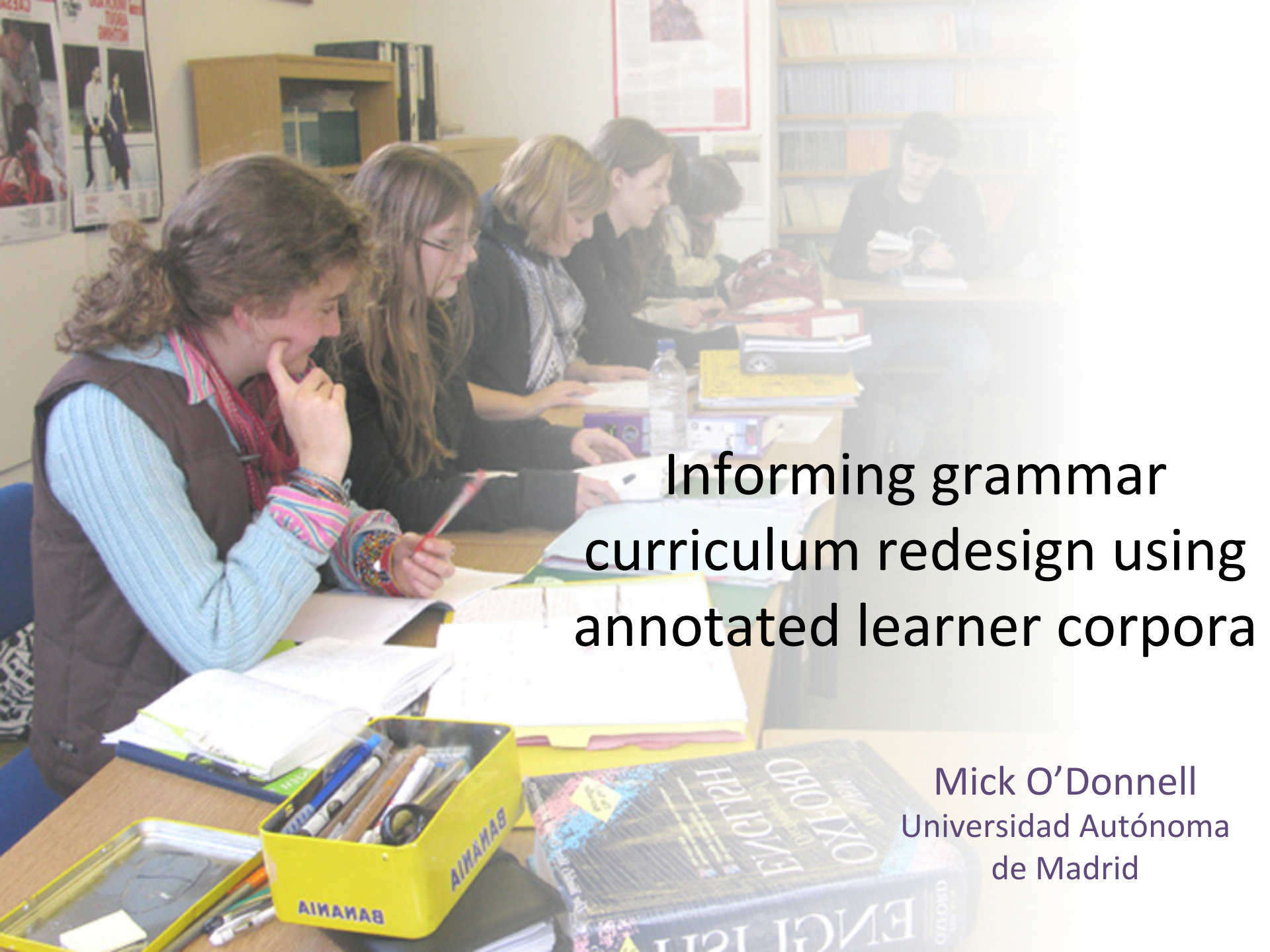


Informing grammar curriculum redesign using annotated learner corpora

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

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 the Spanish Ministerio de Ciencia e Innovación (FFI2009-14436/FILO)
- Runs: January 2010 – June 2013 (but we are applying for a new project)

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.

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

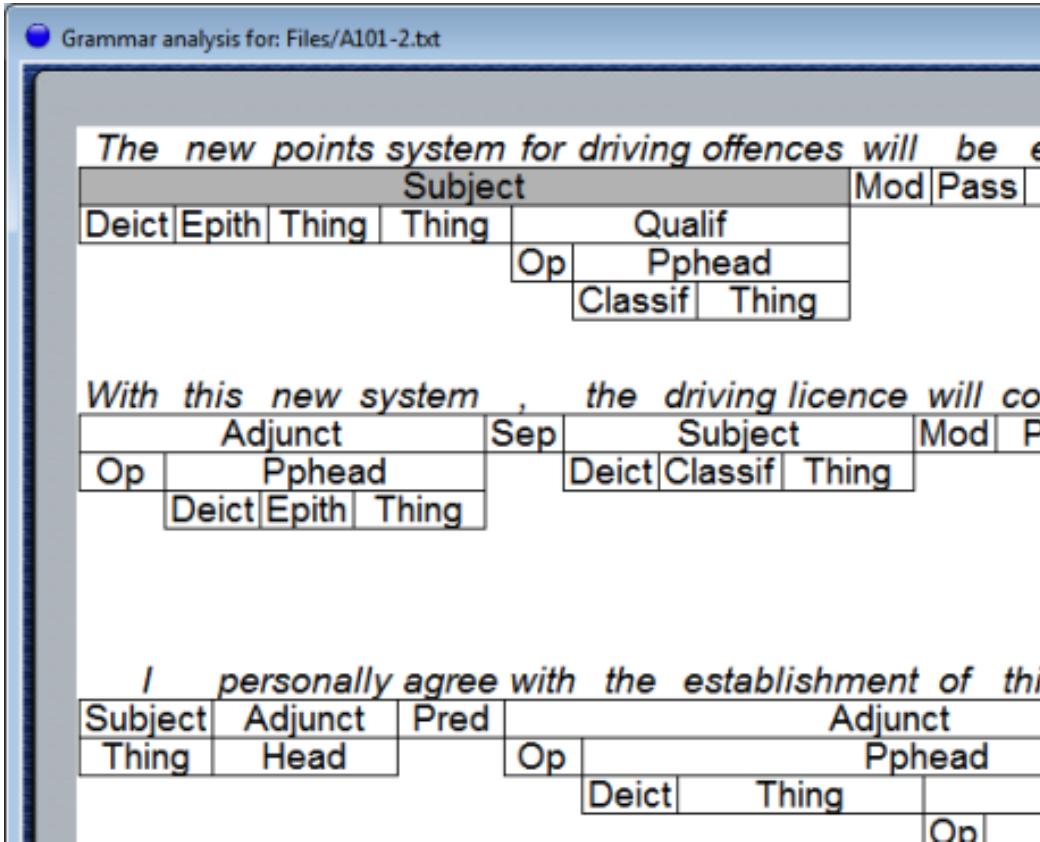
Discovering what learners need

- 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)

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



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

2. Deciding What to teach

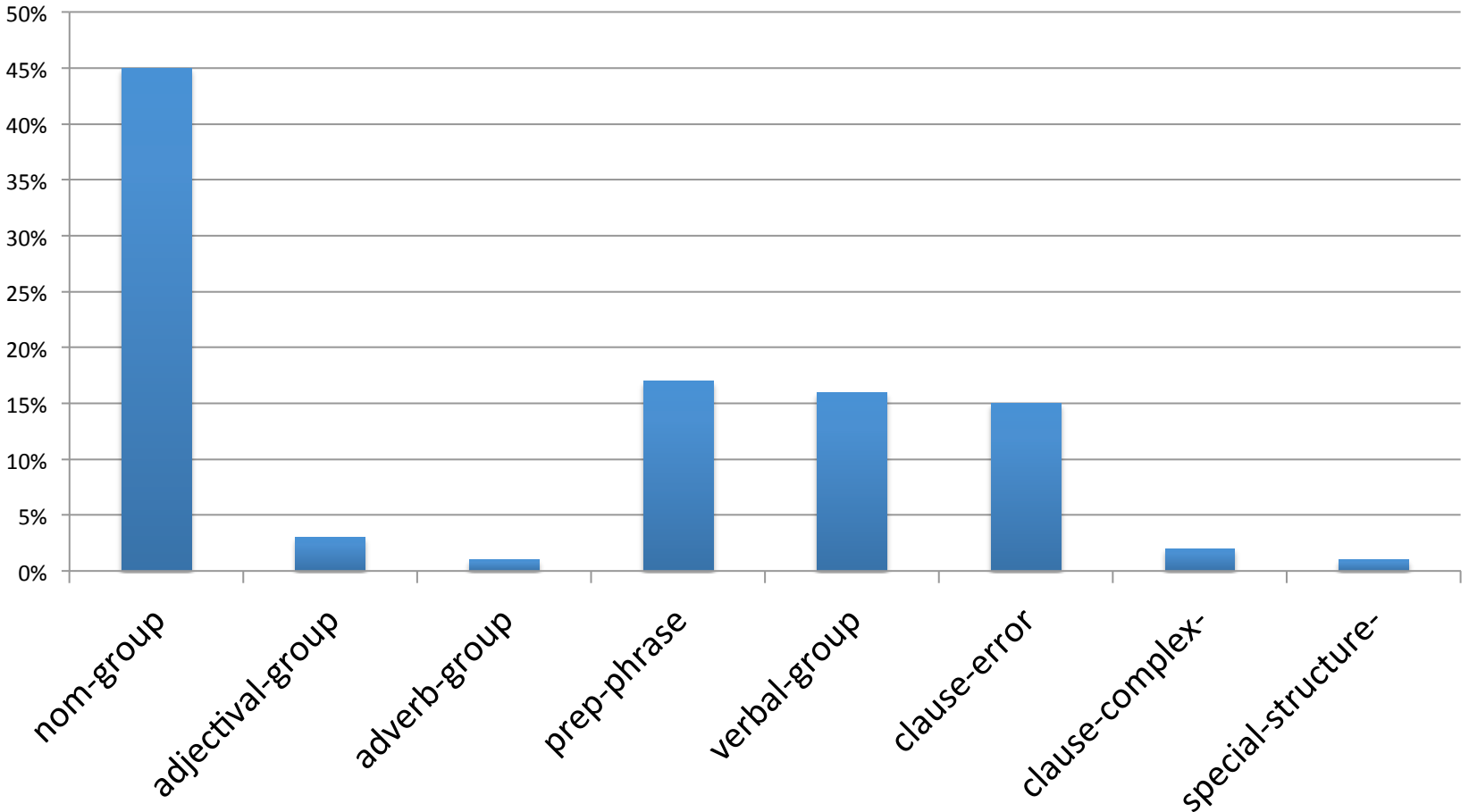
2. Deciding What to teach

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	religión	an (and)	other
Europa	sanity	gobernment	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	opinión	becouse	get
infantil	diary	politic	there	job
habituate	principately	costums	beleive	undeveloped
receptor	evollution	asociation		doing

3. Sequencing Grammar Material

3. Sequencing Grammatical Concepts

- To sequence teaching of grammatical concepts:
  - We need some way to relate each student text 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

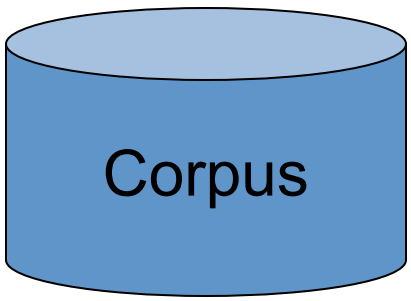
In the Treacle corpora:

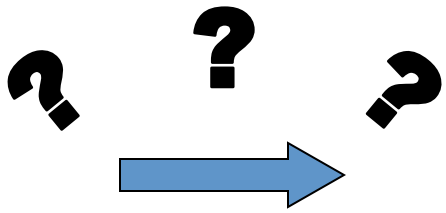
- all learners took the Oxford Short Placement Test within the month of writing.
- Only Grammatical proficiency tested.
- Proficiency score from 0-60.
- CEFR levels estimated from these scores

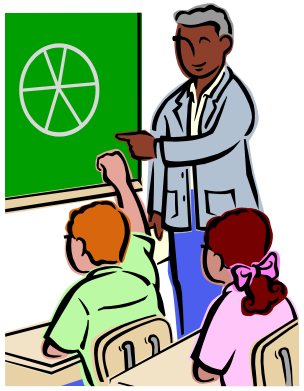
3.2 Using the corpora to sequence concepts

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**?

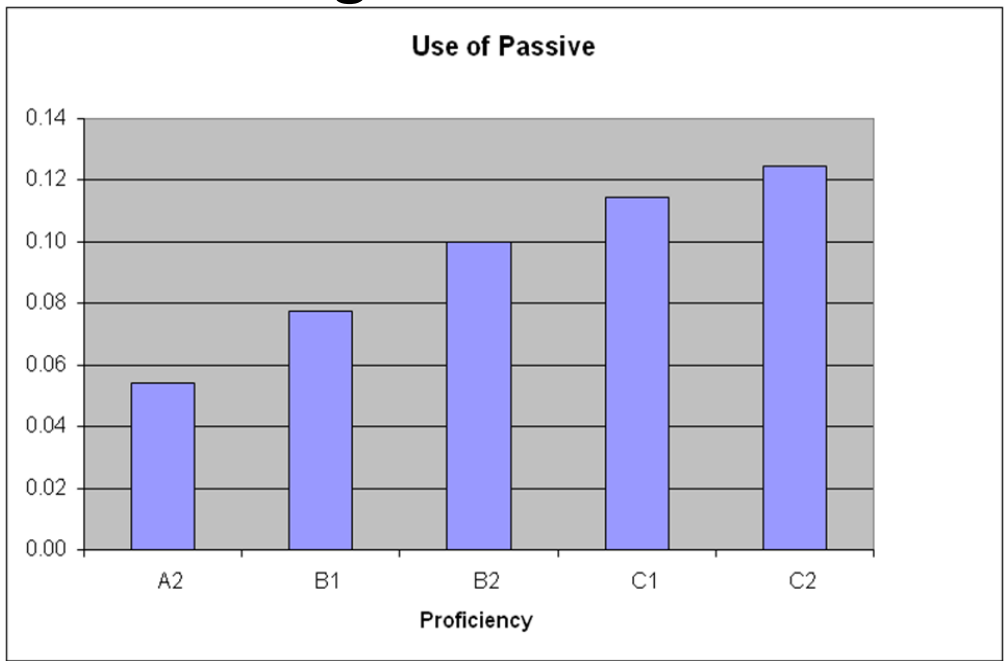


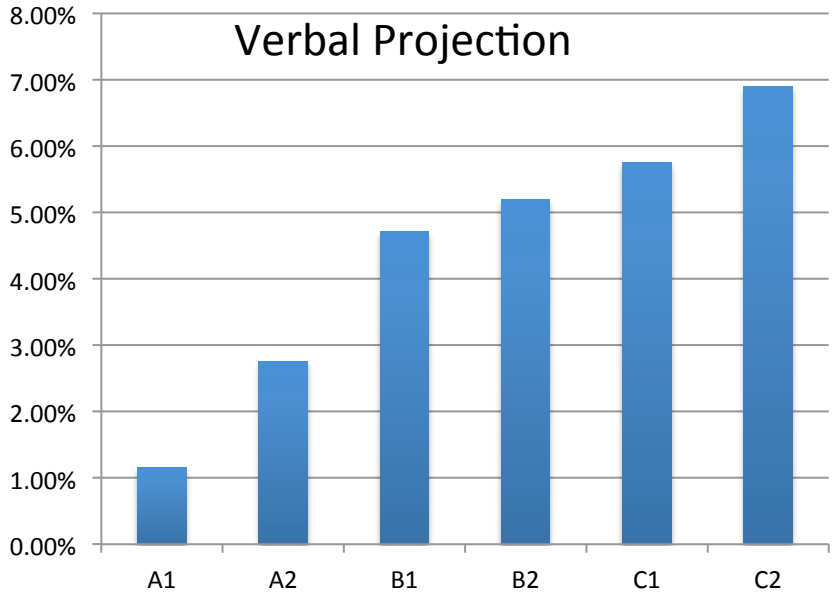




3.2 Using the corpora to sequence concepts (i) **Levels of usage**

- Levels of usage at different proficiency levels 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!

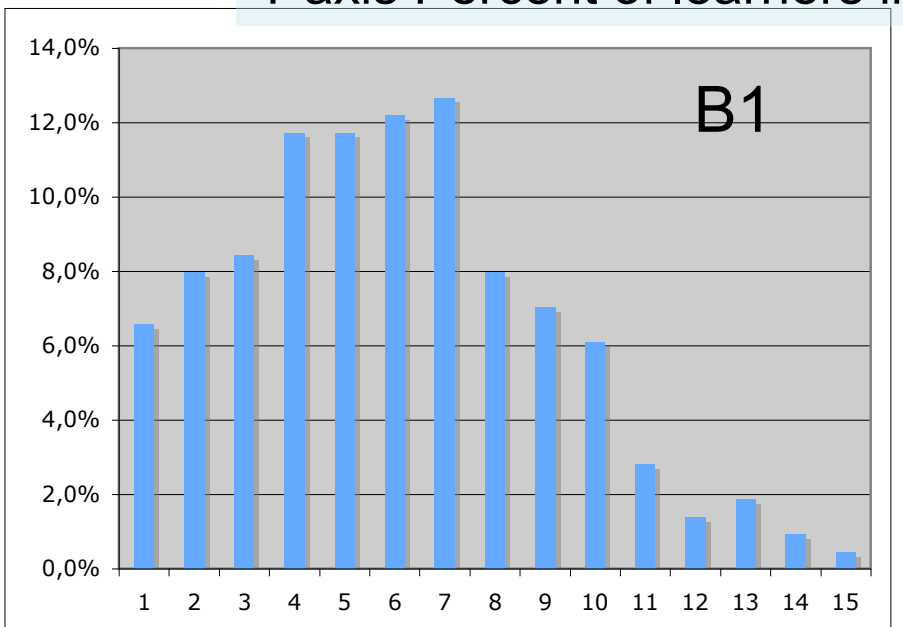


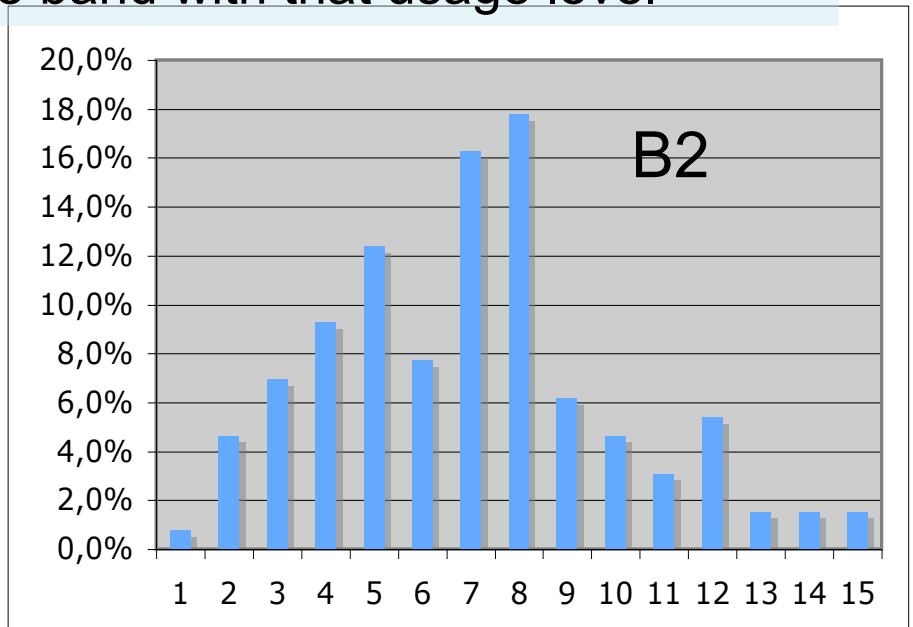


3.2 Using the corpora to sequence concepts (i)
Levels of usage: variation within a level

- 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

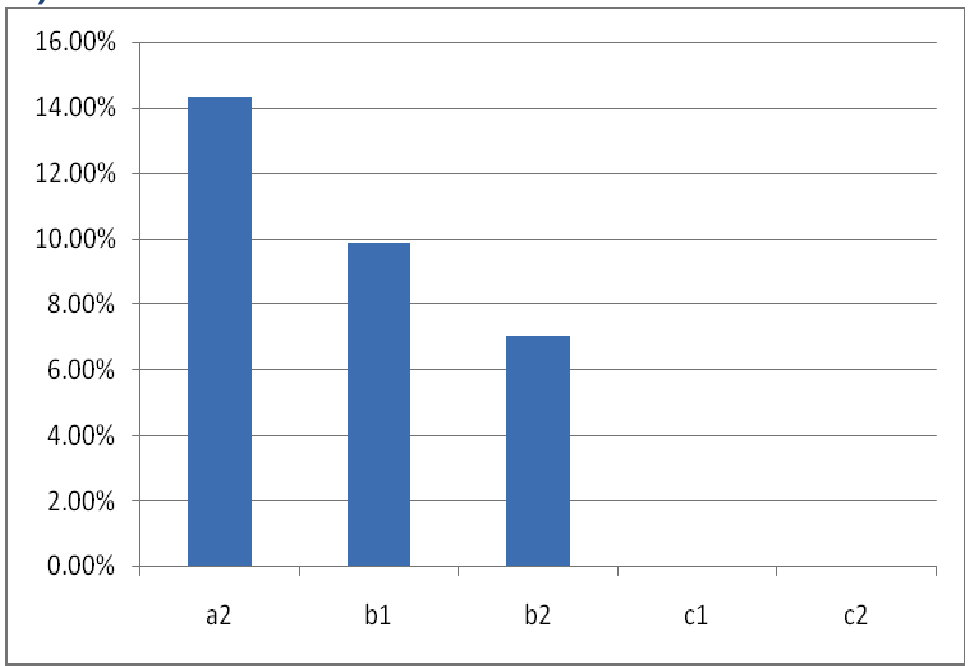




3.2 Using the corpora to sequence concepts (ii)
Onset of Use

- 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(%)

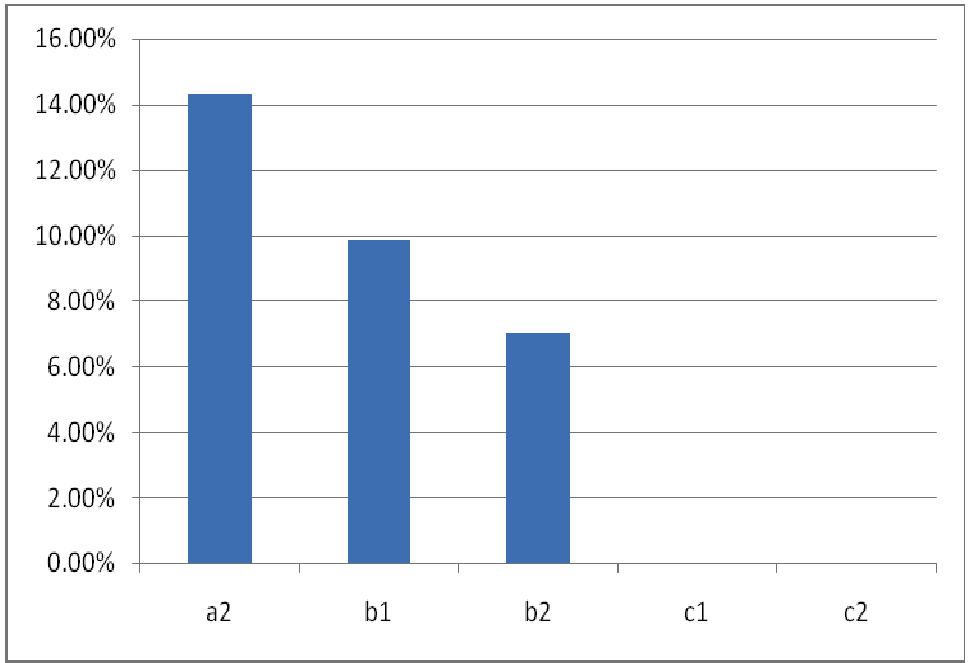


3.2 Using the corpora to sequence concepts (ii)
Onset of Use

We could say to 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(%)

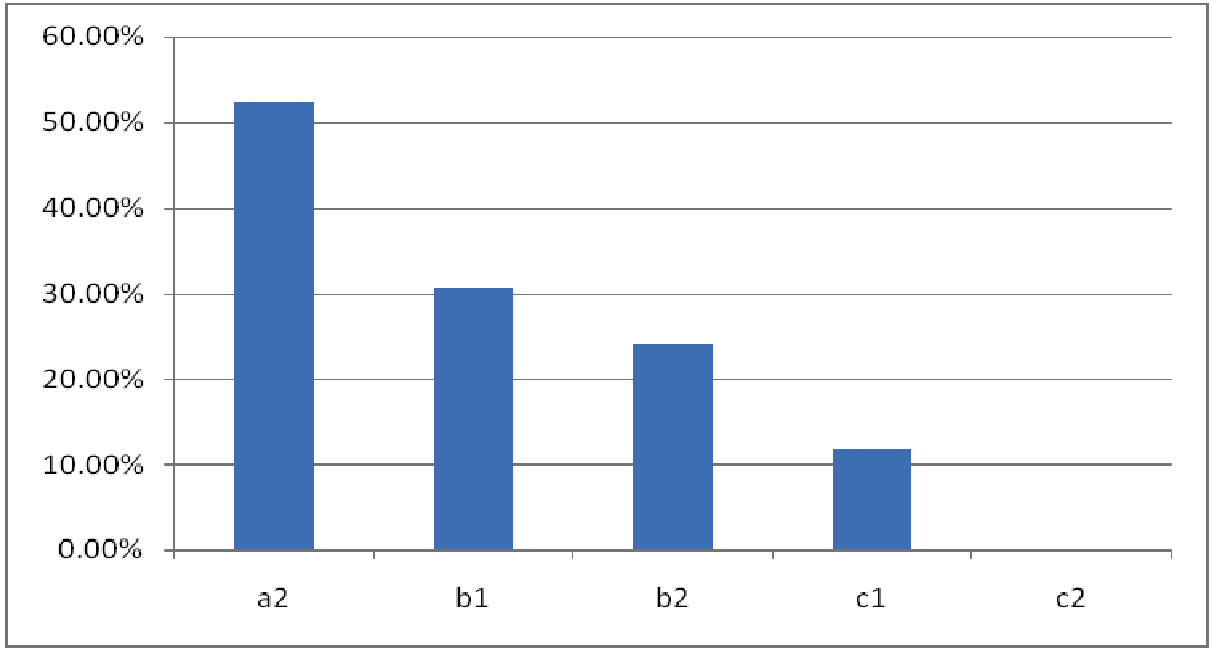


3.2 Using the corpora to sequence concepts (ii)
Onset of Use: problem

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

- This approach only useful 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)
Criterial 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)
Criterial 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

Proficiency levels do not in truth exist:

- they are a convenience created by language professionals to enable us to:
  - ratify our learners,
  - to provide target points for the teaching materials we provide.

- In reality, each language learner learns the foreign language in a unique manner, mastering linguistic concepts in their own time and order.

3.2 Using the corpora to sequence concepts (iii)
Order of Difficulty

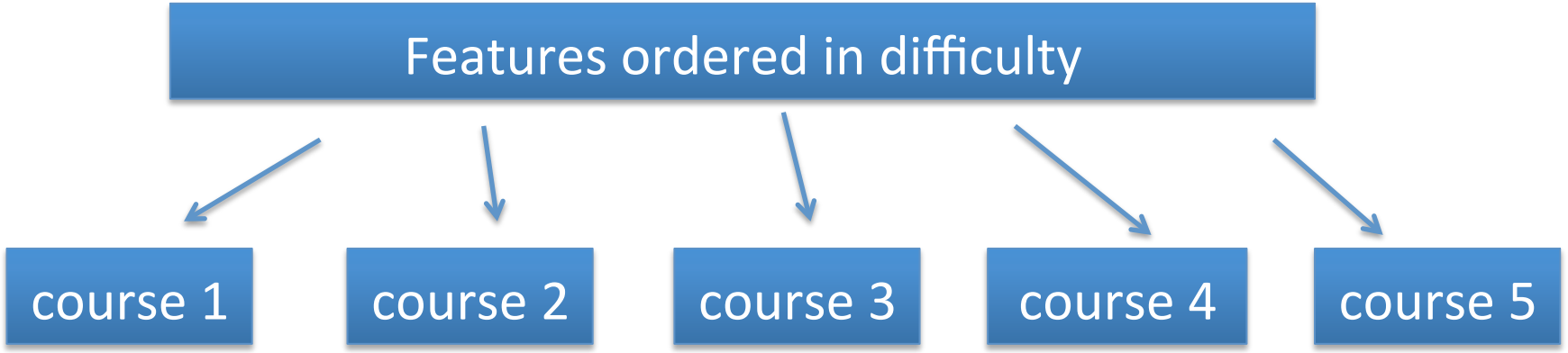
- Learners are all individuals
- However, if we look at learners **as a group**, linguistic features tend to be acquired in a particular order.
- The goal of our work here is to chart the order in which our learners acquire linguistic features.
- This ordering can then be used in curriculum design.

3.2 Using the corpora to sequence concepts (iii)
Order of Difficulty

- We use our learner corpus to chart the relative acquisition order 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.

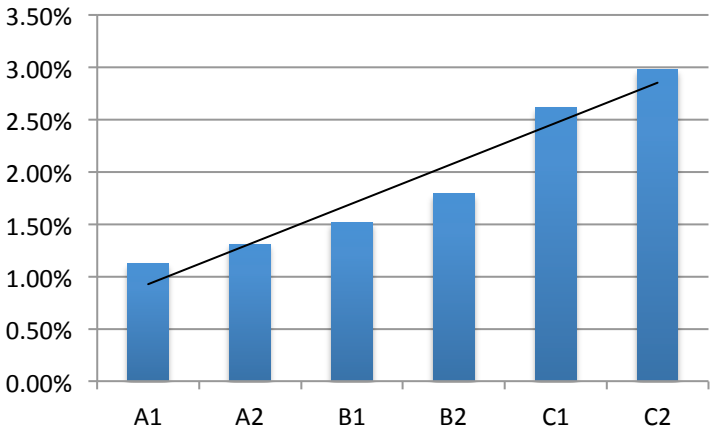
From Order of Difficulty to Teaching

- 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.



3.2 Using the corpora to sequence concepts (iii)
Patterns of feature acquisition

Rising Usage

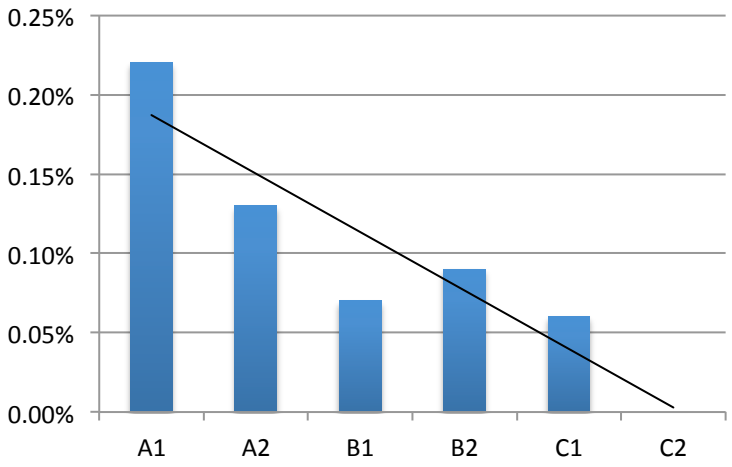


Use of passive

- Most common acquisition pattern.
- Initial 0 or low usage
- Increasing usage with proficiency
- Rise could relate to:
  - acquisiton of the structure (how to produce it)
  - or to acqusition of contexts of use (when to produce it)

3.2 Using the corpora to sequence concepts (iii)
Patterns of feature acquisition

Falling Usage

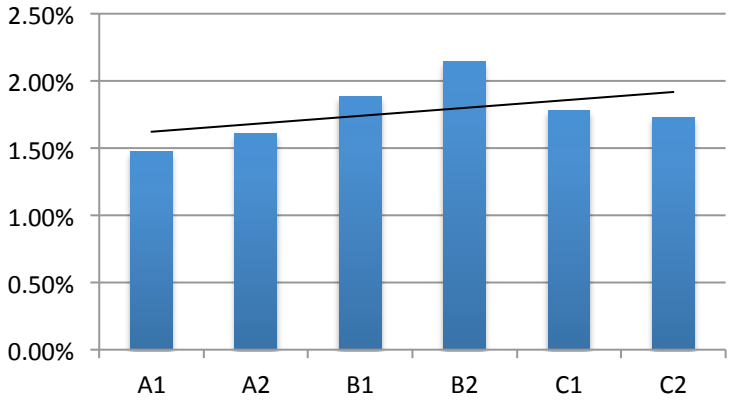


Use of past-progressive aspect

- Initial usage: learners transfer the structure from their L1.
- Falling usage with proficiency, as learners learn L2 contexts of use.

3.2 Using the corpora to sequence concepts (iii)
Patterns of feature acquisition

Rising-Falling Usage



Use of 'will' future forms

- Suggests the structure offers some initial learning difficulty overcome with rising proficiency.
- However, usage later falls.
- Possibly due to:
  - Learning of alternative strategies to express the same meaning
  - Learning L2 contexts of use.

3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

- We assume the following:

1. Structures with rising usage seem to offer some initial difficulty, either in construction of the structure, or with contexts of use.
2. Features with falling or level usage suggest the structure offers no difficulty of construction.

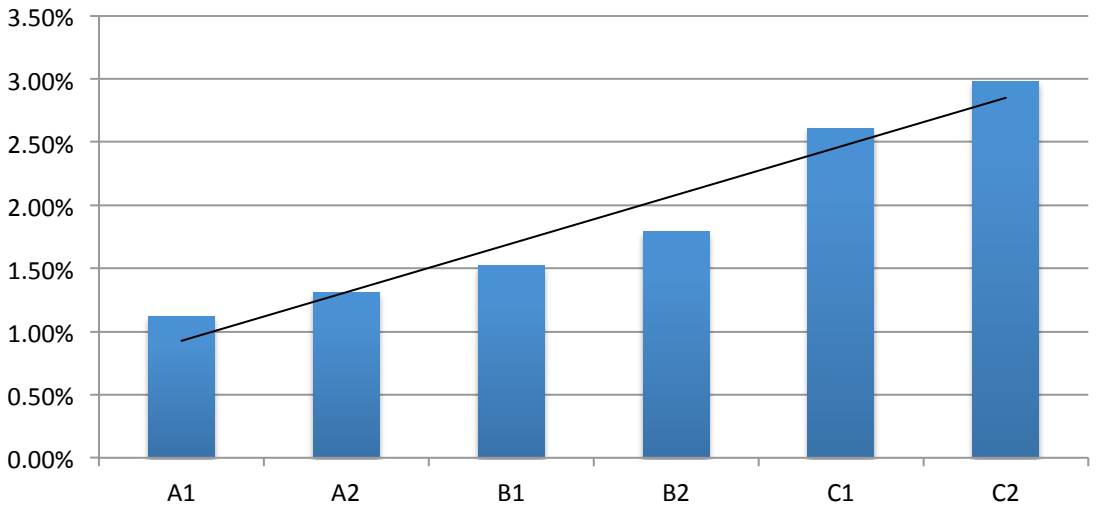
Thus, those structures with rising usage should be placed as more difficult that those with falling usage.

Those with rising-falling usage seem to come between these two.

3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

Line of best fit

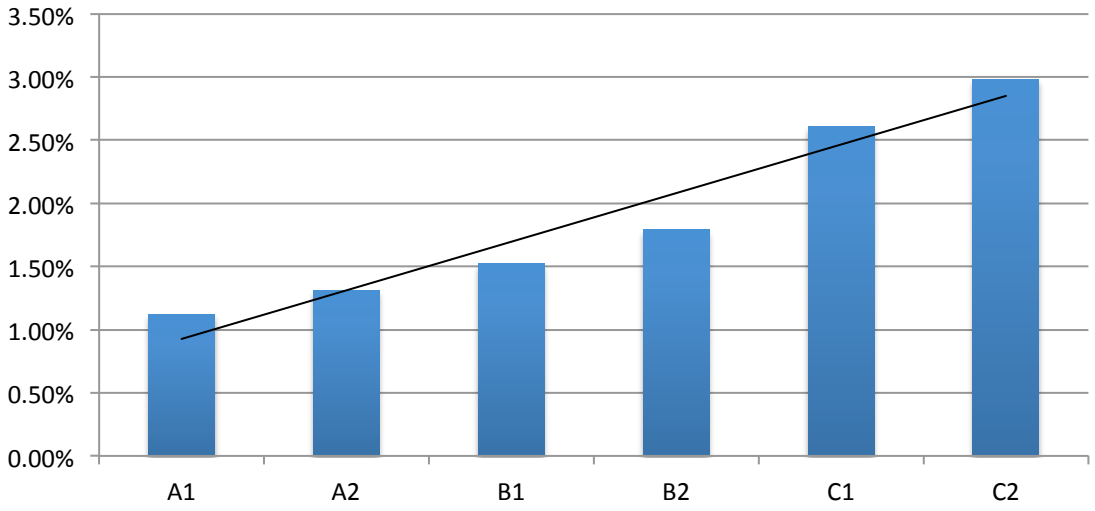
- For these usage graphs, we can fit a line of “best fit” to the data.
- This line can be straight or curved, but for the present discussion, we assume straight.



3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

Line of best fit

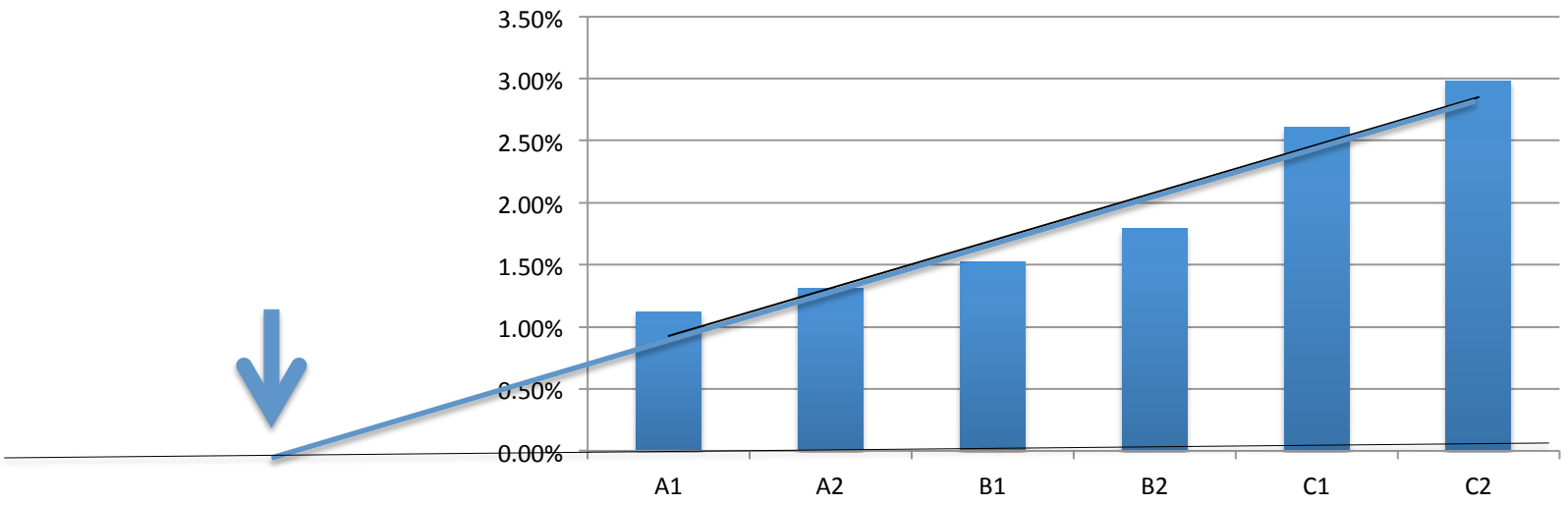
- The line is generally described in terms of 2 variable:
  - Slope (change in Y for a 1 unit change in X)
  - Y-intercept: the value of Y when X=0.



3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

Line of best fit

- For this study, we are also interested in the X-intercept: the theoretical proficiency point where the learner would start to use this structure.



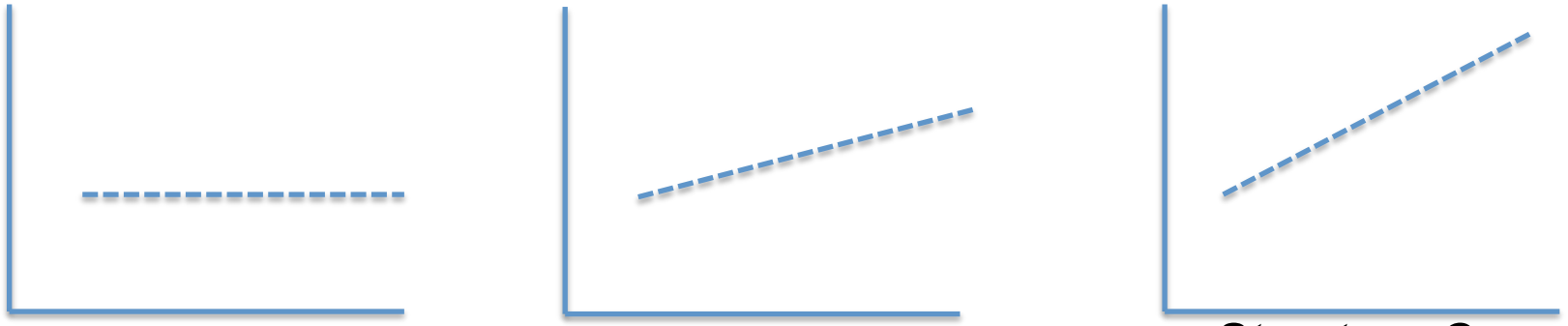
3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

Ordering within rising usage features

- For those structures with rising usage, we can posit two possible ways to measure their difficulty:

1. Slope of the Line of best fit.

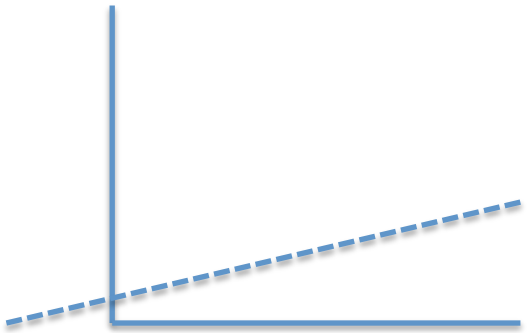
- Where usage does not change with proficiency (A), no learning is taken place.
- Steeper lines suggest the learner is changing their practice more as they progress, and thus that there is a higher level of difficulty involved.



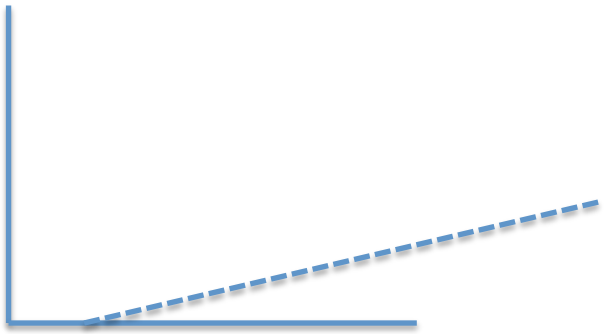
3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

2. X-Intersept.

- Two lines may have similar slope, but different X-intercept.
- The one with the larger x-intercept suggests that learners begin to use it later.
- In general, a higher X-intercept suggests the structure is more advanced.



Structure A



Structure B

3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

Some Results

- To test how well slope and X-intercept order features, we looked at how tense-aspect features would be ordered in difficulty for Spanish learners of English.

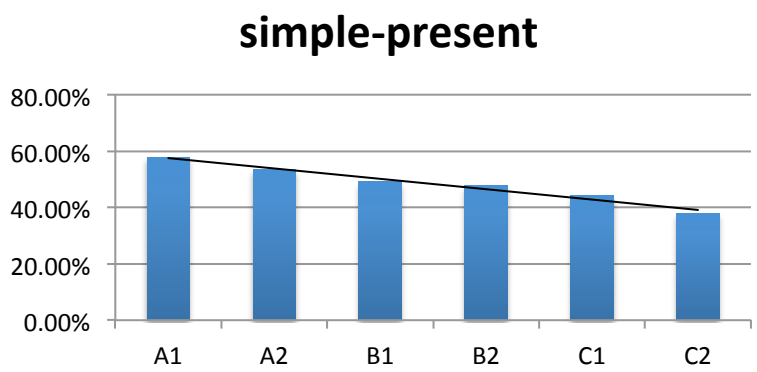
Ordered using slope value only

Tense-Aspect Feature	Slope	X-Intercept
simple-present	-0.00209	394
present-progressive	-0.00022	142
simple-future	-0.00014	266
past-progressive	0.00000	-308
future-progressive	0.00000	-9
modal-progressive	0.00001	-118
past-perfect	0.00003	-7
modal-perfect	0.00003	20
present-perfect	0.00022	-80
simple-modal	0.00023	-191
simple-past	0.00063	-16

3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

Some Results

- Simple-present is the easiest tense to produce, so learnt first.
- Learners move to other tenses as they progress.

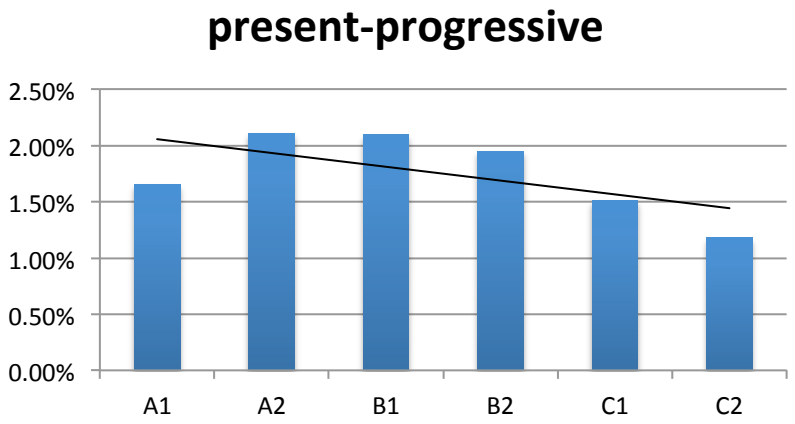


Tense-Aspect Feature	Slope	X-Intercept
simple-present	-0.00209	394
present-progressive	-0.00022	142
simple-future	-0.00014	266
past-progressive	0.00000	-308
future-progressive	0.00000	-9
modal-progressive	0.00001	-118
past-perfect	0.00003	-7
modal-perfect	0.00003	20
present-perfect	0.00022	-80
simple-modal	0.00023	-191
simple-past	0.00063	-16

3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

Some Results

- Present-progressive is the default tense for ongoing action, and similar in English and Spanish.

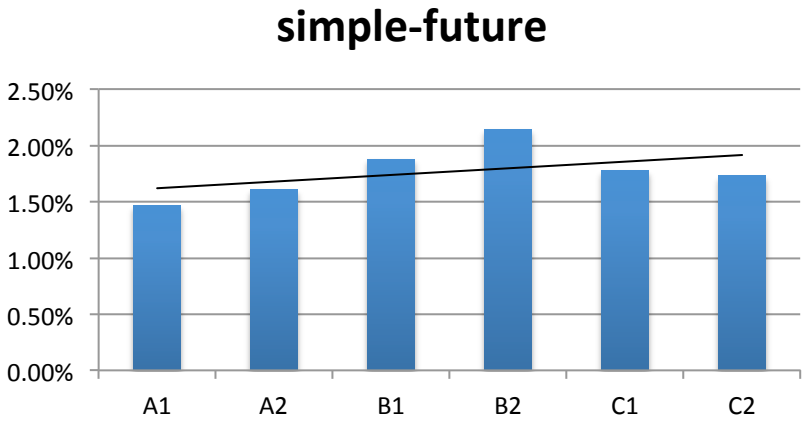


Tense-Aspect Feature	Slope	X-Intercept
simple-present	-0.00209	394
present-progressive	-0.00022	142
simple-future	-0.00014	266
past-progressive	0.00000	-308
future-progressive	0.00000	-9
modal-progressive	0.00001	-118
past-perfect	0.00003	-7
modal-perfect	0.00003	20
present-perfect	0.00022	-80
simple-modal	0.00023	-191
simple-past	0.00063	-16

3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

Simple-future is the easiest means of expressing expected results.

Not identical in Spanish and English, but not difficult.



Tense-Aspect Feature	Slope	X-Intercept
simple-present	-0.00209	394
present-progressive	-0.00022	142
simple-future	-0.00014	266
past-progressive	0.00000	-308
future-progressive	0.00000	-9
modal-progressive	0.00001	-118
past-perfect	0.00003	-7
modal-perfect	0.00003	20
present-perfect	0.00022	-80
simple-modal	0.00023	-191
simple-past	0.00063	-16

3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

Some Results

Progressive tenses developed next.

Tense-Aspect Feature	Slope	X-Intercept
simple-present	-0.00209	394
present-progressive	-0.00022	142
simple-future	-0.00014	266
past-progressive	0.00000	-308
future-progressive	0.00000	-9
modal-progressive	0.00001	-118
past-perfect	0.00003	-7
modal-perfect	0.00003	20
present-perfect	0.00022	-80
simple-modal	0.00023	-191
simple-past	0.00063	-16

3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty

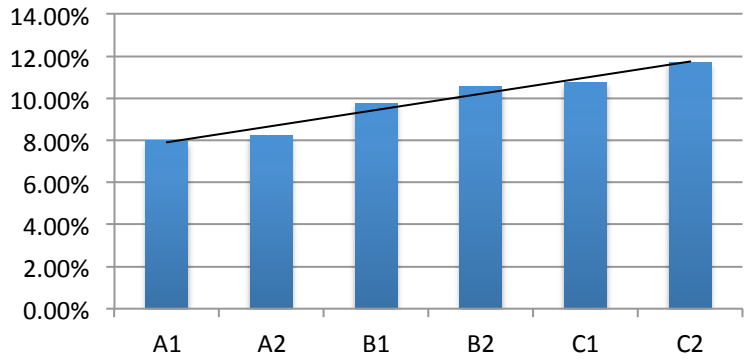
Some Results

Perfect tenses developed almost last

Tense-Aspect Feature	Slope	X-Intercept
simple-present	-0.00209	394
present-progressive	-0.00022	142
simple-future	-0.00014	266
past-progressive	0.00000	-308
future-progressive	0.00000	-9
modal-progressive	0.00001	-118
past-perfect	0.00003	-7
modal-perfect	0.00003	20
present-perfect	0.00022	-80
simple-modal	0.00023	-191
simple-past	0.00063	-16

3.2 Using the corpora to sequence concepts (iii)

Ordering structures in difficulty



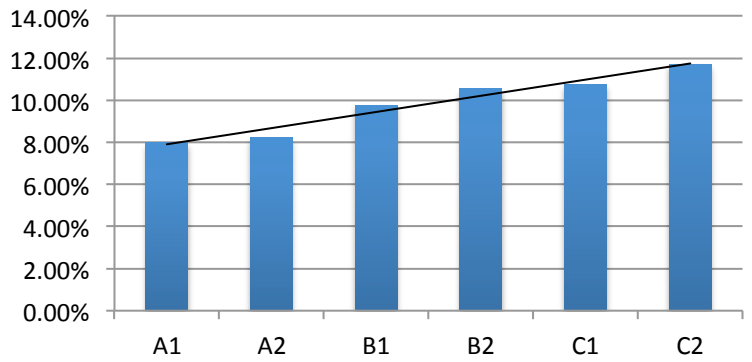
What is simple-modal doing here?

- seems learners do use this more as they progress.
- Probably because students are learning to hedge in their writing.
- Expanding contexts of use!

Tense-Aspect Feature	Slope	X-Intercept
simple-present	-0.00209	394
present-progressive	-0.00022	142
simple-future	-0.00014	266
past-progressive	0.00000	-308
future-progressive	0.00000	-9
modal-progressive	0.00001	-118
past-perfect	0.00003	-7
modal-perfect	0.00003	20
present-perfect	0.00022	-80
simple-modal	0.00023	-191
simple-past	0.00063	-16

3.2 Using the corpora to sequence concepts (iii)

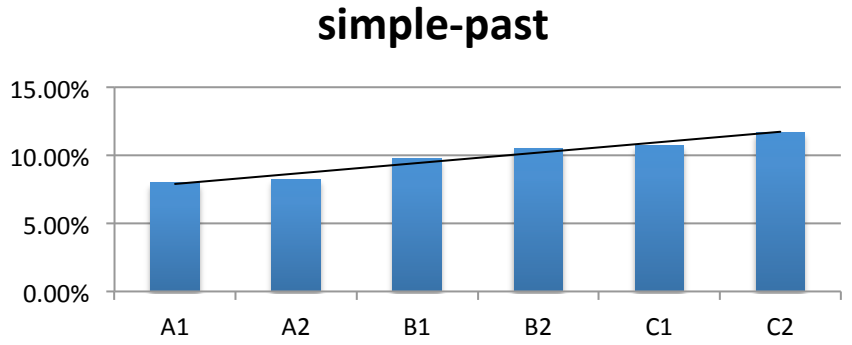
Ordering structures in difficulty



- If the X-Intercept was taken into account, simple-modal would be ranked earlier in the list,

Tense-Aspect Feature	Slope	X-Intercept
simple-present	-0.00209	394
present-progressive	-0.00022	142
simple-future	-0.00014	266
past-progressive	0.00000	-308
future-progressive	0.00000	-9
modal-progressive	0.00001	-118
past-perfect	0.00003	-7
modal-perfect	0.00003	20
present-perfect	0.00022	-80
simple-modal	0.00023	-191
simple-past	0.00063	-16

3.2 Using the corpora to sequence concepts (iii)
Ordering structures in difficulty



- Simple-past is also used increasingly as users progress
- Learners starting to use more evidence in their essays from past events, or quoting people.

Tense-Aspect Feature	Slope	X-Intercept
simple-present	-0.00209	394
present-progressive	-0.00022	142
simple-future	-0.00014	266
past-progressive	0.00000	-308
future-progressive	0.00000	-9
modal-progressive	0.00001	-118
past-perfect	0.00003	-7
modal-perfect	0.00003	20
present-perfect	0.00022	-80
simple-modal	0.00023	-191
simple-past	0.00063	-16

3.2 Using the corpora to sequence concepts (iii)
Order of Difficulty

Deriving Order of Difficulty of syntactic Structures

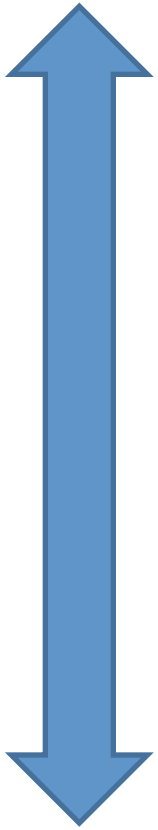
1. For each occurrence of the structure through the corpus, collect the proficiency score of the essay.
2. Average these scores, to give a difficulty index for the error-type.

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

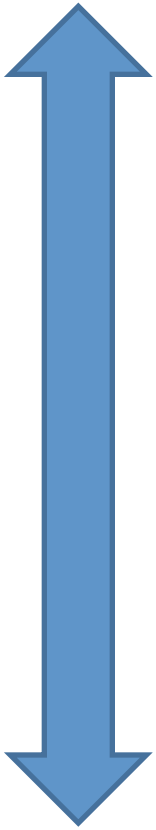


malformation
coinage
false-friend
transferred-spelling
verb-vocab-error
spelling-error
adverb-vocab-error
borrowing
noun-vocab-error
adjective-vocab-error

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

More common with advanced learners

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

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

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.

4. Conclusions (ii)

- We tried to evaluate how well the slope of the Usage vs. Proficiency line matches teacher's intuitions about difficulty of tenses.
- Most of the features are in good order.
- However, issues related to students changing the contexts of use of structures can distort the ordering. (e.g., simple-past)
- Need to combine with evidence from error analysis, which will show where they are working on structural issues vs. context of use issues.

END

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 and optimisation algorithm)

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.