**: *"He likes going to the zoo"*

Present participle clauses as % of all clauses



% of Texts with no present participle clauses

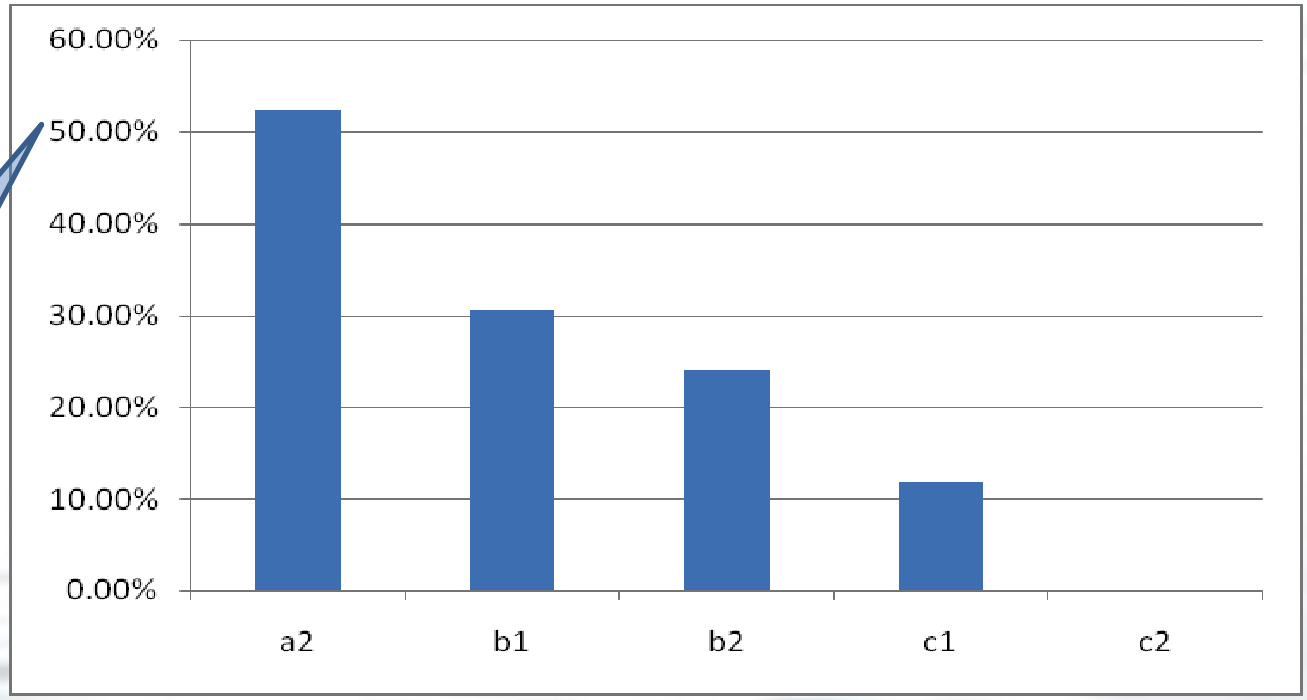


Syntactic Analysis: Extracting Profiles

'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.

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

Syntactic Analysis: Extracting Profiles

- 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.

Recommended Reading List generated

Online Reference Website

Student Writing Correction System

Update estimates of student's acquired concepts

Student Models

Update estimates of concept difficulty

Language Model

Tailored quiz for the student's current needs

Adaptive Quiz System

Thank you for your attention!

- Treacle Web page:

  http://www.uam.es/treacle

- UAM CorpusTool (Free) Macosx, Windows

  http://www.wagsoft.com/CorpusTool