How can Corpus Linguistics help in Foreign Language Teaching/Learning

Mick O'Donnell
Universidad Autónoma de Madrid

Background

- When teaching a Foreign Language, we often depend on course textbooks developed by others.

- These are generally developed for a **generic learner**, and may not specifically fit our own learners.
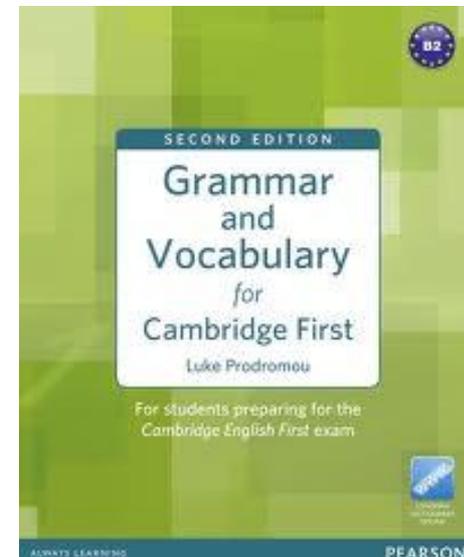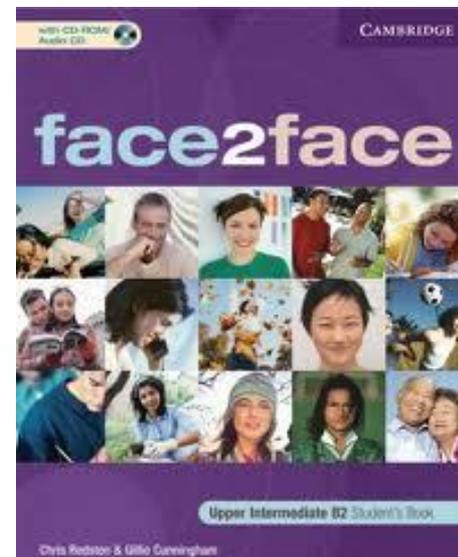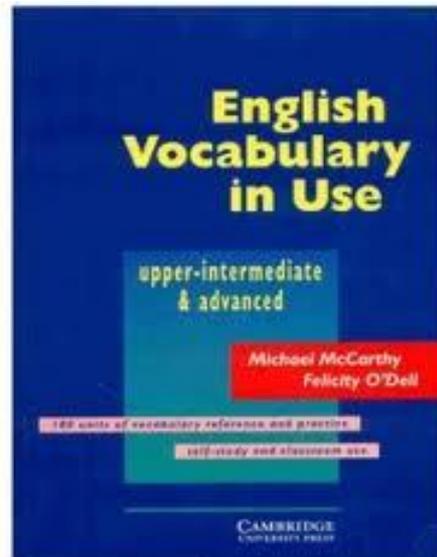
Background

- In relation to a particular mother-tongue group, a Foreign Language course designed for the generic learner will spend too much time on some areas, and not enough time on needed areas.

- As language teachers, we need to save time by focusing mainly on those aspects of the FL that are problematic for our language learners

E.g., for Spanish:

Decir -> Say or *Tell*?

Actualmente ✗→ *Actually*

*La sociedad actual* ✗→ *The Society today*

Background

- We need to discover the real needs of our own students, or students in similar learning contexts.

- One way of doing this: **Analysing Learner Corpora**

- Looking at the language students produce, to see what real problems they are having.

Analysis is not enough!!!

Analysis of what learners actually do

Intervention in the teaching/learning process
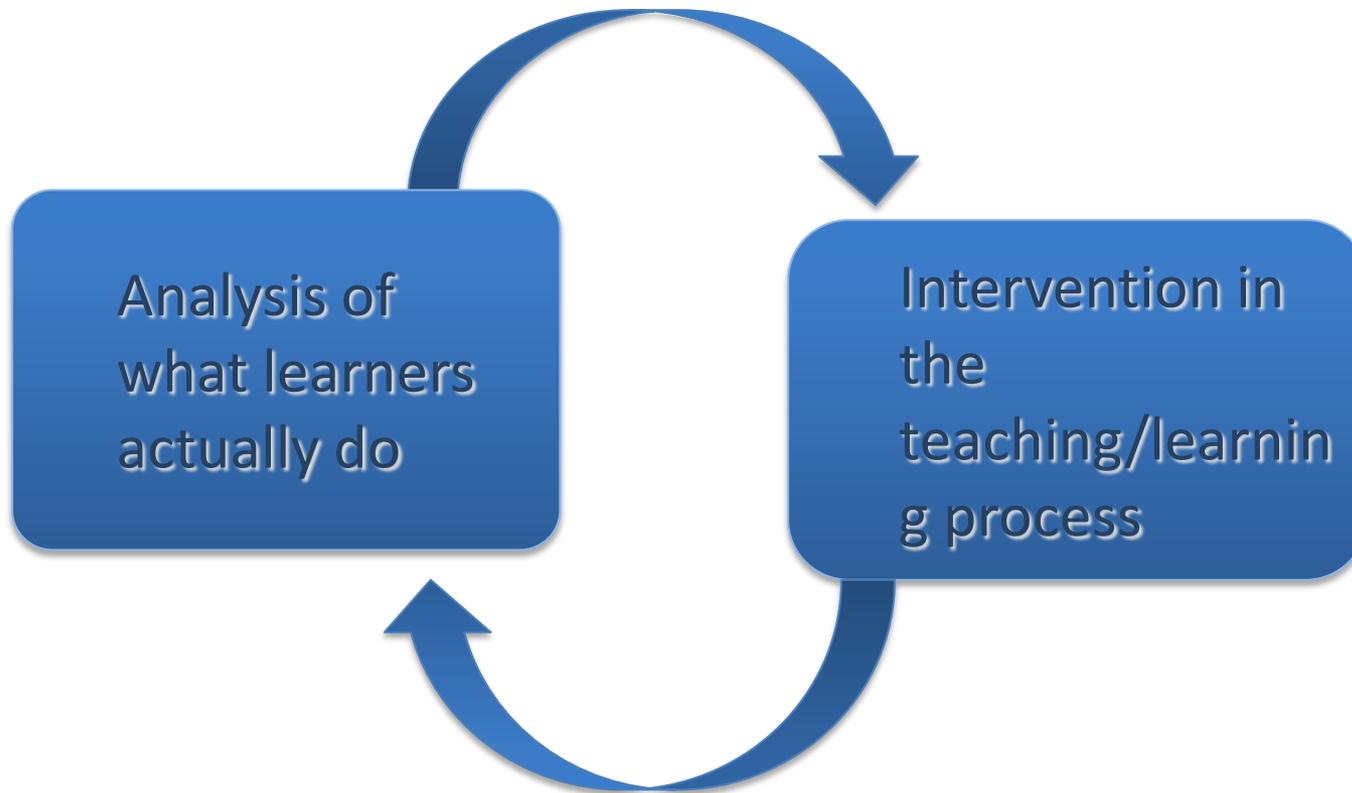
Aims of Talk

- I will show how Corpus Linguistics can be used to improve Foreign Language Teaching (FLT)

- I will start with **Learner Corpora**

- I will then discuss various ways in which **Native corpora** can be used in FLT.

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 – December 2012

1. What is a corpus?

- A "corpus" is a coll[ection of] computerised 'texts', wit[h] metadata

    - Metadata: information [about] the text, who w[rote it,] when, under what [...] what it is about, etc.

- 'Texts' which might be [written] or spoken

- Spoken texts transcribed to [enable] annotation.

1. What is a corpus?

- A corpus is usually **annotated** with linguistic information.

- The actual linguistic information annotated depends on what you are studying

- Typical:

  - Part of Speech (POS)

  - Syntax/Grammar

  - Discourse (e.g., speech acts, generic structure, etc.)

1. What is a corpus?

- **Native Corpus**: a corpus of texts produced by **native speakers**.

  – Good for studying norms in native speakers that learners might approximate

- **Learner Corpus**: a corpus of texts produced by **learners** of a language:

  – Good for studying what learners actually do.

2. Using Learner Corpora

How to get a learner corpus?

- **Use one already collected, e.g.,**

  – **The ICLE Corpus**: (Granger 2003): *International Corpus of Learner English*: contains texts written by learners of English from 16 language backgrounds. (best known of the learner corpora)

  – **Longmans Learners' Corpora**: a 10 million word corpus of texts written by English Learners from around the world.

  – **Cambridge Learner Corpus** (CLC): a large collection of exam scripts written by students taking Cambridge ESOL English exams around the world. Currently over 135,000 scripts.

How to get a learner corpus?

- **WriCLE**: Spanish University learners of English. From Universidad Autonoma de Madrid, 700,000 words with proficiency levels (Contact Paul Rollinson)

- **Br-ICLE** (Brazilian Portuguese portion of ICLE) (Tony Berber Sardinha and others)

- **ComAprend** (Tagnin et al) – Brazilian learners of several languages. Written.

How to get a learner corpus?

- **Spoken Learner corpora:**

  – **LINDSEI**: (Granger 2003): (Louvain International Database of Spoken English Interlanguage) (De Cock et al 1995): Transcripts of spoken interaction, Learners of English.

  – **LINDSEI-BR** (Mello etc.) Brazilian portion of the LINDSEI corpus: Brazilian university learners of English. In process. Spontaneous spoken: narratives, free discussions and descriptions.

Learners of Spanish

- **CEDEL 2** (*Corpus Escrito del Español L2*) (Cristobal Lozano, Universidad de Grenada): Learners of Spanish from around the world: http://www.uam.es/proyectosinv/woslac/cedel2.htm

- **SPLLOC** (Spanish Learner Language Oral Corpus): English learners of Spanish: http://www.splloc.soton.ac.uk/

How to get a learner corpus?

=> **Collect your own**

2. Using a Learner Corpus

2.1 How to get a learner corpus?

- **Collect your own.**

- **Advantages**:
  - You can study the production (written or spoken) of your own students, to see what *they* need.

  - You can control the types of texts collected, and the manner of collecting it.
    - E.g., targeting particular grammatical areas
    - E.g., not allowing access to grammar checkers

  - You can control what data you collect, ensuring you have all the metadata needed for your study.
    - E.g., Age? Gender? Native dialect of Spanish?

Collect your own corpus:

- **But beware!**

  – Lots of work!!!!

  – Many studies get bogged down in the collection and
  time for the analysis.

How to **USE** a learner corpus?

1. Discovering major problems of your learners

2. Curriculum progression: discovering **in what order** material is best taught across a curriculum

3. As a source of student **exercises**.

4. Studying **classroom interaction** strategies

# 2. Using a Learner Corpus

2.2 How to USE a learner corpus?

1. Discovering major problems of your learners

2. Curriculum design: discovering in what order material is best taught across a curriculum

3. As a source of student exercises.

4. Studying classroom interactions

2. Using a Learner Corpus

How to USE a learner corpus?

– Study your student's writing/speaking so as to get a clearer picture of the needs of the students.

– Reformulate your teaching curriculum and methodology on the basis of their true needs.

Study Intervene
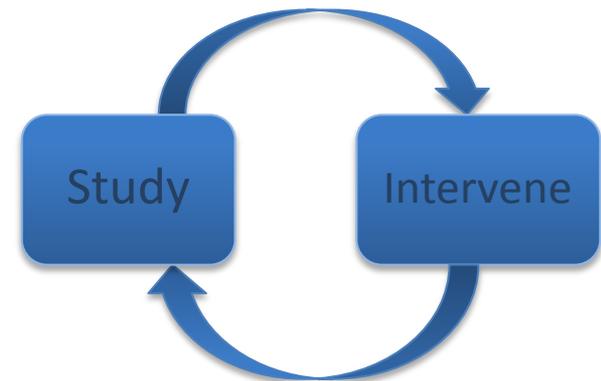
How to discover student needs?

- **Error analysis** is one way to explore the linguistic needs of learners (e.g. Dagneaux et al 1998).

> **Error analysis**: counting the types and frequency of errors that students make in their texts

✗ Who give me th
The man in the
✗ The society tod
✗ Do'nt let it get
Tomorrow, I an

2. Using a Learner Corpus

In the TREACLE project, we error coded:

- 300 student essays
- 110,000 words:
- 16,000 errors across
- Across 6 proficiency levels

Allows us to see:

- Major problem areas for the learners as a whole
- How problem areas evolve as they progress

Error coding process

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

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.
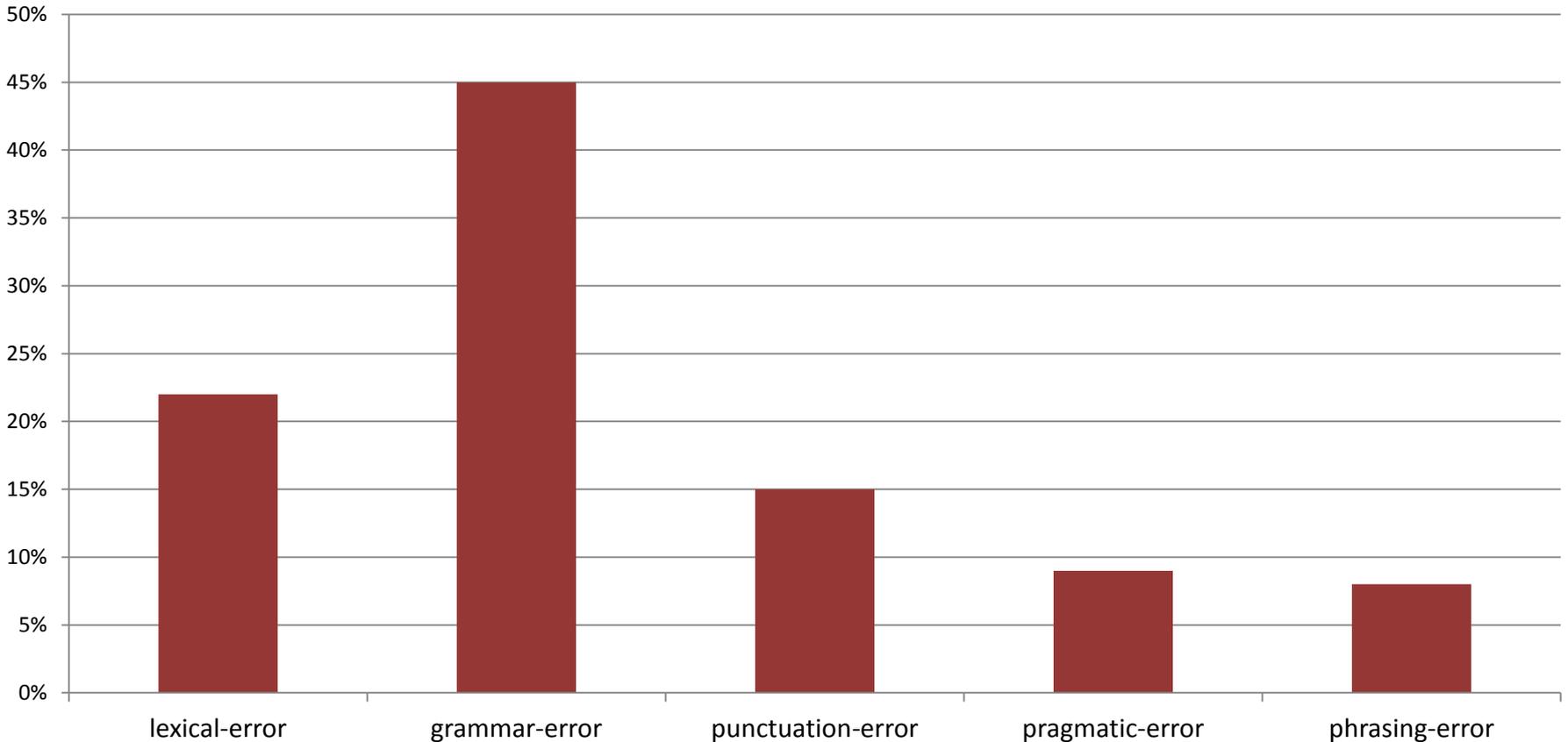
Error Annotation: Grammatical Errors
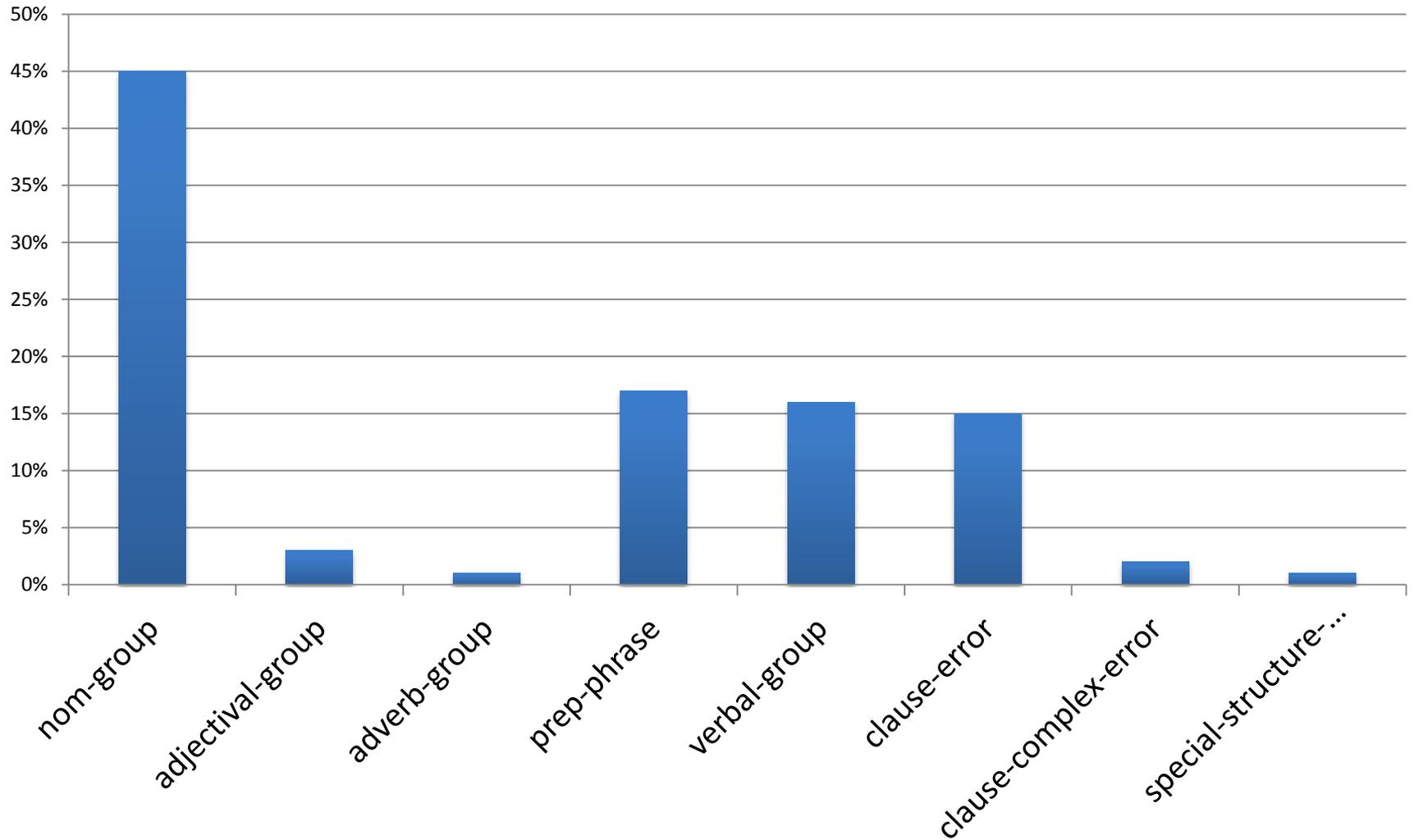
Subtypes of Nominal-group error
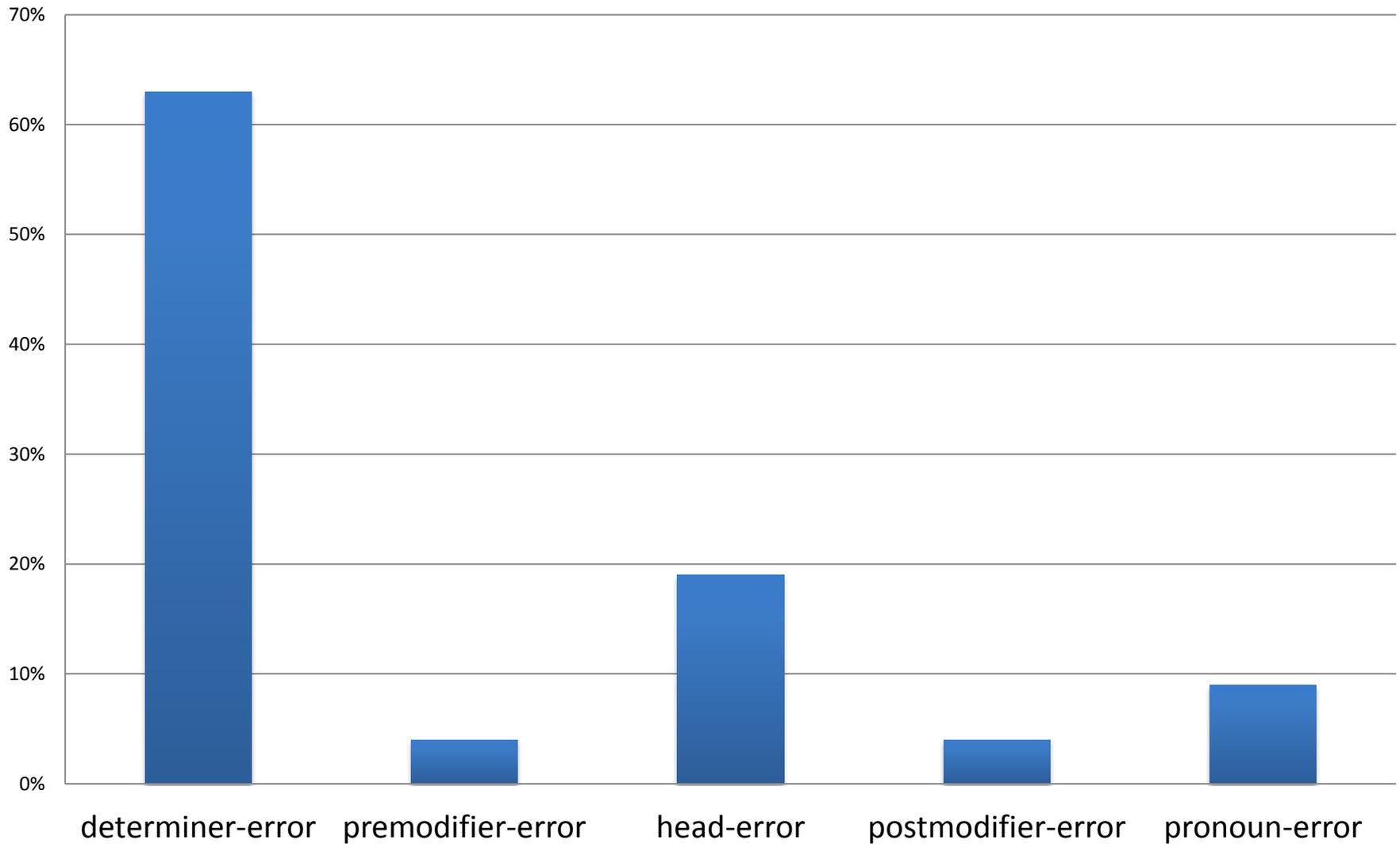
Most common Lexical Errors

Spelling errors			Transfer errors			Wordchoice errors		
inmigration	76	8.00%	actually	10	3.44%	persons	43	3.17%
inmigrants	69	7.44%	optative	5	1.72%	other	23	1.80%
live	20	2.11%	supposes	5	1.72%	work	17	1.25%
inmigrant	15	1.58%	fomenting	5	1.72%	works	17	1.25%
religión	14	1.47%	course	4	1.37%	do	13	0.96%
ilegal	11	1.16%	cannon	4	1.37%	make	13	0.96%
whit	11	1.16%	important	4	1.37%	win	11	0.81%
wich	10	1.05%	sanity	3	1.03%	job	10	0.74%
gobernment	9	0.95%	asignature	3	1.03%	have	9	0.66%
lifes	9	0.95%	poblation	3	1.03%	take	8	0.59%

2. Using a Learner Corpus

How to USE a learner corpus?

1. Discovering major problems of your learners

2. Curriculum design: discovering in what order material is best taught across a curriculum

3. As a source of student exercises.

4. Studying classroom interactions

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.

General Errors (by type)
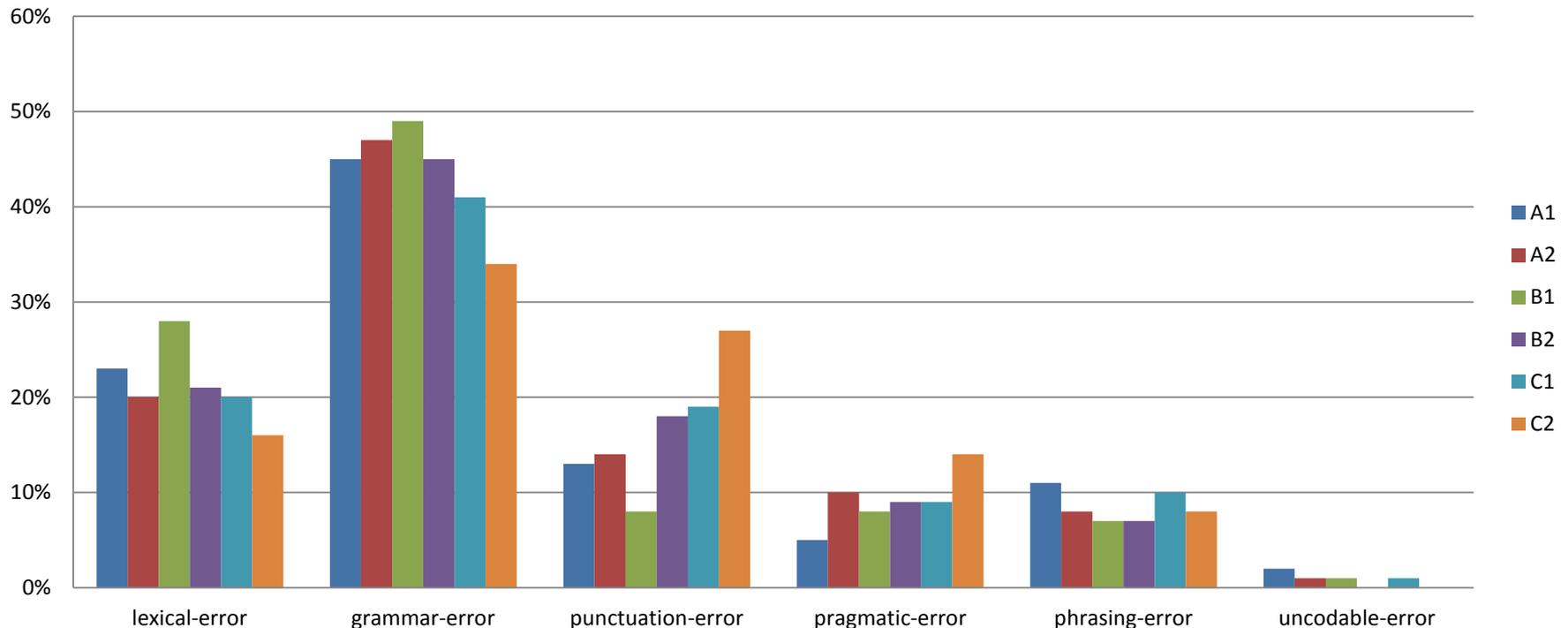
Error Annotation: When to teaching

- Initial attention on the nominal group and the prep phrases.

- As students progress, more attention needed on clause structure issues.
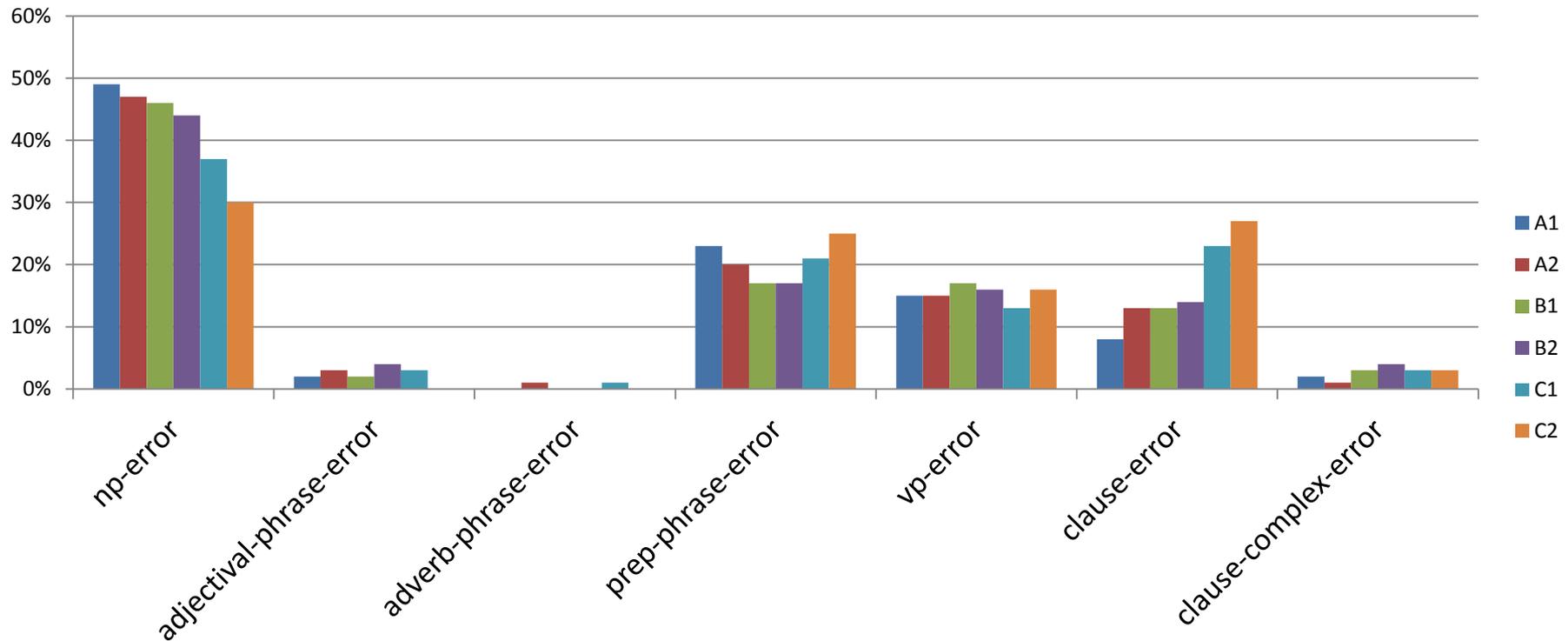
unnecessary-capitalisation incorrect-comparative-formation

other-phrasing-error incorrect-modal-for-context
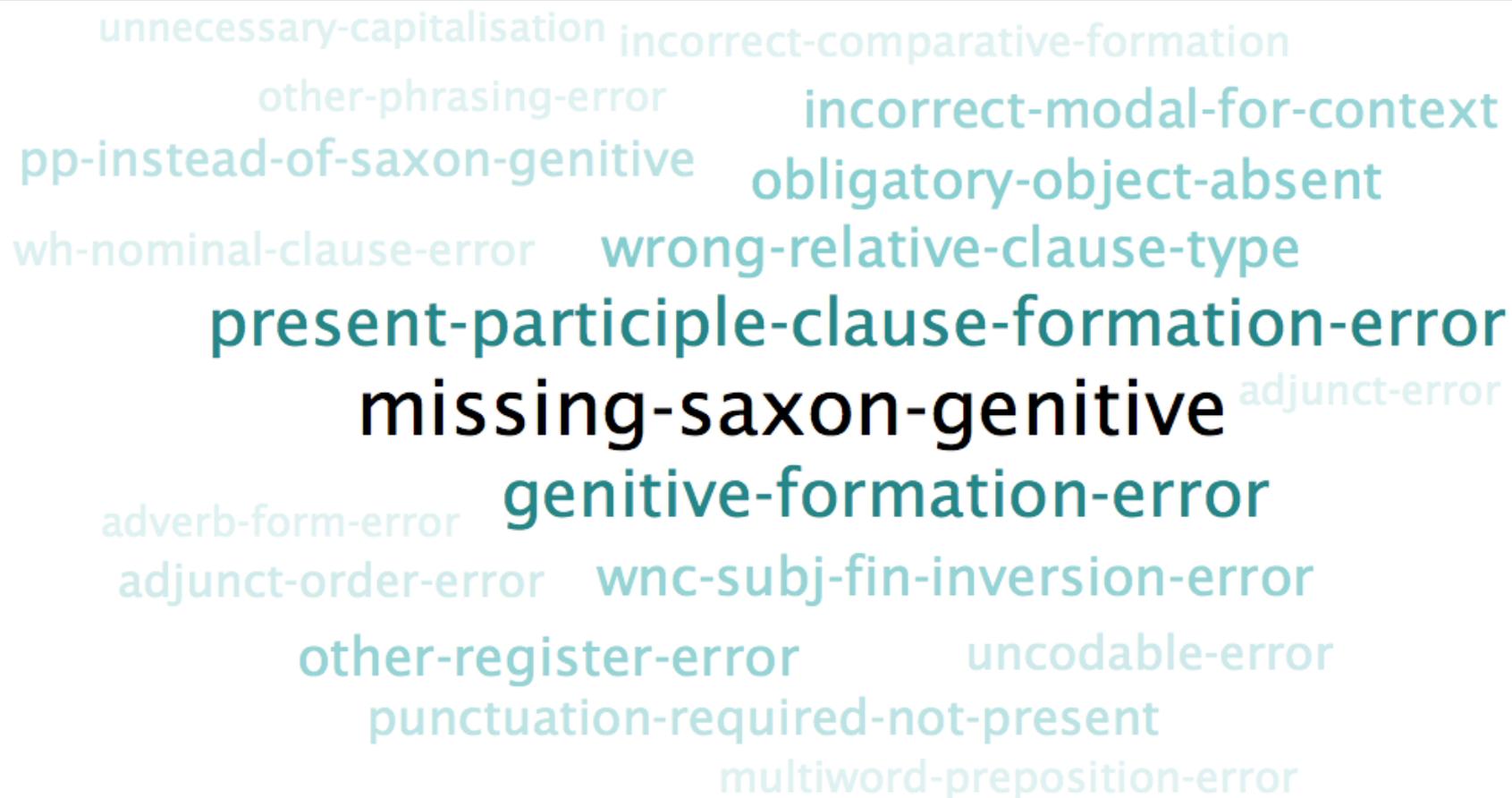
pp-instead-of-saxon-genitive obligatory-object-absent

wh-nominal-clause-error wrong-relative-clause-type

present-participle-clause-formation-error

missing-saxon-genitive adjunct-error

genitive-formation-error

adverb-form-error

adjunct-order-error wnc-subj-fin-inversion-error

other-register-error uncodable-error

punctuation-required-not-present

multiword-preposition-error

But just looking at errors is not enough!

- Students may avoid certain structures, so they don't appear as errors.

    – E.g., Phrasal verbs don't appear much in our error counts.

    – Students seem to avoid them, using alternative forms.

- Alternative approach: compare what they do syntactically to what natives do.

We thus push a two-pronged approach:

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

Automatic syntactic tagging of corpus to see what structures students are **attempting**; and what they are **avoiding** using

Only both together give the full picture!

The new points system for driving offences will be established in Spain before summer o▲

Subject							Mod	Pass	Pred	Adjunct		
Deict	Epith	Thing	Thing	Qualif						Op	Pphead	Op
				Op	Pphead						Thing	
					Classif	Thing						

Thing

O

With this new system , the driving licence will consist of a number of points that c

Adjunct		Sep	Subject			Mod	Pred	Adjunct			Conj
Op	Pphead		Deict	Classif	Thing			Op	Pphead		
	Deict	Epith	Thing						Deict	Thing	Qualif
										Op	Pphead
											Thing

I personally agree with the establishment of this new law , as I feel tha

Subject	Adjunct	Pred	Adjunct			Sep	
Thing	Head		Op	Pphead		Conj	
				Deict	Thing	Qualif	Subject	Pred
					Op	Pphead	Thing
					Deict	Epith	Thing	Cor

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

Assigned

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

Gloss

Comment:

TENSE	FINITENESS	VERB-TYPE
simple-present	simple-finite	intranstive-verb
present-perfect	finite-with-connector	monotransitive-verb
present-progressive	relative-clause	ditransitive-verb
simple-past	that-clause	ergative-verb
past-progressive	wh-nominal-clause	relational-verb
past-progressive	infinitive-clause	verbal-verb
simple-modal	pres-participle-clause	mental-verb
modal-perfect	past-participle-clause	
modal-progressive		

MODALITY	DO-INSERTION	POLARITY
nonmodal-clause	do-inserted	positive-polarity
true-modal-clause	no-do-inserted	negative-polarity
future-clause		

PROCESS TYPE	VOICE	MOOD
material-clause	active-clause	declarative-clause
verbal-clause	passive-clause	imperative-clause
mental-clause		interrogative-clause
relational-clause		

- UAM CorpusTool 3.0 adds automatic Transitivity analysis with participant roles (not yet released)

Gingrich	launched	a blistering attack	on Romney	in what	is	essentially
Actor	Process	Goal	Circumstance			

				Carrier	Process	Circumstance

Syntactic Analysis

- **Grammatical profiles**: the degree to which each proficiency level uses each kind of structure

- From these we can get an idea when teaching a structure is beneficial.

Use of Passive
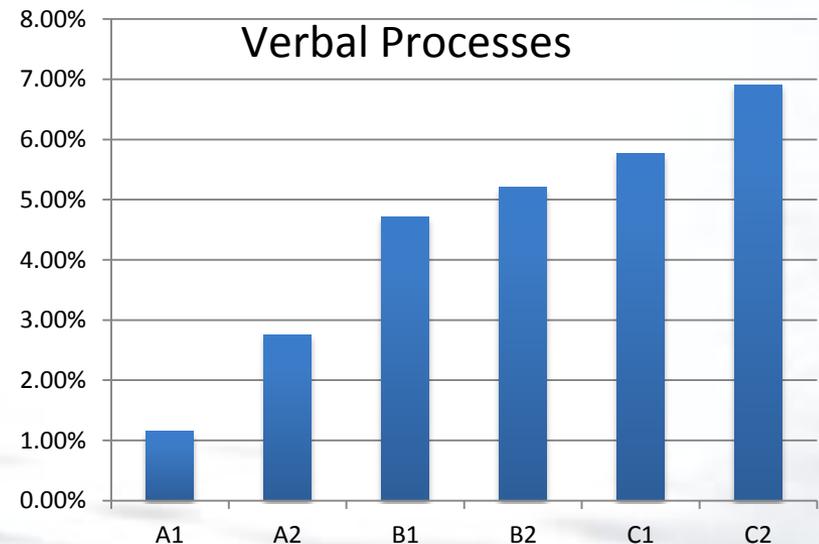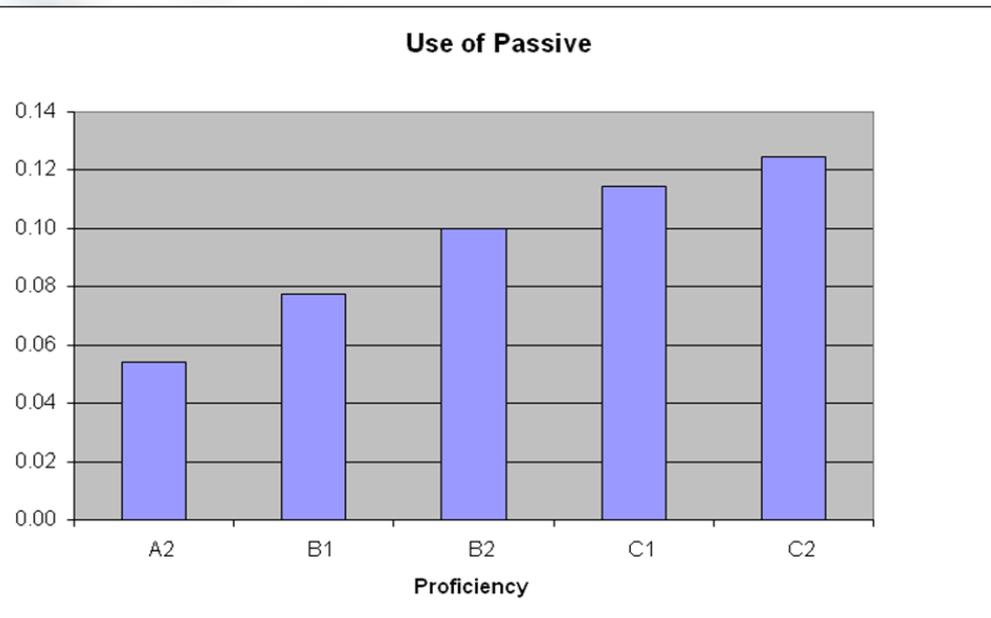
Verbal Processes

5. Results (iii): Verbal Processes

- Verbal Passives: very clear increase in passive with verbal processes! Up to 26%!!!

- Main increase in "It could be argued that..." type structures (postponed Verbiage Subject)

- Students learning to distance themselves from their claims.

verbal-passive

	A2	B1	B2	C1	C2
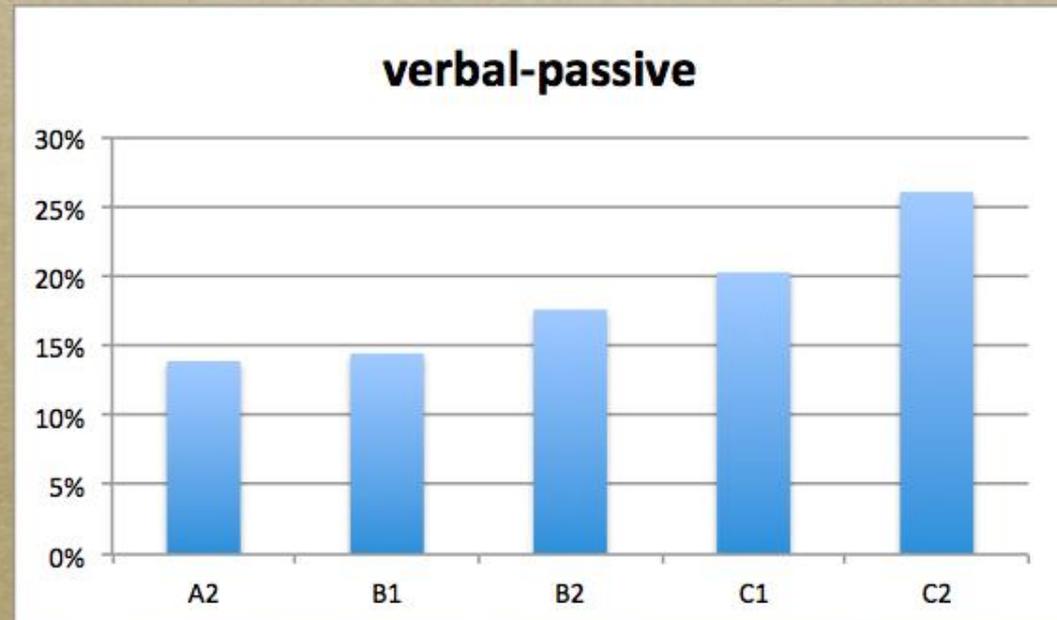
2. Using a Learner Corpus

- INTERVENTION: As a result of these studies, we have already changed the way we teach our first year students.

  ➢ Earlier emphasis on Nominal groups, leaving clause grammar for a bit later.

  ➢ Explicit teaching of the criteria for including or not an article in front of nouns

  ➢ Exercise sets to take home on determiners and preposition usage.

  ➢ Etc.
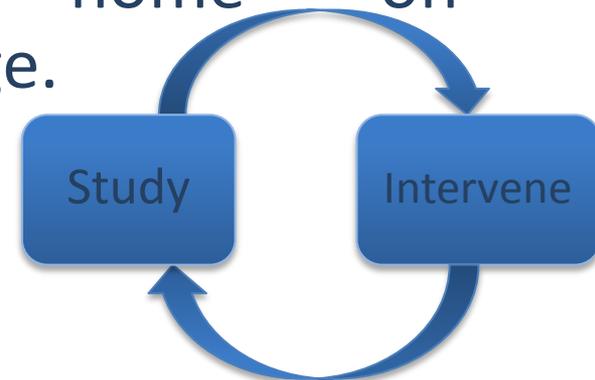
Study

Intervene

2. Using a Learner Corpus

How to USE a learner corpus?

1. Discovering your student's needs

2. Curriculum design: discovering when material is best taught across a curriculum
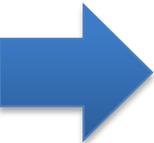
3. As a source of student exercises.

4. Studying classroom interactions

Deriving student exercises from a learner corpus

- Corpus software allows you to search for student sentences with a particular error type, or with a particular grammatical structure.

CorpusTool 2.8.10

| Project | Search | AutoCode | Statistics | Explore | Options | Help |

Enter Search Query Below:

| sentence | + | containing | anywhere | determiner-present-not-required | + | Show | Show a | Save | Help |

P Only Partially Coded ***** Segment with Comment

867 matches

	P	C	Pretext	Contained	PostText
🔍	F		n see that the amount of cars within	the	population is increasing.
🔍	F		to me that if people used more often	the	public transport there would be less traffic in the roads.
🔍	F		lthough a lot of people already use	the	public trasport normally , it is not enough because there
🔍	F		The aim of those delays is that	the	cars do not go in at the same time in the city.
🔍	F		So that the problem of	the	traffic is still there.
🔍	F		There are also problems like	the	public transport strikes.
🔍	F		ough they have cars, prefer to take	the	public transport, take the car instead because of the strik
-	F	-	o take the car have to be waiting in	the	traffic jams for hours because of the big amount of cars i
🔍	F		Summing up, I think that	the society	is very concerned about this problem because it is some

2.2.1 Exercises on determiners (iii)

What is wrong with this text:

In the last months, in Spain, many people were asked about the most important social problems in our society. Some people think the main worrying social issue is the terrorism; other people think that the problem is the drugs.

However, some sociological investigations realized in Spain have demonstrated that 51% think that the unemployment is the worst problem, the next is the price of housing and the third important problem is the domestic problem.

2.2.1 Exercises on determiners (iii)

What is wrong with this text:

In the last months, in Spain, many people were asked about the most important social problems in our society. Some people think the main worrying social issue is ~~the~~ terrorism; other people think that the problem is ~~the~~ drugs.

However, some sociological investigations realized in Spain have demonstrated that 51% think that ~~the~~ unemployment is the worst problem, the next is the price of housing and the third important problem is the domestic problem.

2.2.1 Exercises on determiners (iii)

Fill in each gap with either "a", "an", "the" or "-"

In ＿＿＿ recent months in Spain, many people were asked about ＿＿＿ most important social problems in our society. Some people think ＿＿＿ main issue is ＿＿＿ terrorism; ＿＿＿ other people think that ＿＿ problem is ＿＿＿ drugs.

2. Using a Learner Corpus

- Build online exercises using Hot Potatos, Moodle, etc.

- Questions relating to the highest error categories in your class:

6 ✍	Which of the following sentences are grammatically correct?	
Marks: -- /2	Choose at least one answer.	☐ All my two brothers are coming.
		☐ I like no of them.
		☐ He has much chance of winning.
		☐ I have seen others which I like less.
		☐ I bought five clothes at the sale.
		☐ Each of them have good qualities.

2. Using a Learner Corpus

- **Data-driven learning**: give students access to the error tagged corpus, and ask them to derive the rules for use of the grammatical category

CorpusTool 2.8.10

| Project | Search | AutoCode | Statistics | Explore | Options | Help |

Enter Search Query Below:

sentence + containing anywhere determiner-present-not-required +   Show   Show a   Save   Help

P Only Partially Coded * Segment with Comment   867 matches

	P	C	Pretext	Contained	PostText
🔍	F		n see that the amount of cars within	the	population is increasing.
🔍	F		to me that if people used more often	the	public transport there would be less traffic in the roads.
🔍	F		lthough a lot of people already use	the	public trasport normally , it is not enough because there
🔍	F		The aim of those delays is that	the	cars do not go in at the same time in the city.
🔍	F		So that the problem of	the	traffic is still there.
🔍	F		There are also problems like	the	public transport strikes

2. Using a Learner Corpus

How to USE a learner corpus?

1. Discovering your student's needs

2. Curriculum design: discovering when material is best taught across a curriculum

3. As a source of student exercises.

4. Studying classroom interactions

2. Using a Learner Corpus

Studying classroom interactions

T: Before we go on, can you um tell me how I decided when to start a new paragraph

Chn: [hands up]

T: because that's something that some people at the beginning of the year had a bit of trouble with knowing when to begin a new paragraph. Adrian, what do you think?

A: Um, when you like change the subject or like you, you go on time, place, time that just happened, or … that's about it.

T: That's excellent. That's a really good answer. Does someone want to add to that? Can you think of something else to say. Okay Julia

2. Using a Learner Corpus

Studying classroom interactions

- To see which teacher strategies are most effective.

- To identify strategies used by successful students which could be taught to other students

- To identify students who participate and those who don't.

- Etc.

2. Using a Learner Corpus

Exploring effective teacher strategies :

1. Do good teachers ask more questions of students than other teachers?

2. Do students learn better when teachers give supportive feedback?

3. Do good teachers frame their questions by first providing background information?

Etc.

3. Using Native Corpora

Some NATIVE corpora

- **English:**
  - **BAWE (British Academic Written English):** 2760 essays from University students across different faculties and years. Very good reference corpus for university-level learners of English.

    http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/
  - **MICASE**: Transcripts of Spoken academic events.

- **Spanish (Chile):**
  - **PUCV-2006** (Giovanni Parodi, Rene Venegas): (481 texts, 60 million words)

- **Portugese (Portugal):**
  - **DEEB** (Carlos Gouveia, Lisbon) 4,800 Portuguese exam texts from 4 school years.

3. Using Native Corpora

Why use Native Corpora:

- **Studying core vocabulary**: Studying the vocabulary and phrases that are most essential in a given genre or domain.

    - "in other words", "this essay will address"

- **Reference corpus**: Studying the normal patterns of grammatical usage of native writers as a means of locating under-usage and over-usage of your students.

- **Data-driven learning**: teaching students to use a native corpus as a resource to discover rules of correct/preferred usage.

Studying core vocabulary: Keywords

- Discovery of "keywords" in a given target text type, as a starting point for Vocabulary classes.

- E.g., what are the key words to teach Computer Engineers?

Studying core vocabulary: Key Phrases

- What common phrasings will help a learner appear to be competent in a register or genre?

  – N-grams, "key phrases", "lexical bundles"

Table 6
Most frequent 50 4-word bundles in four disciplines (Bold = item occurs in 4 disciplines, italic = items occurs in 3 disciplines)

Biology	Electrical engineering	Applied linguistics	Business studies
in the presence of	**on the other hand**	**on the other hand**	**on the other hand**
in the present study	as shown in figure	**at the same time**	**in the case of**
on the other hand	**in the case of**	*in terms of the*	at the same time
the end of the	is shown in figure	on the basis of	*at the end of*
is one of the	it can be seen	in relation to the	on the basis of
at the end of	as shown in fig	**in the case of**	**as well as the**
it was found that	is shown in fig	in the present study	the extent to which
at the beginning of	can be seen that	*the end of the*	*the end of the*
as well as the	*can be used to*	*the nature of the*	significantly different from zero
as a result of	the performance of the	*in the form of*	are more likely to
it is possible that	as a function of	**as well as the**	the relationship between the

3.2   USES of Native Corpora:

– Studying core vocabulary: Studying the vocabulary and phrases that are most essential in a given genre or domain.

- "in other words", "this essay will address"

– **Reference corpus**: Studying the normal patterns of grammatical usage of native writers in the same genre as a means of locating under-usage and over-usage of your students.

– **Data-driven learning**: teaching students to use a native corpus as a resource to discover rules of correct/preferred usage.
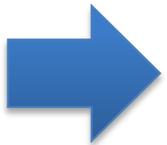
Data Driven Learning

- **Key idea**: learners learn best by being confronted by real examples of text, and through repeated exposure, come to understand normal usage of vocabulary or grammar.

- Reading is good, but exposures to instances of vocabulary/syntax structures may be so far apart that previous exposures are forgotten.

- A concordancer can show a learner multiple real contexts of vocabulary or structures at one time.

- Learner can use this information to figure out meanings/usage rules.

Data Driven Learning

1	sing itself to such possibilities,	however	, A Bend in the River, for all its a
2	rine of diminished responsibility.	However	, a central element in the common-la
3	targets by wearing a Celtic scarf.	However	, a closer look at the supposedly mo
4	probably unhelpful (Simpson 1976).	However	, a cognitive therapy approach to he
5	. Even apart from such exceptions,	however	, a contract not made in the require
6	ischarge in that area. More often,	however	, a discharge will have been noticed
7	ferent countries. From 1860 on,	however	, a dramatic increase in the number
8	ient in committing suicide. It is,	however	, a fine line. between aiding suicid
9	eir anthropology. It also implied,	however	, a further difference from the earl
10	may well seem perverse. There is,	however	, a good deal of information in the
11	ir and construction of body cells.	However	, a high-protein diet need not be ex
12	cribed to mere accident. There is,	however	, a large variety of killings which
13	ndment Act to be unconstitutional.	However	, a law dealing with the national si
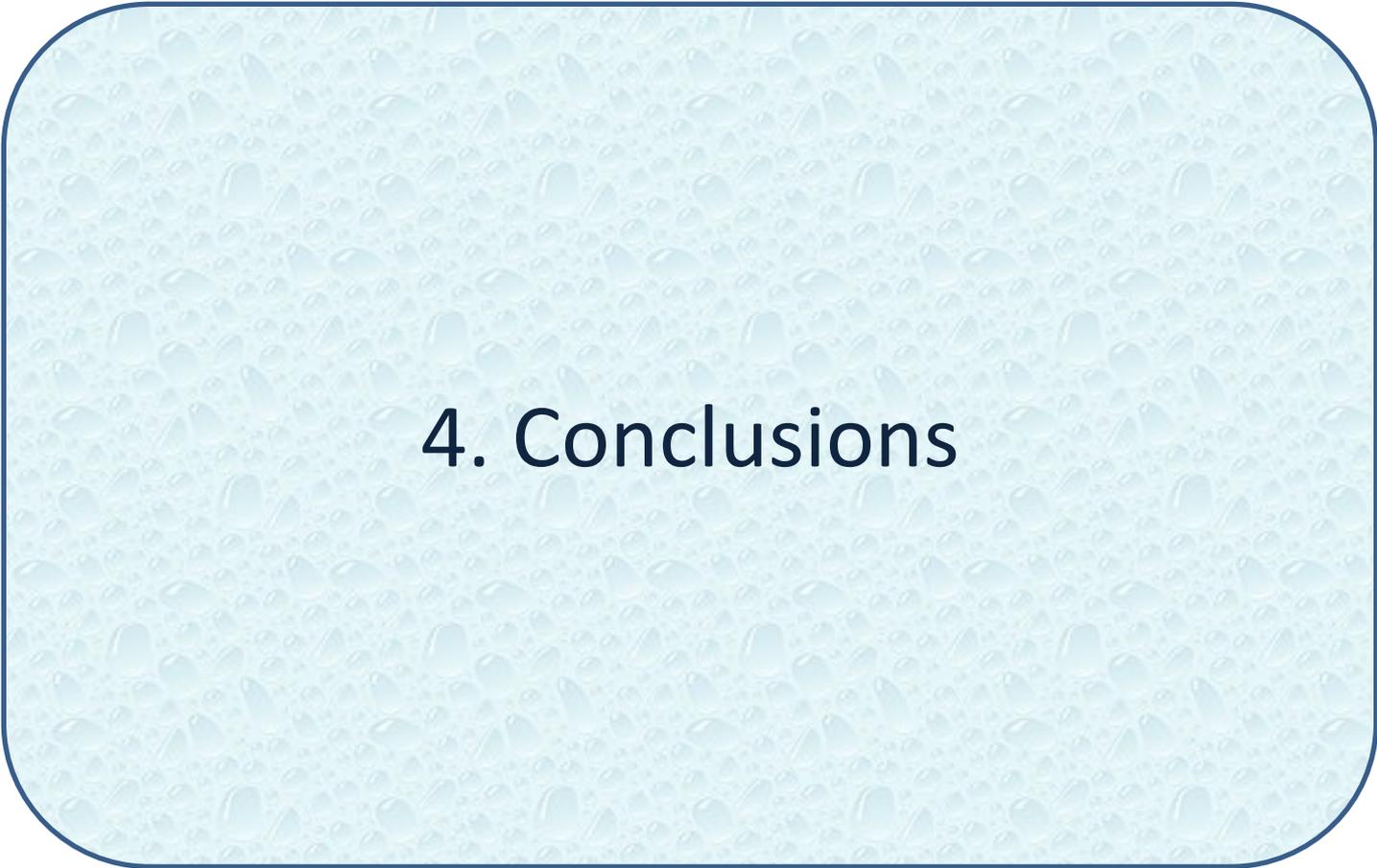
4. Conclusions

4. Conclusions

- Language Learners from different mother-tongue contexts face different needs from those from other language backgrounds.

- Many resources for FLT are targeted at a generic student, and don't directly address the needs of specific students

- We can use Learner Corpora for our specific community in order to discover the specific needs of our students

- We can re-shape our course materials to better target these specific needs

4. Conclusions

- Error analysis is a good way to clearly see what problems our students have

- Studies across proficiency levels allow us to see where certain topics become more critical.

- Automatic grammatical analyses allow us to profile grammatical abilities at each level.

- We can thus reformulate our teaching curriculum to best suite the needs of our students.