A photograph of a classroom where several students are seated at desks, focused on their work. They are surrounded by books, notebooks, and stationery. In the foreground, a yellow tin labeled 'BAHANAB' is filled with pens and pencils. A large book titled 'OXFORD ENGLISH' is visible on the desk. The background shows a bookshelf and a poster on the wall.

From Learner Corpora to Curriculum Design: an empirical approach to staging the teaching of grammatical concepts

Mick O'Donnell
Universidad Autónoma
de Madrid

Road map

1. Goal: Redesign grammar education with learner corpora
2. Deciding **what** to teach
3. Deciding **when** to teach
4. Conclusions



1. Orientation

The TREACLE Project

- Project: TREACLE



Teaching
Resource
Extraction from an
Anotated
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 Politècnica de Valencia
- Funded by the Spanish Ministerio de Ciencia e Innovación (FFI2009-14436/FILO)
- Runs: January 2010 – June 2013 (but we are applying for a new project)

Goal of our research

- We are studying the linguistic production of our learners to gain insight into:
 - **what they need to learn** (mainly in terms of grammar and vocabulary)
 - **In what order** they need this material.
- Goal is to use these insights to **change the way we teach** English grammar over the 4 years of our English degree



Discovering what learners need

- **Learner corpora** can tell us a lot about what our students need to learn:
 - Manual **Error Analysis** to show what structures or vocabulary they are currently struggling with
 - Automatic **Syntactic Analysis** to reveal what students are actually attempting (and not attempting)

Treacle Corpora

- The project uses two learner corpora:
 - ☞ **WriCLE** corpus: 500,000 words (521 essays) collected by Paul Rollinson at UAM (1st year and 3rd year of English Studies)
 - ☞ **UPV Learner Corpus** 150,000 words of shorter texts by ESP students at Universidad Politecnica de Valencia.

Proficiency level of each writer measured by giving **Oxford Quick Placement Test** at same time.

Corpus: Annotation

Our Corpus is:

- Syntactically parsed (based on Stanford parser):

- 700,000 words
- 1,330 texts
- 30,000 sentences

- Error Coded:

- 300 student essays
- 110,000 words
- 16,000 errors

Grammar analysis for: Files/A101-2.txt

<i>The new points system for driving offences will be e</i>										
Subject								Mod	Pass	
Deict	Epith	Thing	Thing		Qualif					
				Op	Pphead					
					Classif	Thing				

<i>With this new system , the driving licence will co</i>										
Adjunct					Sep	Subject			Mod	P
Op	Pphead					Deict	Classif	Thing		
	Deict	Epith	Thing							

<i>I personally agree with the establishment of th</i>									
Subject		Adjunct		Pred		Adjunct			
Thing	Head				Op	Pphead			
						Deict	Thing		



2. Deciding What to teach

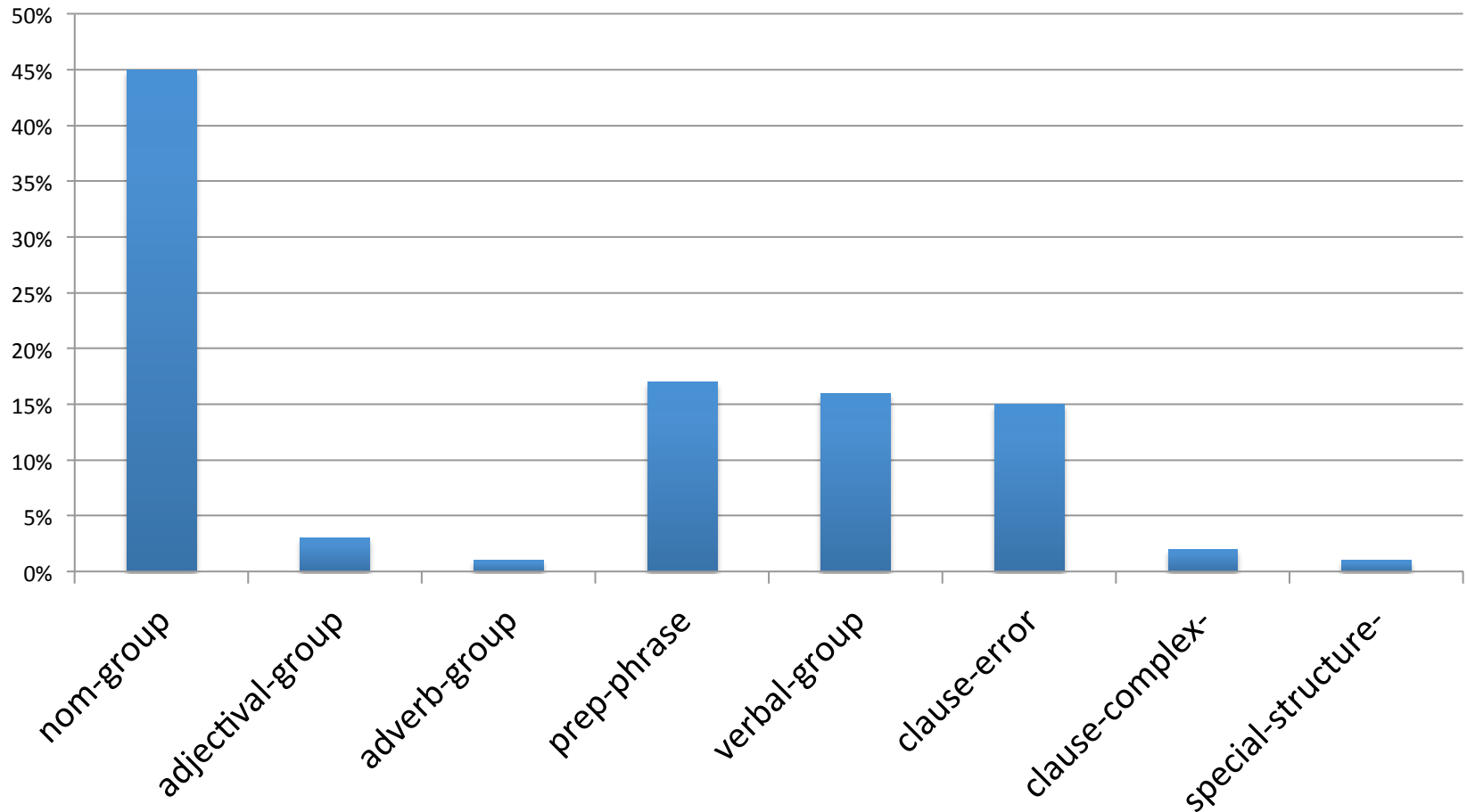
2. Deciding What to teach

When deciding **what to teach**, Learner corpora researchers:

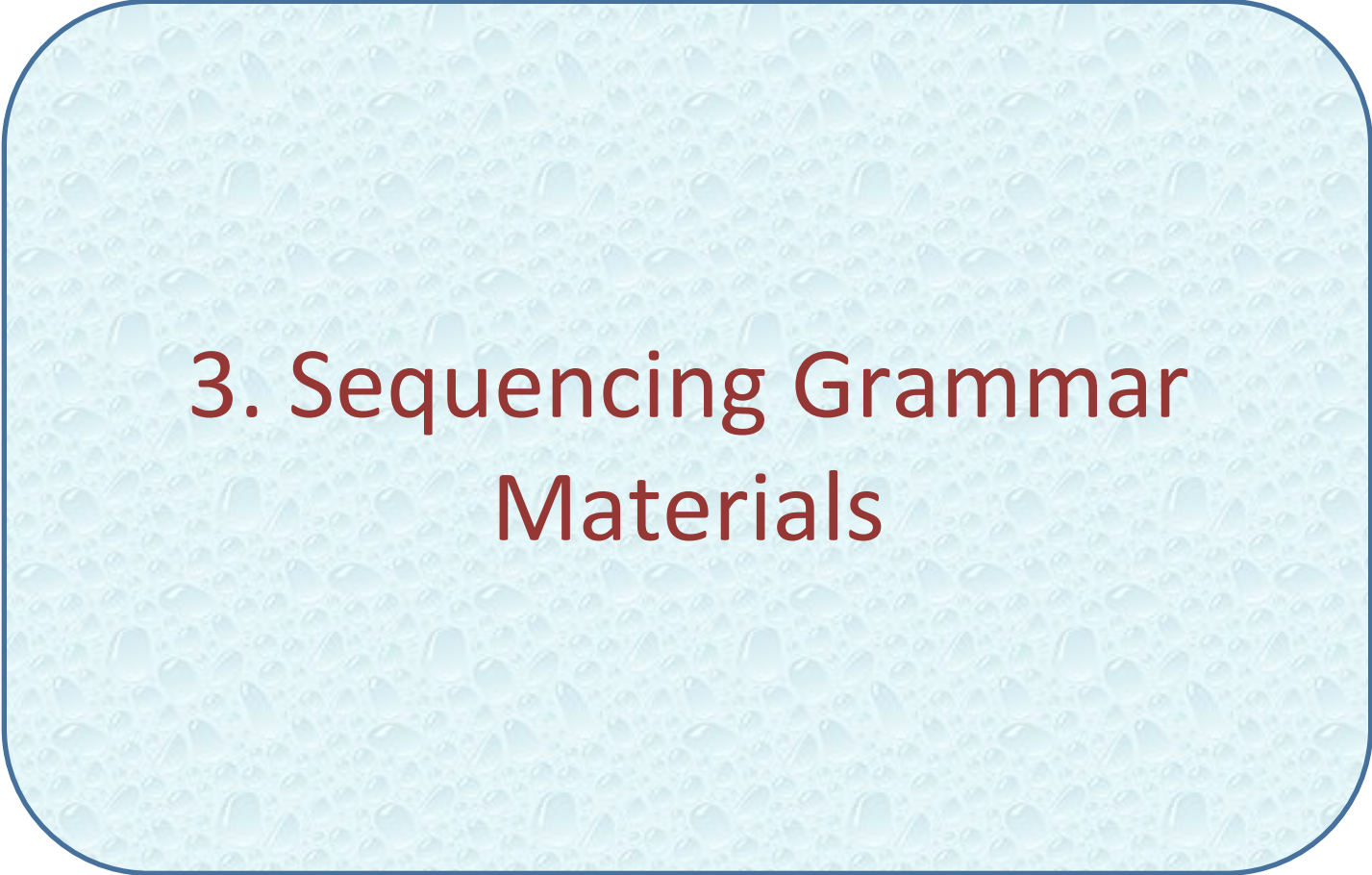
- Compare **level of usage** of vocabulary or syntactic structures to native writers.
 - Where learners under-use, more attention needed on this item.
 - Where learners over-use, teaching of alternative structures recommended
E.g., if modal-auxiliaries used more by learners than natives, teach adverbial and adjectival alternatives.
- Explore **errors** made by a group of errors to identify phenomena that need more work.

2. Deciding What to teach: Using error data

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



Transfer errors			Intralingual Errors	
Borrowing	Coinage	Transferred spelling	Spelling	Wordchoice
carril-bici	determinated	inmigration	live	persons
laboral	optative	inmigrant	whit	work
España	fomenting	ilegal	wich	be
ONGs	course	religi3n	an (and)	other
Europa	sanity	gobovernment	the	do
temporal	poblation	posibilities	a	make
mas	form	cicles	lifes	economical
hachis	displacements	adiction	countrys	win
mundial	asignature	tipes	life foreing	have
conducta	desesperation	opini3n	becouse	get
infantil	diary	politic	there	job
habituate	principately	costums	beleive	undeveloped
receptor	evollution	asociation		doing



3. Sequencing Grammar Materials

3. Sequencing Grammatical Concepts

- To sequence teaching of grammatical concepts:
 - We need some way to relate each instance of writing to the proficiency of the writer.
 - Ideally, each text in the corpus should have metadata indicating the proficiency level of the writer.

3.1 Sources of Evidence of Proficiency

Means of assessing grammatical proficiency:

1. Proficiency Exams (e.g., First Certificate):

- Test whether a learner is proficient at the designated level.
- Just written component generally used.
- Can use “pass vs. fail” or raw scores.
- Evidence of grammatical sequencing by taking results from a number of exam levels
- E.g., English Profile have data from several levels.

3.1 Sources of Evidence of Proficiency

Means of assessing grammatical proficiency:

1. Proficiency Exams (e.g., First Certificate):

- **Problem 1:** Scores represent many areas of language ability apart from grammar, e.g., overall structure, clarity of argument, etc.
- **Problem 2:** scores for distinct level exams cannot be compared.

3.1 Sources of Evidence of Proficiency

Means of assessing grammatical proficiency:

2. Score in a Placement test (e.g., Oxford Placement Test):

- Provides a single score for all learners
- Scores can be divided into CEFR levels
 - e.g., Oxford Placement Test 135-149 -> B2
- Can test just grammatical proficiency.

3.1 Sources of Evidence of Proficiency

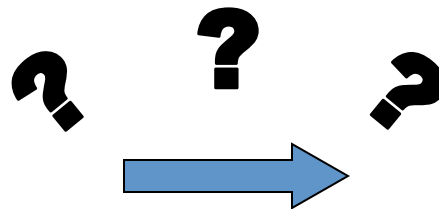
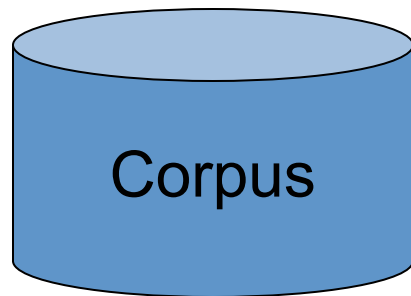
In the Treacle corpora:

- all learners took the **Oxford Short Placement Test** within the month of writing.
- Only Grammatical proficiency tested.
- Proficiency score from 0-60.
- CEFR levels estimated from these scores

3.2 Using the corpora to sequence concepts

So, we have an learner corpus with lots of annotations, and proficiency scores, but...

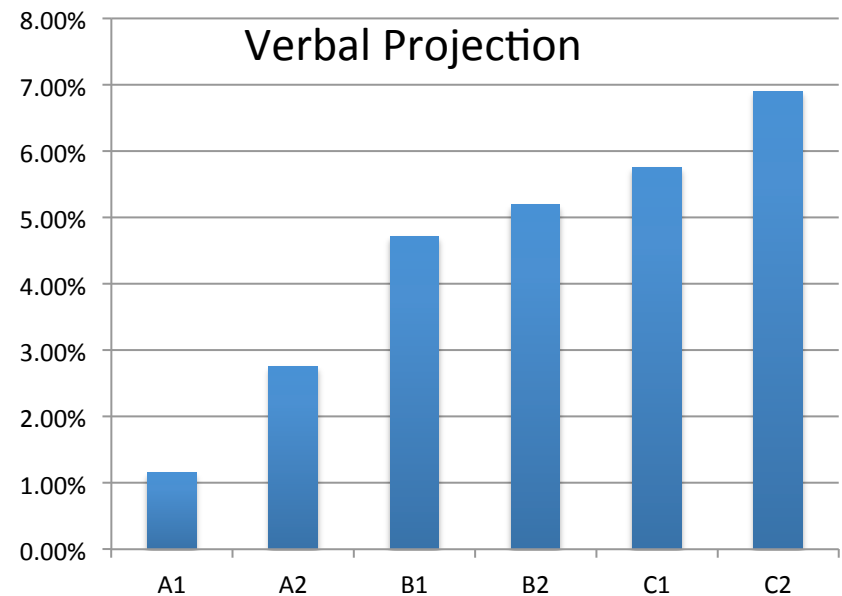
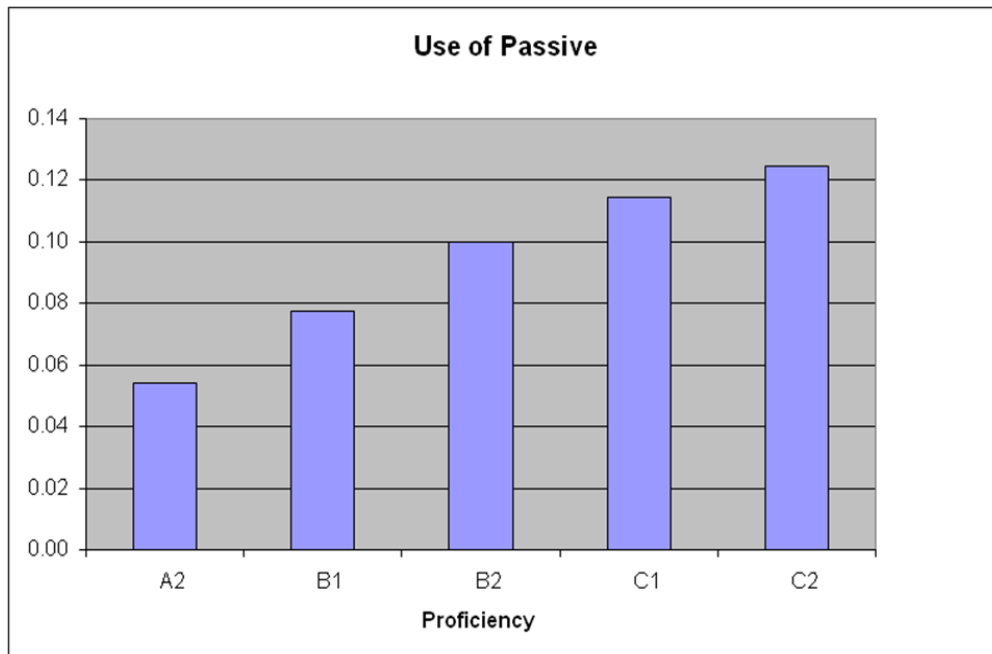
How do we use the corpus to inform us as to **how to sequence grammatical concepts?**



3.2 Using the corpora to sequence concepts (i)

Levels of usage

- Levels of usage at different are not too useful:
 - Where in the increasing use of a feature does one draw the line and say: this is where this should be taught!



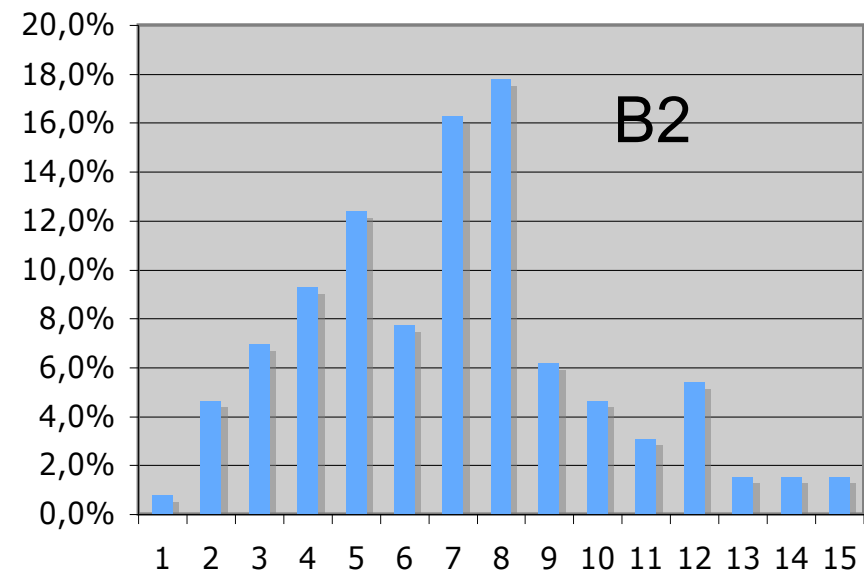
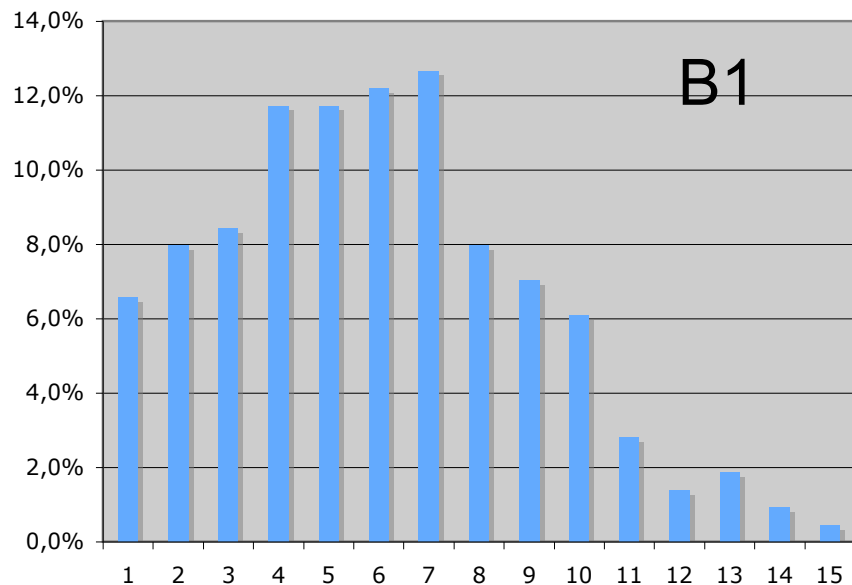
3.2 Using the corpora to sequence concepts (i)

Levels of usage: variation within a level

- Even within a level, students are not the same.
- These graphs show that, for learners in the same proficiency band, the degree of use of a syntactic feature can vary widely.

X-axis: Percent of clauses in the text which are Passive

Y-axis: Percent of learners in the band with that usage level

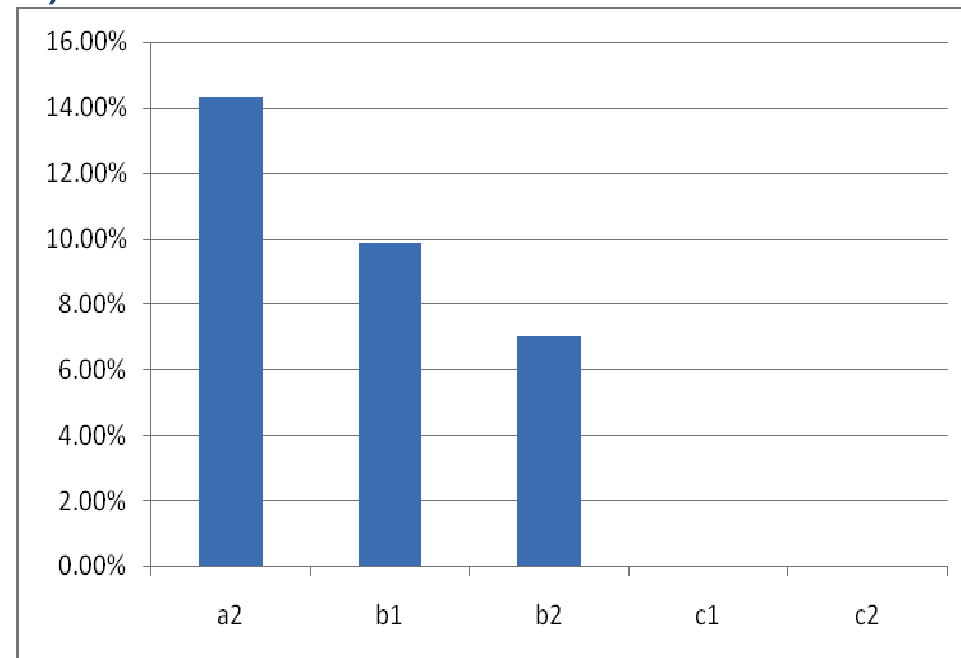
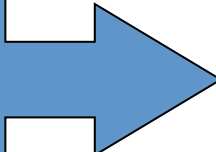


3.2 Using the corpora to sequence concepts (ii)

Onset of Use

- Better to ask whether a learner 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 use the feature **at all** (at each level)

Texts which
don't use
present participle
clauses(%)



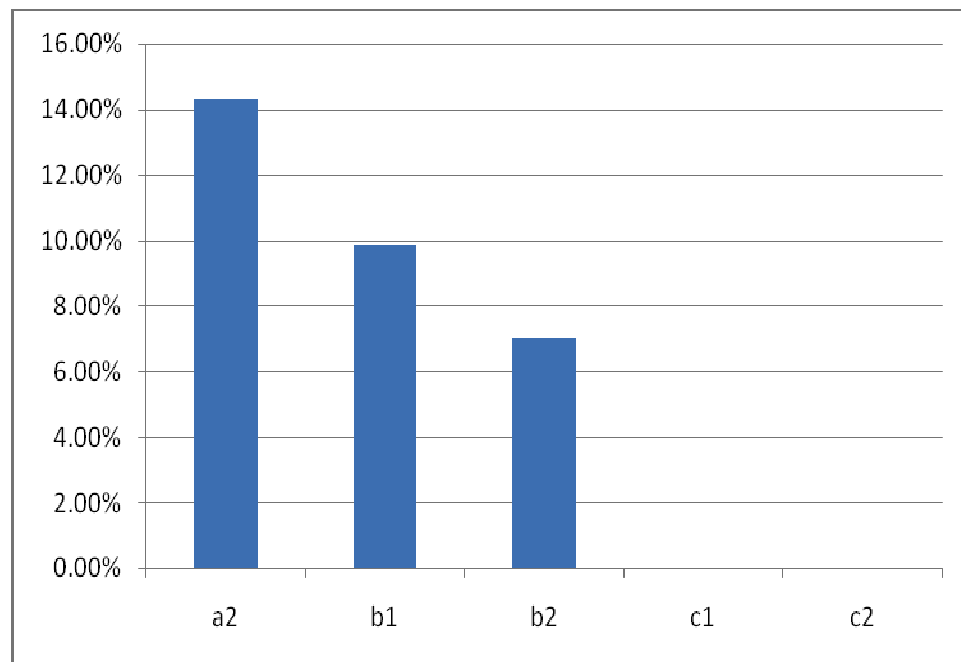
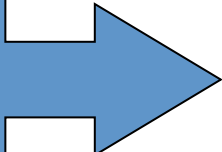
3.2 Using the corpora to sequence concepts (ii)

Onset of Use

We could say:

- start teaching a structure when early adopters are experimenting with it (e.g., 10% of learners)
- More conservative students have not yet started to use it.

Texts which
don't use
present participle
clauses(%)



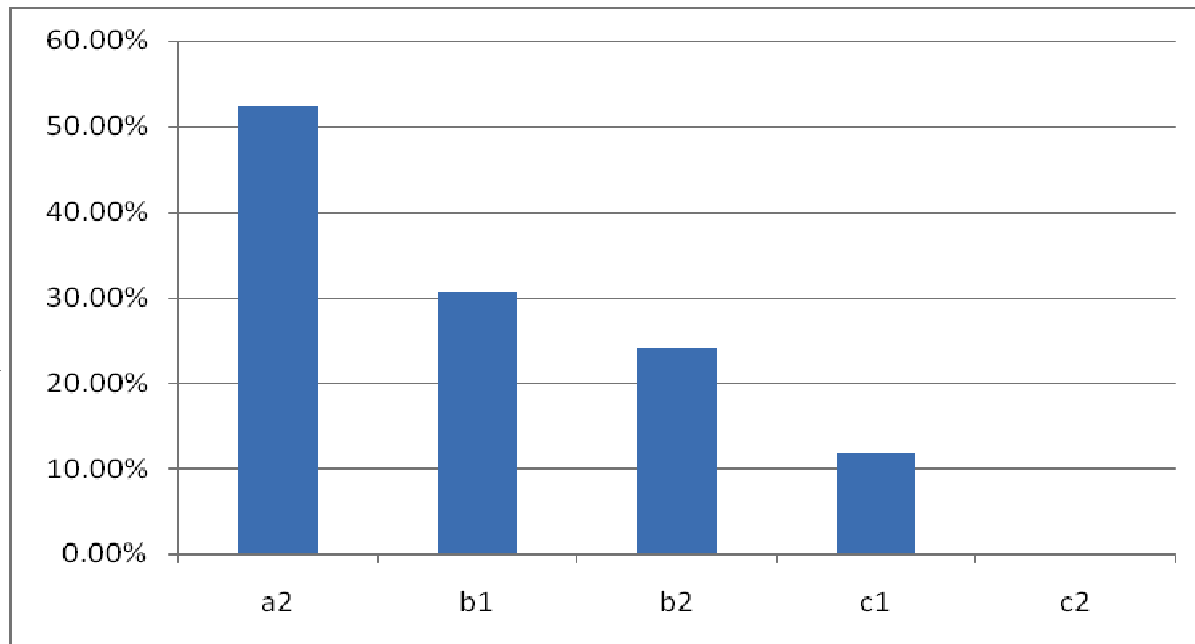
3.2 Using the corpora to sequence concepts (ii)

Onset of Use: problem

Problem: Each text needs to be long enough so that we should statistically expect occurrence of the structure:

- Can thus only be used for more common structures, passive, relative-clauses, etc.
- No use for occurrence of more marked structures (Clefts, etc.)

Texts which
don't use
past participle
clauses (%)



3.2 Using the corpora to sequence concepts (iii)

Criterial Features approach

Hawkins et al claim to be able to identify ‘criterial features’ of each proficiency level:

“certain linguistic properties that are characteristic and indicative of L2 proficiency at each level”

3.2 Using the corpora to sequence concepts (iii)

Criteria Features approach

Hawkins and Buttery: “Positive linguistic properties are correct properties of English that are acquired at a certain L2 level and that generally persist at all higher levels”

BUT:

- In actual practice, things are blurred.
- Acquisition does not happen suddenly between levels.
- Rather, in each successive L2 level, a higher number of learners exhibit the feature.
- The question remains: how many learners need to exhibit the feature to say that the feature is criterial for that level?

3.2 Using the corpora to sequence concepts (iii)

Criterion Features approach

Example of problem

Hawkins and Buttery: “For example, new verb co-occurrences that appear at B1, such as the ‘ditransitive’ NP-V-NP-NP structure (*she asked him his name*), are criterial for [B1, B2, C1, C2]; “

In our corpus: we have instances of this at A2 level, e.g.,
“...since the mother give him the opportunity to live”
“the actual system gives children a common education...”
“...make this law an enemy of every smoker...”

3.2 Using the corpora to sequence concepts (iii)

Criterial Features approach

- Similarly with errors (“negative linguistic properties”)
- Errors of a given type do not magically disappear at a given level.
- The incidence of the error falls with increasing proficiency, but the disappearance is gradual.

3.2 Using the corpora to sequence concepts (iii)

Order of Difficulty

- I believe that grammatical features do not inherently belong to particular levels of proficiency
- Rather, each linguistic feature can be acquired at different points of the learning process, depending on the individual experiences of the learner.
- However, grammatical concepts do exhibit more or less difficulty for learners of a given L1.
- Consequently, grammatical concepts will be –
when viewed over a population of learners –
acquired in a particular order.

3.2 Using the corpora to sequence concepts (iii)

Order of Difficulty

- We thus use our learner corpus to chart the relative difficulty of grammatical concepts.
- We do not try to fix these concepts to proficiency levels.
- Rather, we just produce an ordering of grammatical concepts in relation to each other.

3.2 Using the corpora to sequence concepts (iii)

Order of Difficulty

Deriving Order of Difficulty of **Errors**

1. For each occurrence of the error, collect the proficiency score of the essay.
2. Average these scores, to give a difficulty index for the error

3.2 Using the corpora to sequence concepts (iii)

Order of Difficulty

Deriving Order of Difficulty of **Errors**

1. For each occurrence of the error, collect the proficiency score of the essay.
 2. Average these scores, to give a difficulty index for the error.
- This gives us a number between 1 and 60 (for our corpus, between 29 and 49).
 - The more often an error is made by a lower proficiency learner, the lower this score will be.
 - If the error bothers advanced learners, then the difficulty score will be higher.

3.2 Using the corpora to sequence concepts (iii)

Order of Difficulty

Deriving Order of Difficulty of syntactic Structures

1. For each occurrence of the structure through the corpus, collect the proficiency score of the essay.
2. Average these scores, to give a difficulty index for the error-type.

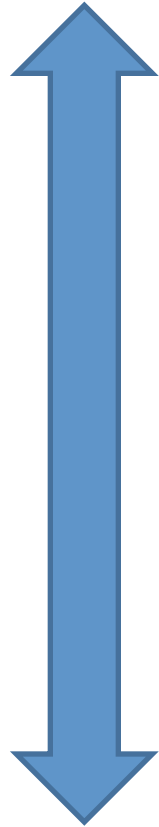
3.2 Using the corpora to sequence concepts (iii)

Order of Difficulty

- We don't expect the actual score to mean anything by itself: we do not associate the feature with this proficiency level.
- However, this process allows us to see which features are on average acquired later than other features.
 - for any two errors, the error with the lower score is made more often by lower proficiency learners, and is thus something that probably should be taught first.

Lexical Errors in terms of apparent difficulty

More common
with basic
learners

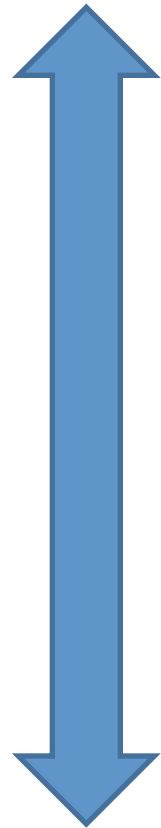


malformation
coinage
false-friend
transferred-spelling
verb-vocab-error
spelling-error
adverb-vocab-error
borrowing
noun-vocab-error
adjective-vocab-error

More common
with advanced
learners

Lexical Errors in terms of apparent difficulty

More common with basic learners



malformation
coinage
false-friend
transferred-spelling
verb-vocab-error
spelling-error
adverb-vocab-error
borrowing
noun-vocab-error
adjective-vocab-error

With the exception of borrowing, Transfer errors are more common for beginners, while later, intralanguage errors predominate.

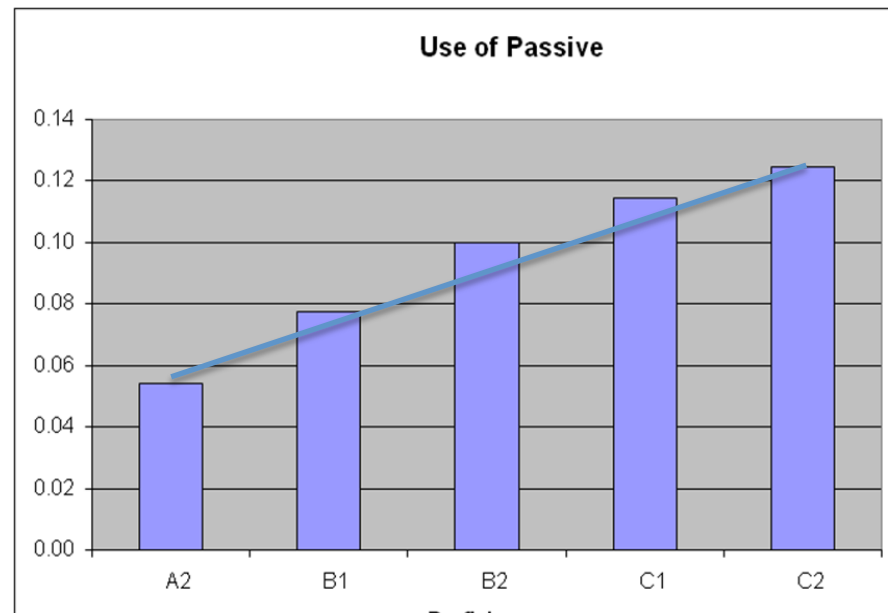
More common with advanced learners

Borrowings at advanced levels: more explicit mention of Spanish institutional terms: “Fiscal Jefe”

3.2 Using the corpora to sequence concepts (iii)

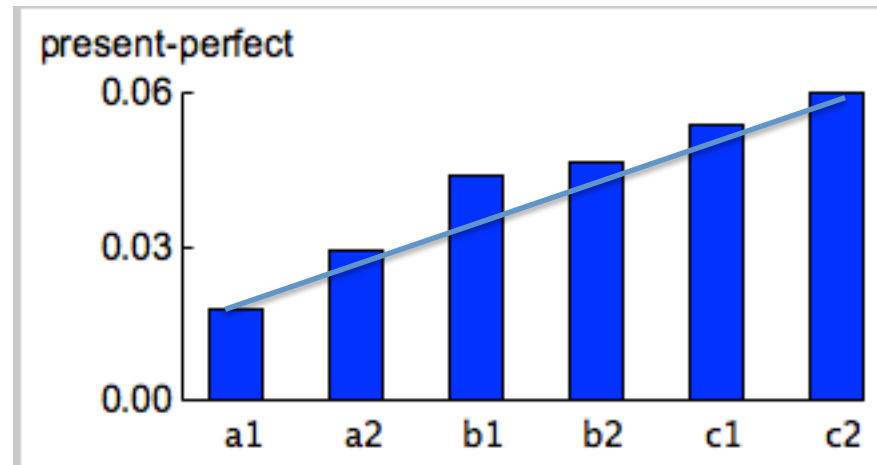
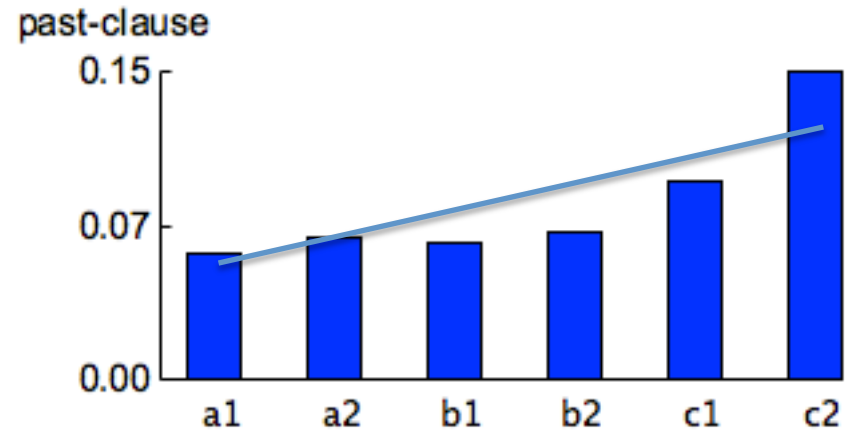
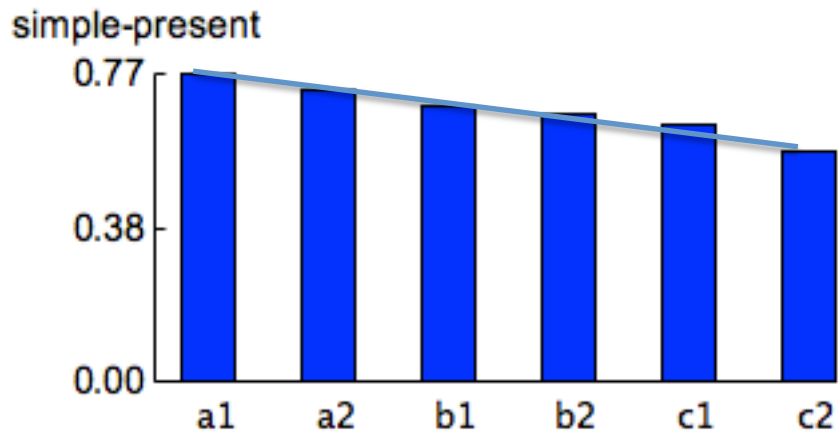
Slope of Proficiency line

- The slope of the “usage vs proficiency” chart indicates increasing or decreasing difficulty.
- We can rank syntactic features in terms of this slope
- Features with low or negative slope are not problematic beginner learners
- Features with more slope are those which continue to be learnt until higher levels.



3.2 Using the corpora to sequence concepts (iii)

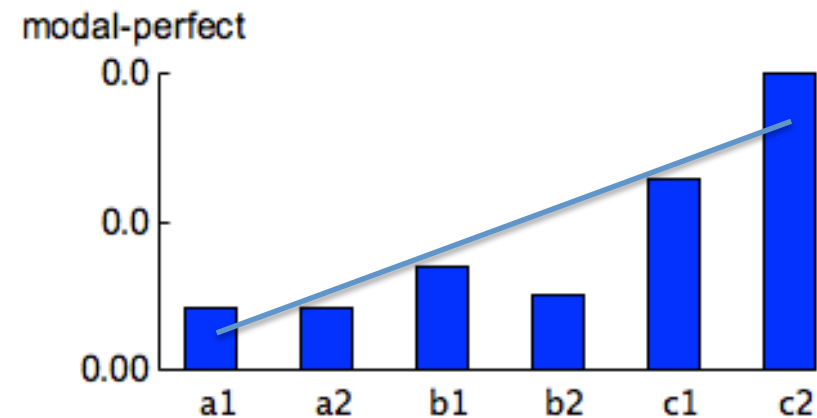
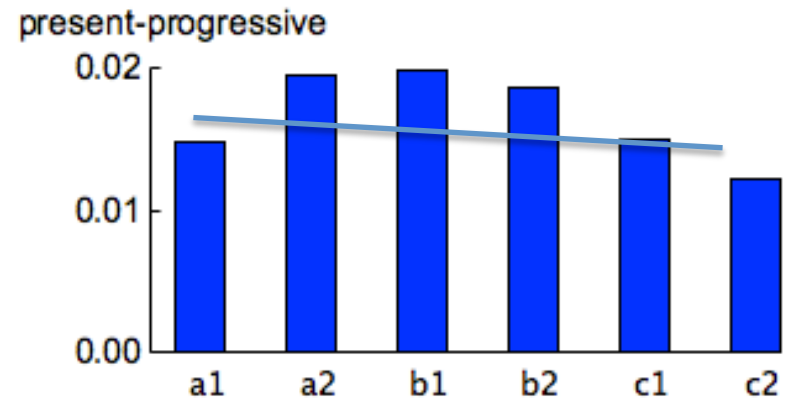
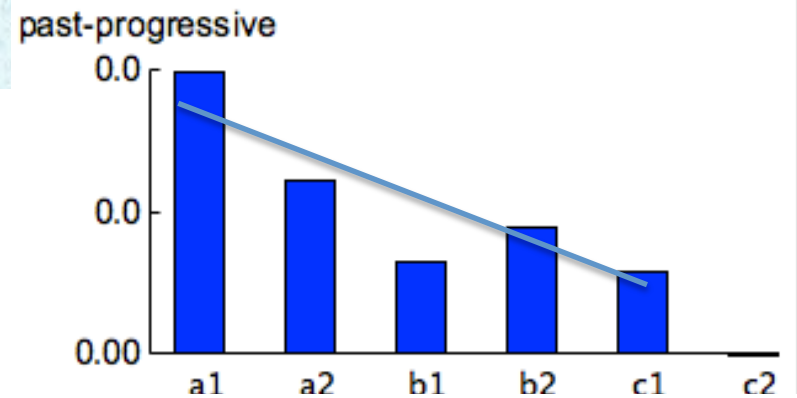
Slope of Proficiency line



3.2 Using the corpora to sequence concepts (iii)

Slope of Proficiency line

- **Results for tense-aspects**
(% change in use per point of prof.)
 - simple-future -1.3%
 - present-progressive -1.2%
 - past-progressive -0.5%
 - simple-present -0.3%
 - present-perfect 0.7%
 - future-progressive 0.9%
 - simple-modal 1.0%
 - simple-past 1.1%
 - present-progressive-perfect 1.2%
 - modal-progressive 1.4%
 - past-perfect 1.6%
 - modal-perfect 5.215%





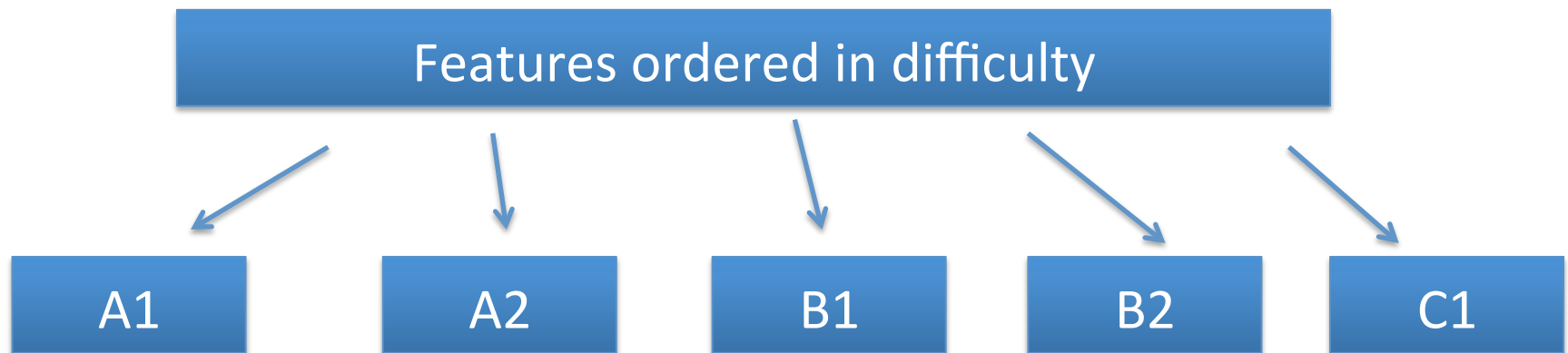
4. Conclusions

4. Conclusions

- Error annotation and syntactic annotations can show us what students need to learn
- But to see in what sequence they need to be taught this material is more difficult
- This paper has explored various methods for sequencing grammatical concepts based on learner data.
- Currently, the slope of the “usage vs. proficiency” curve is a good indicator of syntactic difficulty: higher values should be taught later.

Not covered here

- Dividing concepts into courses:
 - Given a list of grammatical concepts sequenced by difficulty, we can divide this list into equal-size subsets to be taught in each course within the sequence of courses in the degree.



Not covered here

- Dividing concepts into courses:
 - Prior to splitting the list, some shifting around of concepts to ensure that thematically related concepts are taught in the same course.
 - (using an optimisation algorithm)