Using a learner corpus to explore evolving learner proficiency

Mick O'Donnell
Universidad Autónoma de Madrid

Aim

- This talk will describe work using a corpus of Learner English to study second language development
  - How do Spanish University learners of English develop their use of English linguistic resources as they move towards native competence
  - To discover indications as to where we need to focus our teaching of grammar.
- Using various forms of automatic corpus annotation

Motivation

- A better understanding of the process of how learners acquire a second language can help us better design both traditional and online learning systems.

- More concretely, the study can tell us which areas of the language are most critical for a learner from a particular mother-tongue/target language pair.

  - **Critical language features**: defined quantitatively, the aspects of language which most frequently are problematic for the pool of learners.

Methodology

Two pronged attack:

1. Study the **errors** of learners to identify the linguistic features that learners most frequently get wrong.
   - ➡ Treacle Project (16,600 errors coded and studied)

2. Study the **patterns of grammatical choices** made by learners, to identify overuse and underuse of particular features (Present talk)
   - ➡ Non-use may identify lack of acquisition
   - ➡ Over- or under-use may relate to different contexts of use between languages.

Corpus

- **Learner corpora**:
  - **WriCLE Corpus** (Rollinson & Mendikoetxea, 2010) 556 essays by Spanish University learners of English
    - The **UPV Learner Corpus,**150,000 words of shorter texts by ESP students. (Andreu et al 2010)
- **Native Corpus**: 74 **BAWE** Sociology Essays (similar questions by English natives)

Sample text

Inmigration is a problem that almost every European country must deal with. Specifically in Spain there are one million of inmigrants with documentation, so it have to be more than one an a half actually, including those who have not got papers. The truth is that there are places where inmigration is not a problem for anybody but there are other places where people think foreigners will let them without work; or they think they do not want their children to be in the same school as inmigrants. In this essay I am going to discuss the main viewponts about inmigration in Spain.

To begin with, there are some people who believe that inmigrants make our society grow up, so they are in favour of inmigration in this country. Many people think that race variety might be a way to build a world without wars.

4. Methodology

Can learner English be parsed reliably?:

- Actually, yes, with something like 80% reliability on each clause feature (some more, some less)
- This is enough to see trends.
- Each level has its own problems:
  - Low level learners make more lexical and grammar mistakes, which may throw the parser
  - Higher level learners write better text but write longer sentences, which are harder for the parser to parse.

Proficiency Level

- **Proficiency level** (A1, A2, …, C2) associated with each learner essay from placement test. (Oxford Quick PT)

Related Work

Study is in a sequence of studies of learner language using the same learner corpus:

Error Analysis
(e.g., MacDonald et al. 2011)

Lexical Errors
(Mediero Durán 2013)

Modality
(Garcia 2012)

Transitivity
(O'Donnell 2012)

Tense-Aspect
(O'Donnell 2013)

Theme
(O'Donnell 2014)

Article Errors
(Dotti/O'Donnell 2014)

Word choice errors
(Dotti 2014)

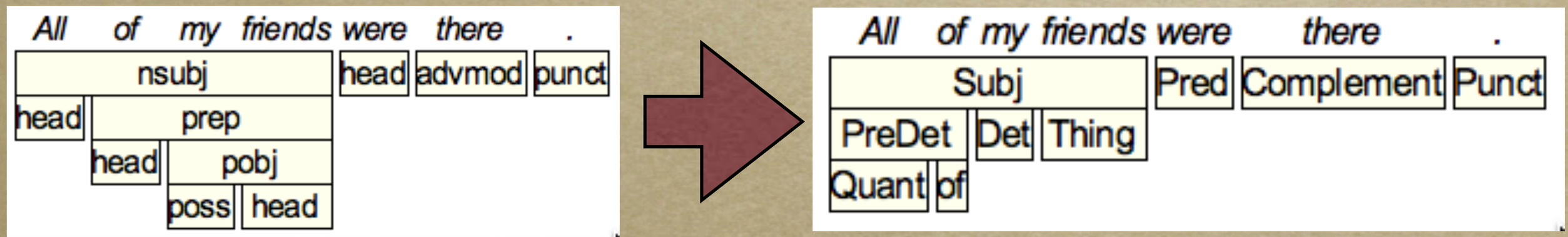
Quantification
(O'Donnell 2015)

3. The Data: Annotation

- All texts automatically parsed within UAM CorpusTool (O'Donnell, 2008)
- Uses Stanford Parser (Klein and Manning 2003) to syntactically annotate each tree.
- Stanford parse is transformed into a richer corpus annotation:
  - Transformation towards more appropriate tree structure.
  - Featurisation of linguistic aspects of interest.

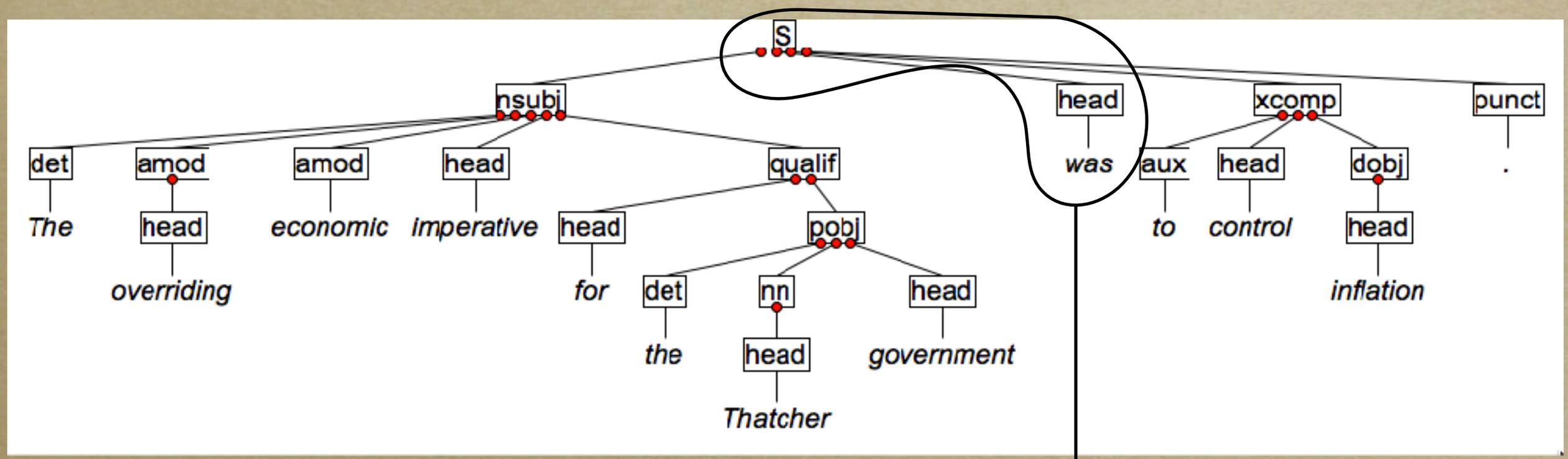
3. The Data: Tree Transformation

- The Stanford parse makes decisions as to syntactic structure which may not correspond to what one wants.
- We thus apply a sequence of tree transformation operations to produce the analyse we need.



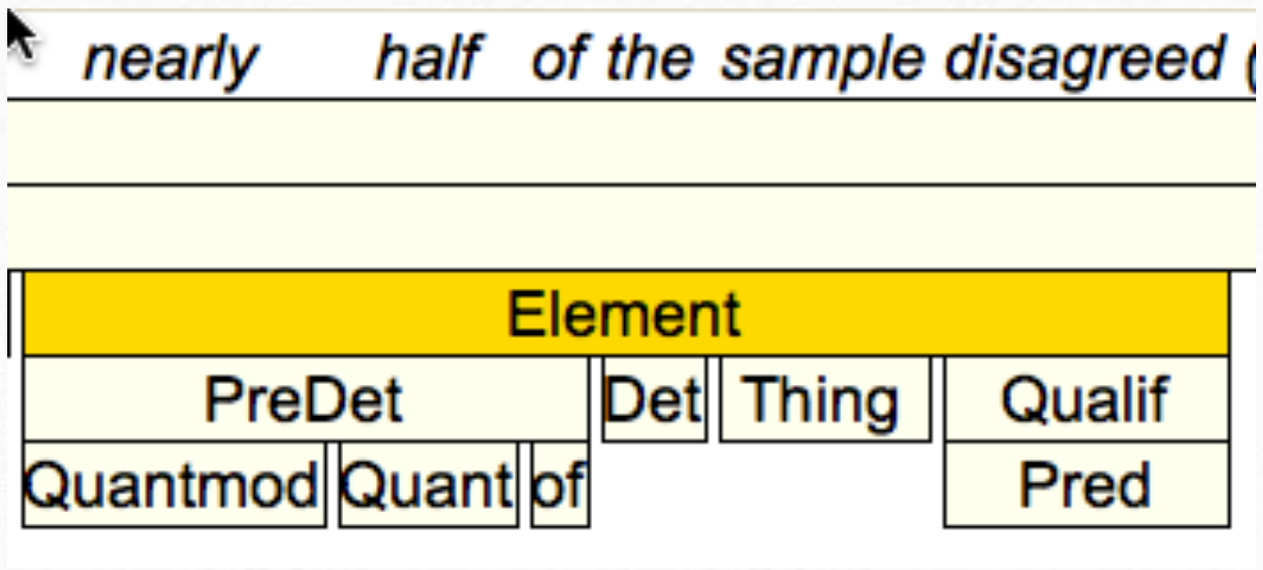
3. The Data: Featurisation

- Syntactic parsers provide only minimal information about each constituent (one class, or one class and one role category):
- For corpus analysis, we often need to 'featurise' the structure, labelling lexico-structural configurations of interest:



Noun Phrase

- **Noun Phrase Structure**:
  - **predetermined** (*all the children, all of the children*)
  - **determiner type** (none, the, many, another, etc.)
  - **premodification** / **postmodification**
  - **Kind**: proper, common, pronominal
  - **Extensive quantification features**
  - **count vs. mass nouns**
  - **abstract vs. concrete nouns**
  - **nominalised heads** (*the run, the dismissal*, etc.)



Quantification

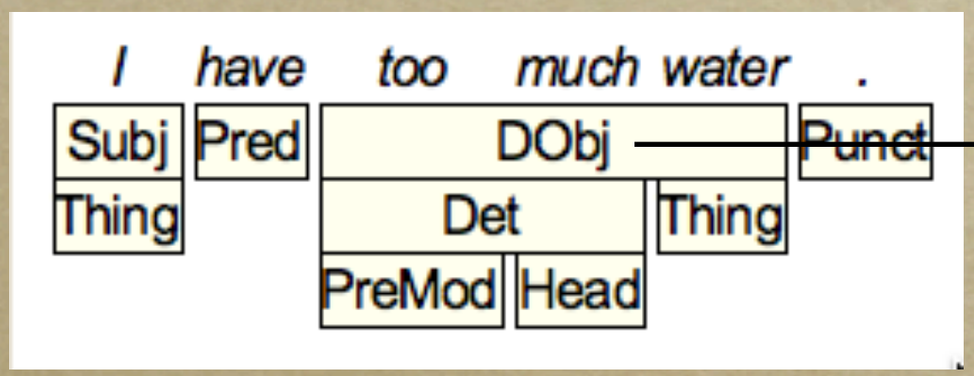
- Use of terms such as “few”, “many”, “much” “a lot of” “all of’, etc.

- A key area of English acquisition for Spanish natives as many intralingual differences:

  - “mucho”-> “much water”; “many apples”
  - “all my friends” but not “many my friends”
  - Complex rules: “X I have any water”
    but ✓”I don’t have any water”

3. The Data: Featurisation

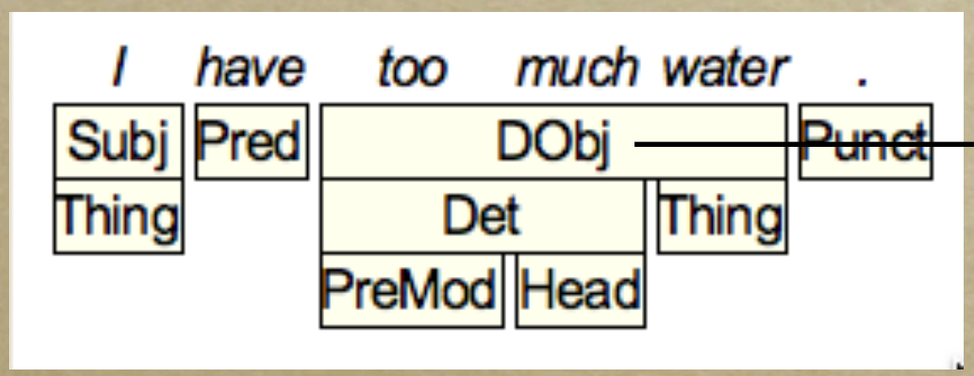
- UAM CorpusTool's internal code supplies features to each nominal group (noun phrase):



group
nominal-group
common-group
not-predetermined-group
determined-group
quantifier-determined
much-determined
not-premodified-group
not-postmodified-group

3. The Data: Featurisation

- UAM CorpusTool's internal code supplies features to each nominal group (noun phrase):



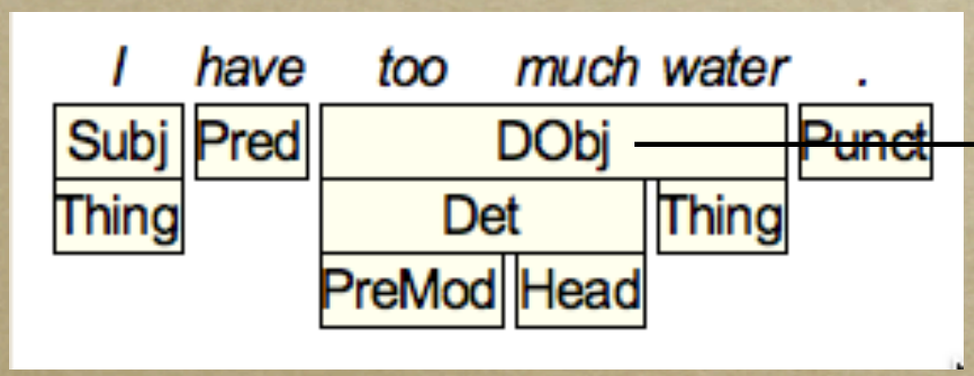
group
nominal-group
common-group
not-predetermined-group
determined-group
quantifier-determined
much-determined
not-premodified-group
not-postmodified-group

Internal Context:

intensified-quantification
noncount-group
singular-group
concrete-group

3. The Data: Featurisation

- UAM CorpusTool's internal code supplies features to each nominal group (noun phrase):



group
nominal-group
common-group
not-predetermined-group
determined-group
quantifier-determined
much-determined
not-premodified-group
not-postmodified-group

Internal Context:

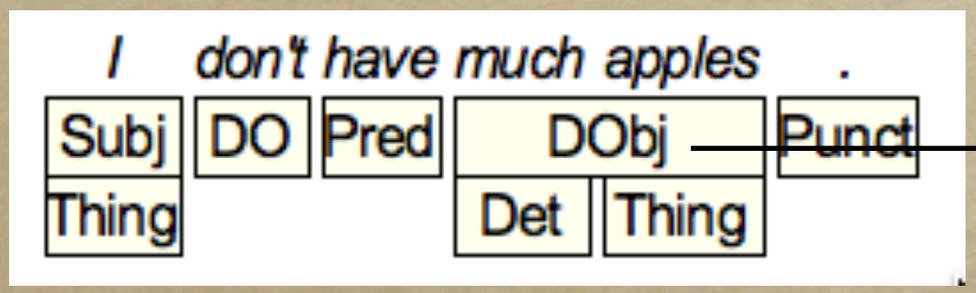
intensified-quantification
noncount-group
singular-group
concrete-group

External Context:

in-positive-context
in-statement

3. The Data: Error Detection

- Quantification Errors also detected automatically:
  (i) agreement errors



group
nominal-group
common-group
not-predetermined-group
determined-group
quantifier-determined
much-determined
not-premodified-group
not-postmodified-group

Error Features

noncount-quant -with-count-noun

Internal Context:

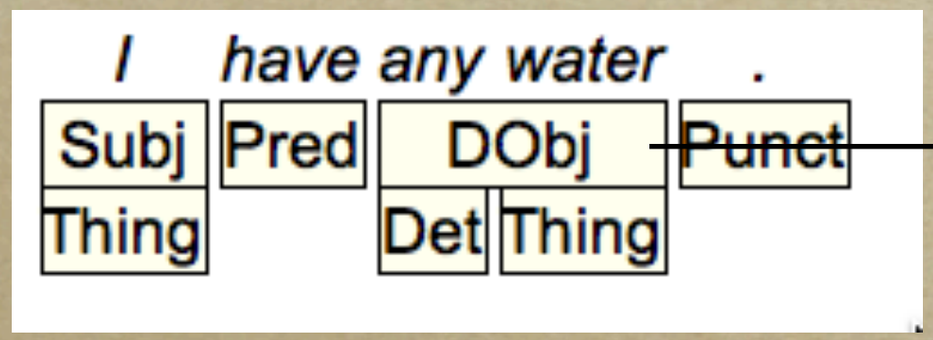
not-intensified-quantif
count-group
plural-group
concrete-group

External Context:

in-negative-context
in-statement

3. The Data: Error Detection

- Quantification Errors also detected automatically:
  (ii) Context errors



group
nominal-group
common-group
not-predetermined-group
determined-group
quantifier-determined
any-determined
not-premodified-group
not-postmodified-group

Error Features: **much-any-in-positive-statement**

Internal Context:
not-intensified-quantif.
mass-group
singular-group
concrete-group

External Context:
in-positive-context
in-statement

3. The Data: Identifying negative

- *Quantifiers like "much" and "any" are possible in negative contexts but not always in simple positive statements:*
  - *X I have much money*
  - *I don't have much money.*

- *UAM CorpusTool searches upwards for any containing constituent which includes negativity:*
  - *"not" in verbal group: I don't have much money.*
    I don't think he has much money.
  - *Negative Subject: Nobody has much money.*
    Neither student has much money.
    None of them has much money.

Part 2: Linguistic Model

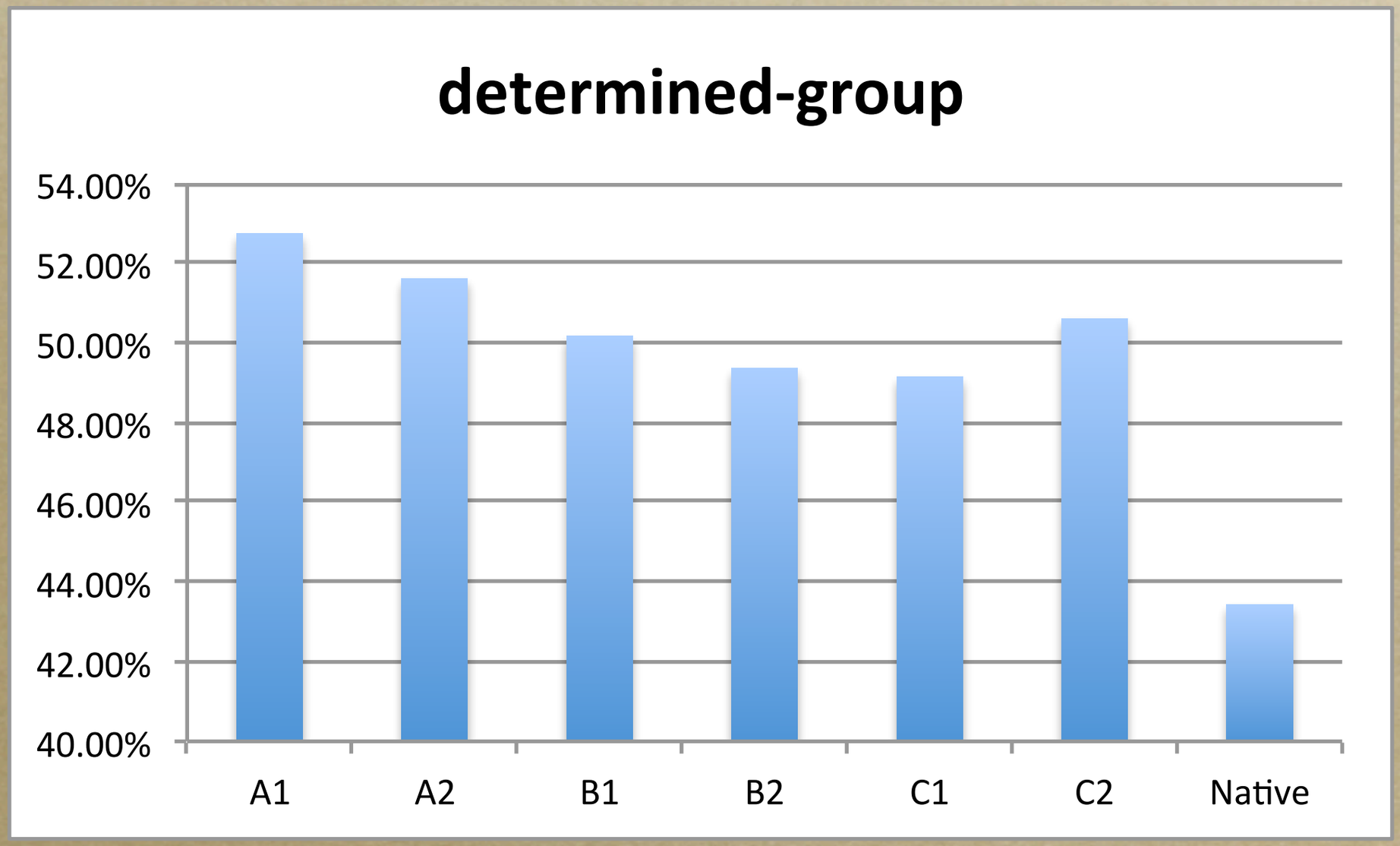
- Ignored quantification in Premodifier slot (open class):
  - *e.g., my two children, seven dogs*

all	*the*	*best*	*jokes*	*in one book*
PreDet	*Det*	*Premod*	*Head*	*PostMod*

Included

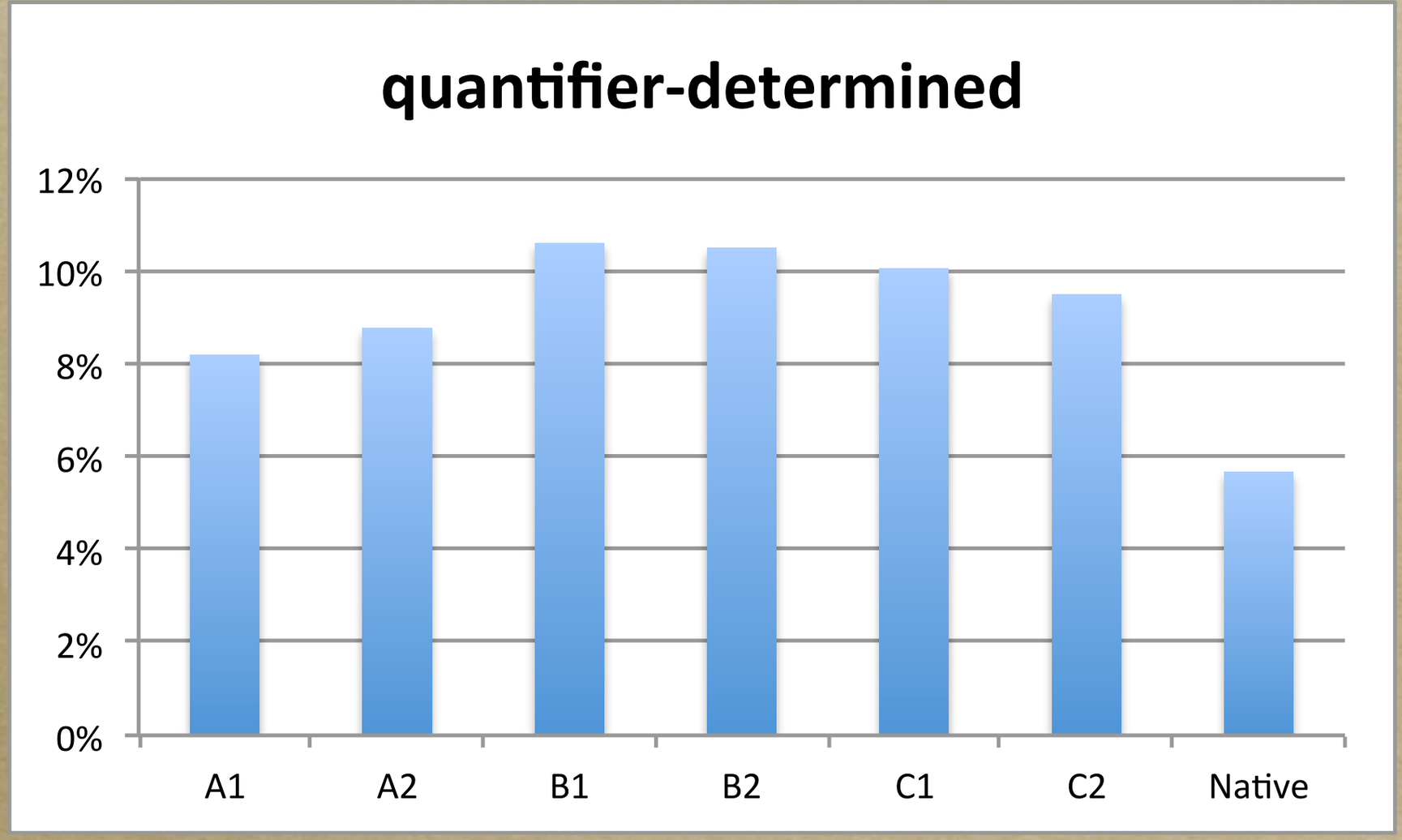
PART 4: Results
(i) Use of Quantifiers in Determiner slot

- Spanish learners over-produce determined common phrases
- (Graph: % of common noun phrases with determiner slot)

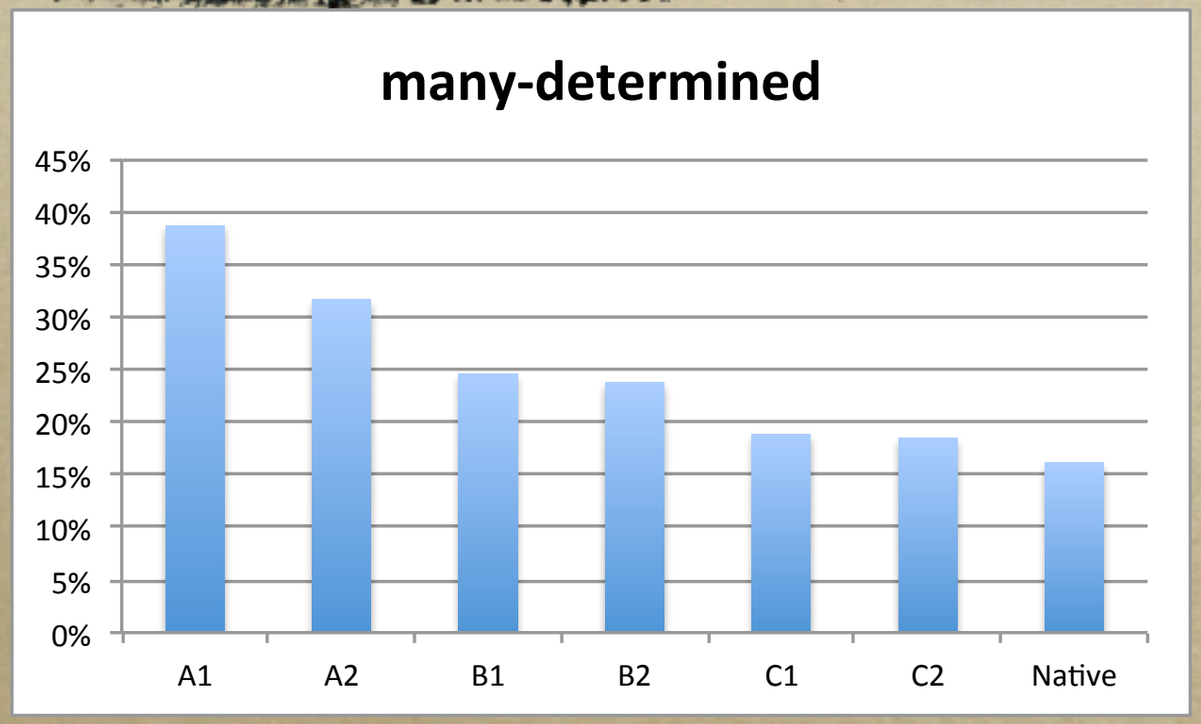


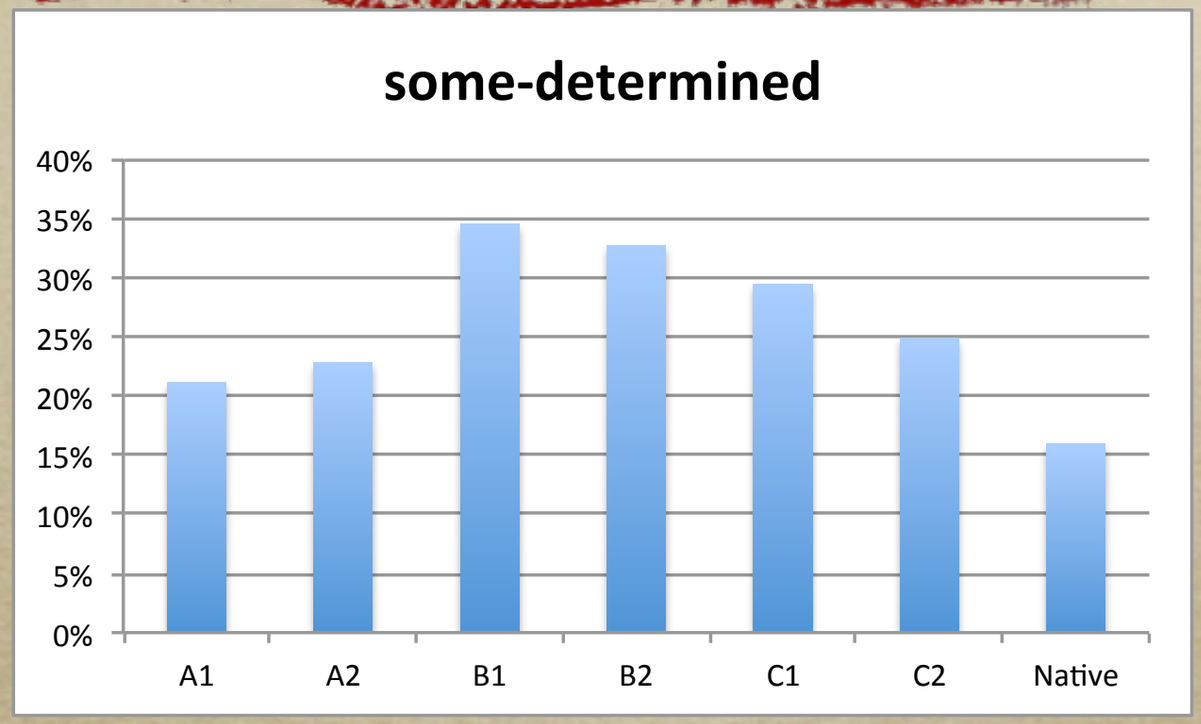
PART 4: Results
(i) Use of Quantifiers in Determiner slot

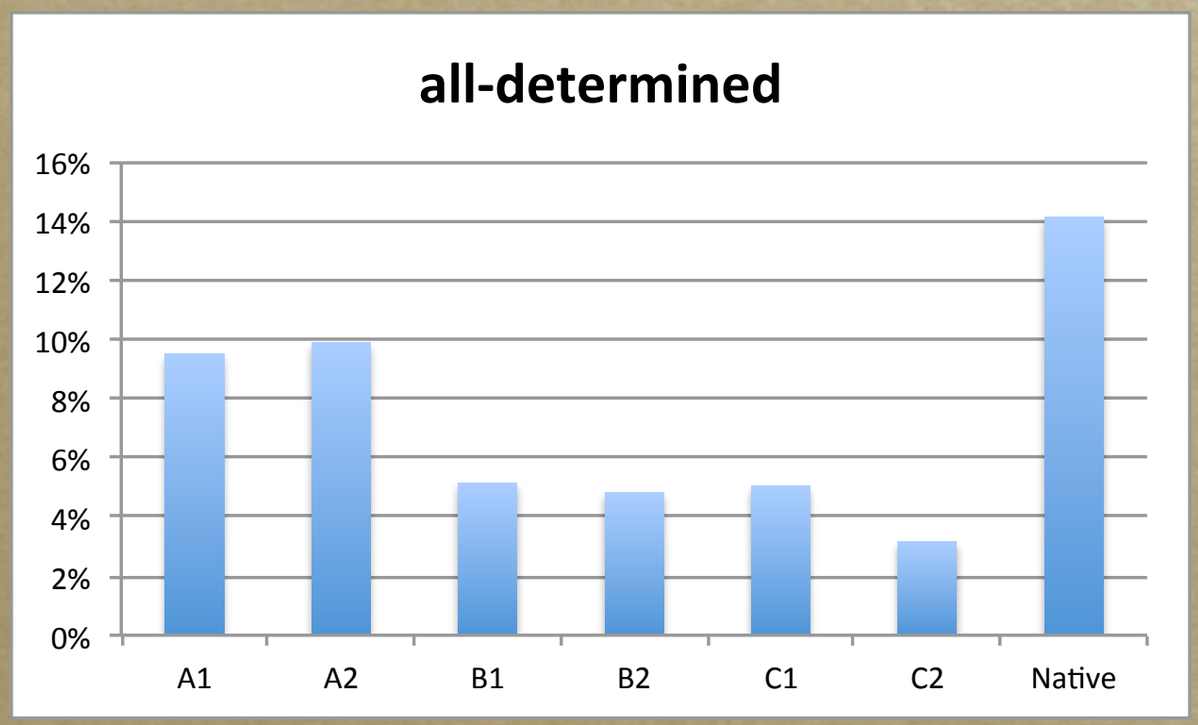
- Of the determined NPs, our learners use more quantifier determination than natives.
- E.g., “both reasons”, “no profit”, “many people”.

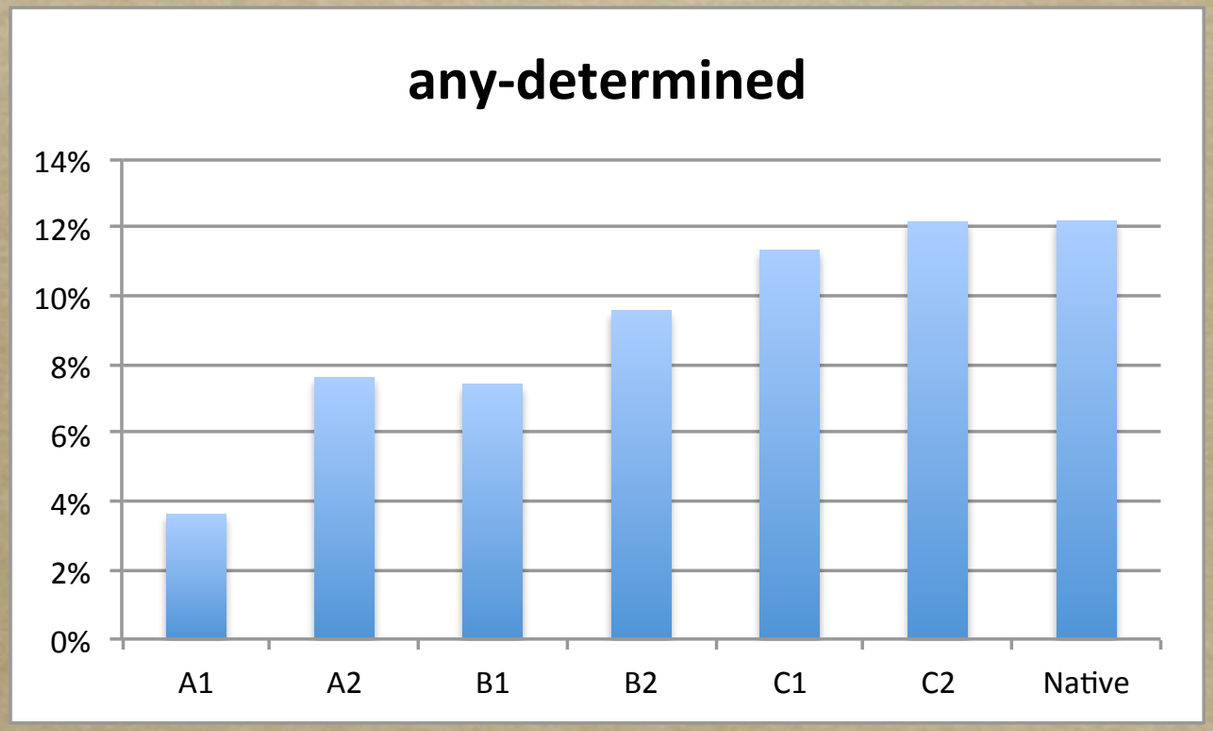


PART 4: Results
(i) Types of quantifier in Det slot

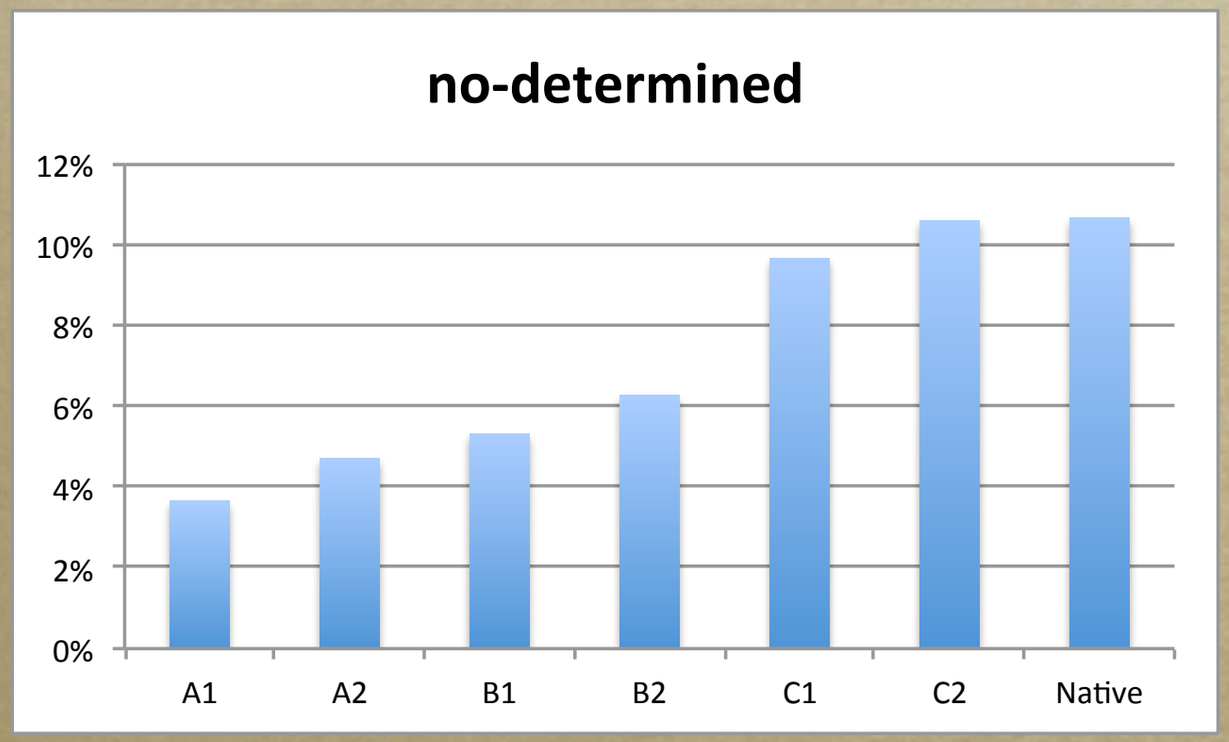


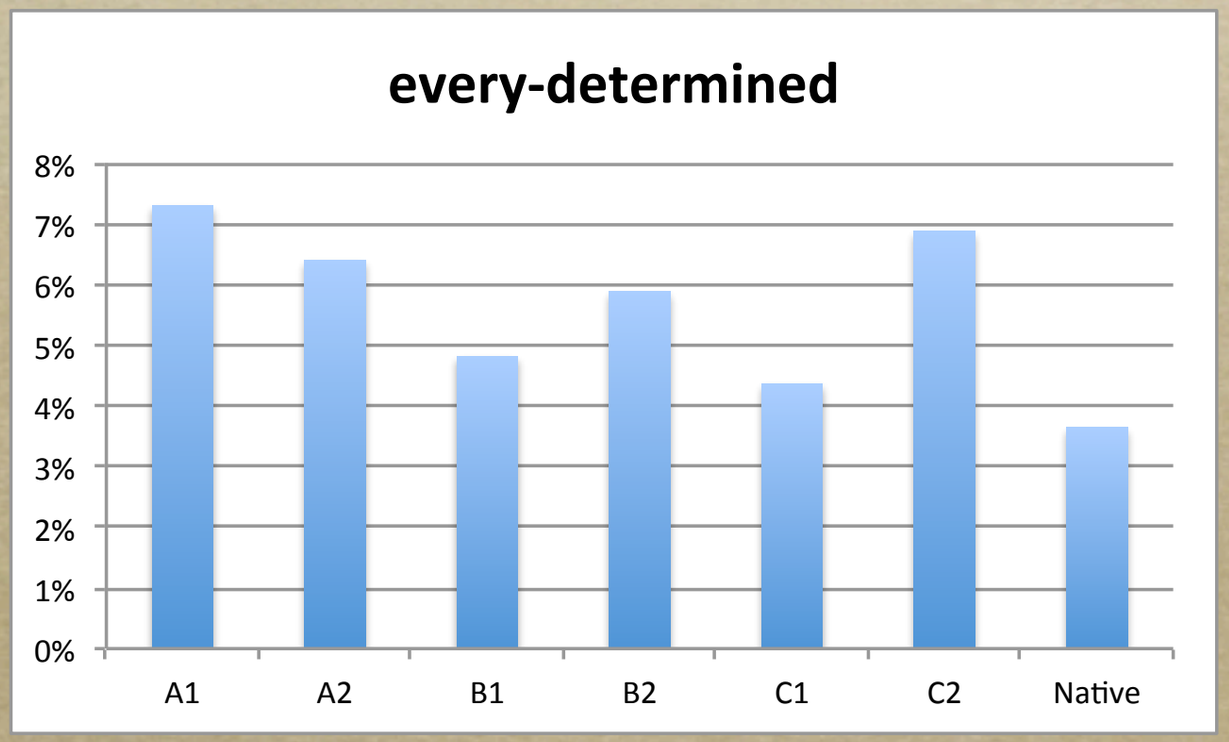






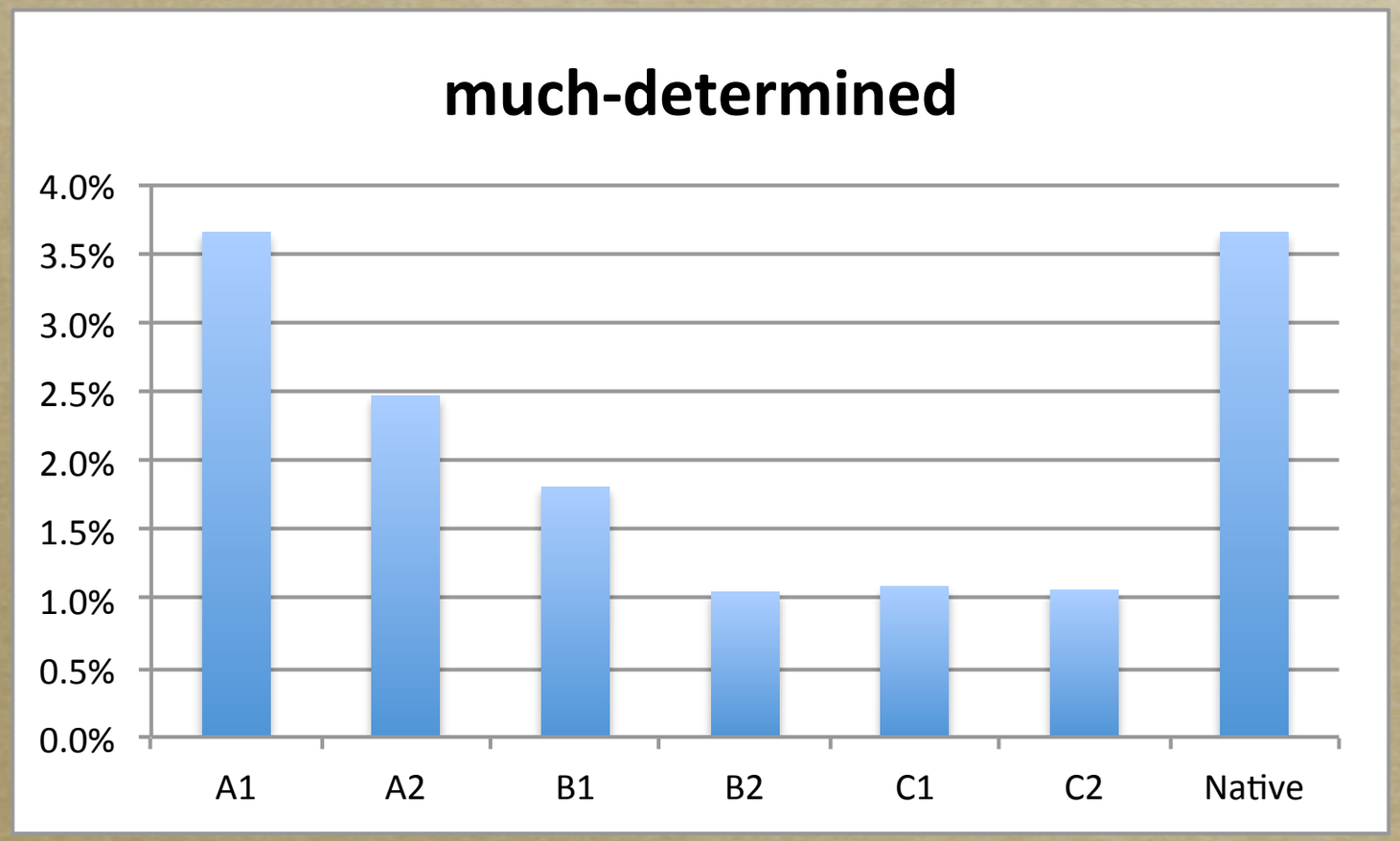
PART 4: Results
(i) Types of quantifier in Det slot





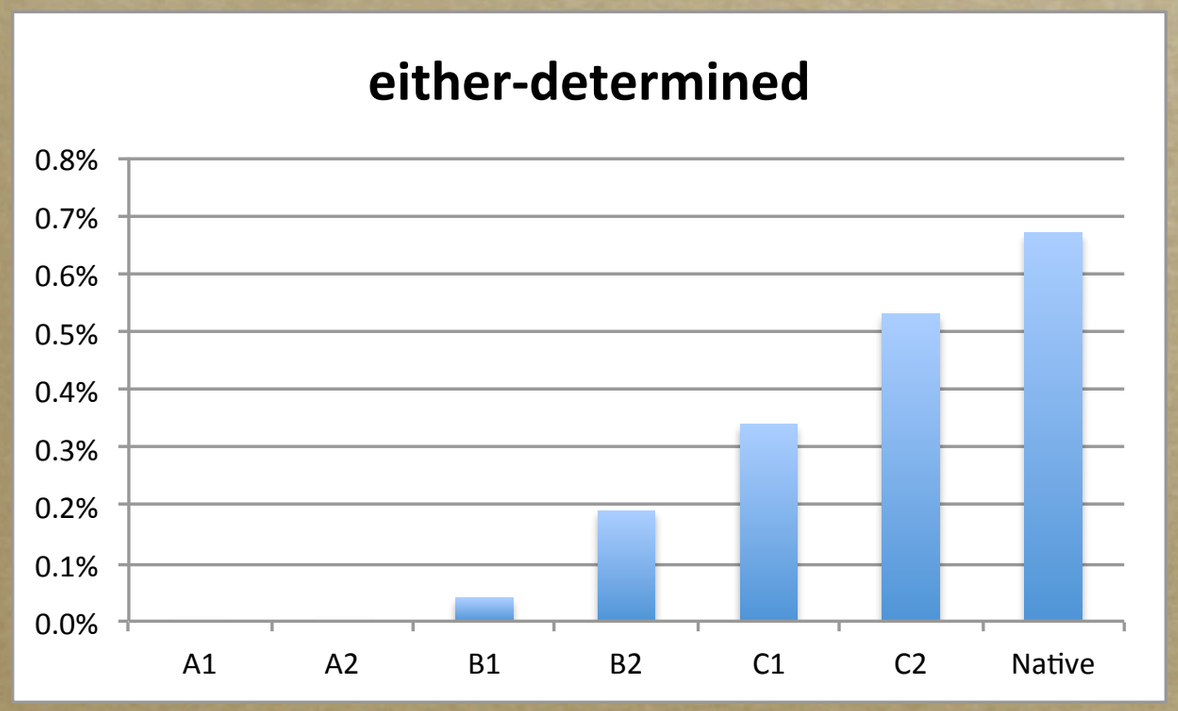
PART 4: Results
(i) Types of quantifier in Det slot

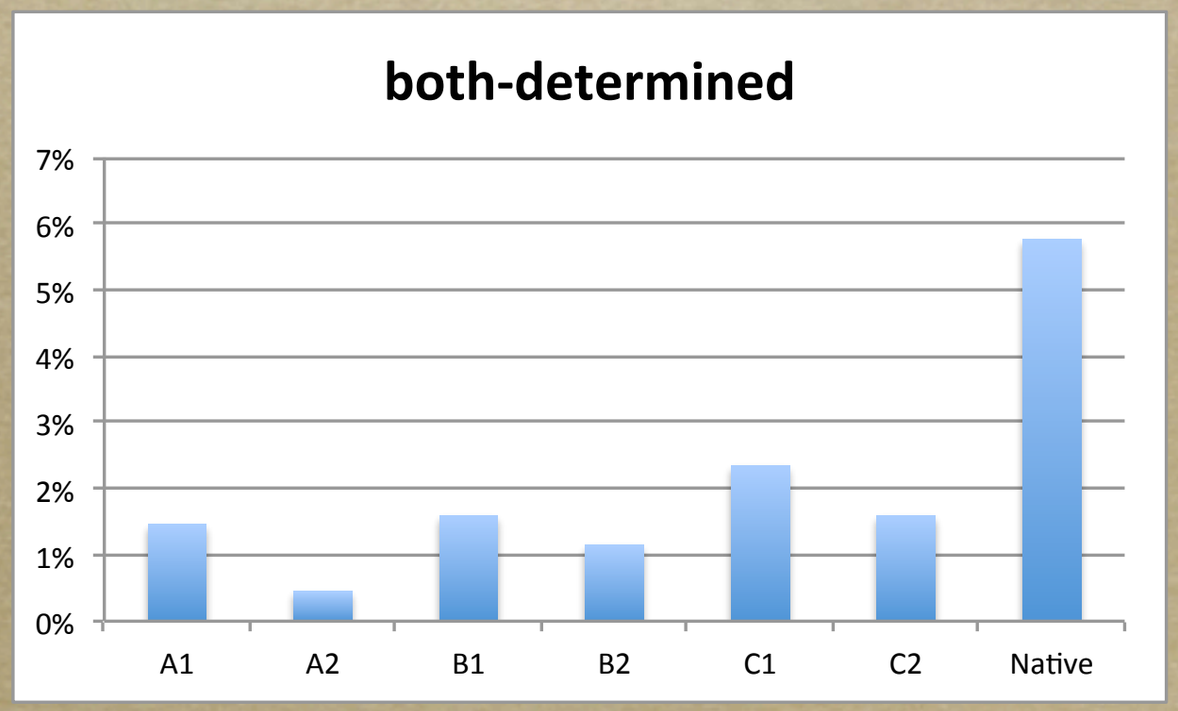
- Special Case: “Much” is wrongly used by many Spanish learners of English, since it has complex rules governing its use.
- Advancing learners appear to avoid using it, to avoid errors.



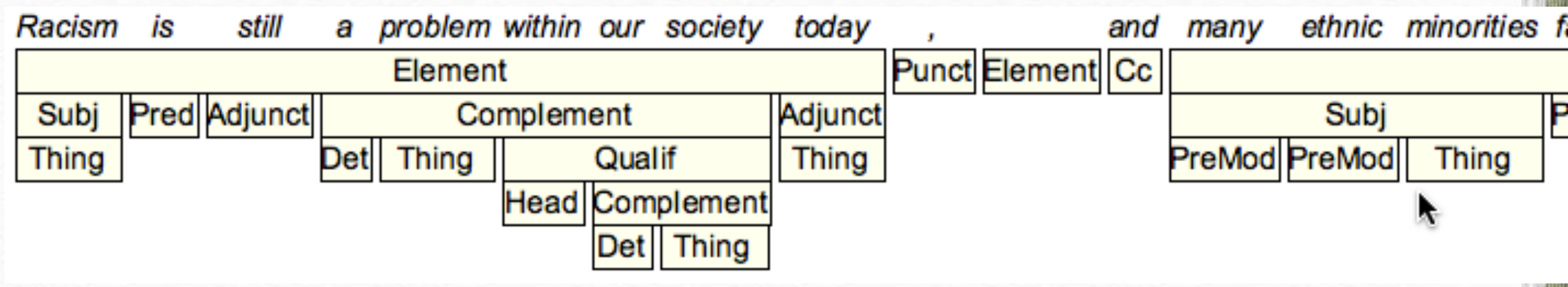
PART 4: Results
(i) Types of quantifier in Det slot

- Dual Determiners: clear that Spanish learners don't use these appropriately.
- While 'either' seems to be acquired with proficiency, both seems not to be properly acquired.



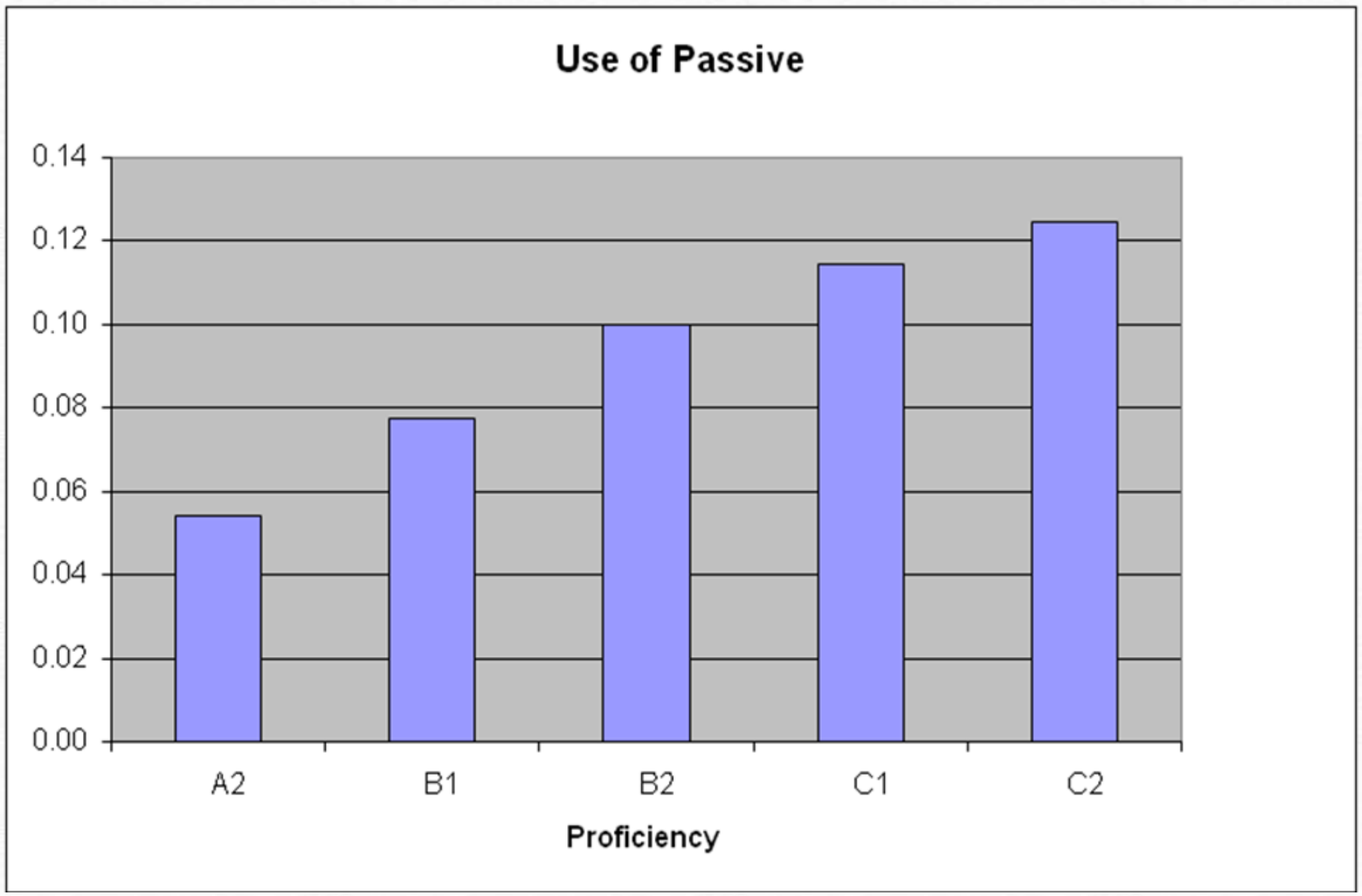


Basic Grammar



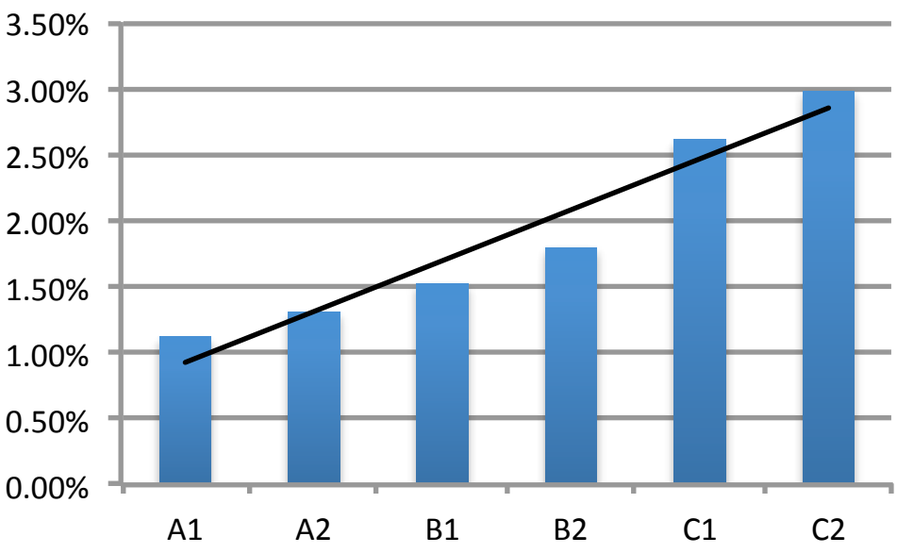
- Clause Features:
  - **Voice** (active vs passive)
  - **Tense-Aspect** (simple-present, past-perfect, etc.)
  - **Mood** (declarative, interrogative, imperative)
  - **Finiteness** (finite, infinitive-clause, past-participle-clause, present-participle-clause, relative-clause, that-clause, etc.)
  - **Marked Sentence Structure**: it-cleft, extraposition, there-existential, etc.

Voice


Use of Passive
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
A2
B1
B2
C1
C2
Proficiency

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

Rising Usage



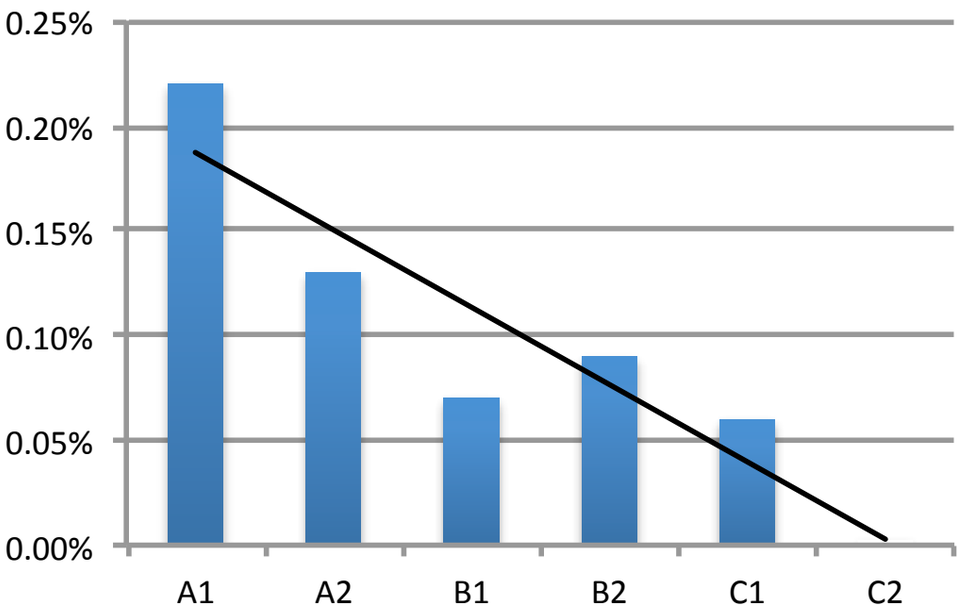
Use of passive

- 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

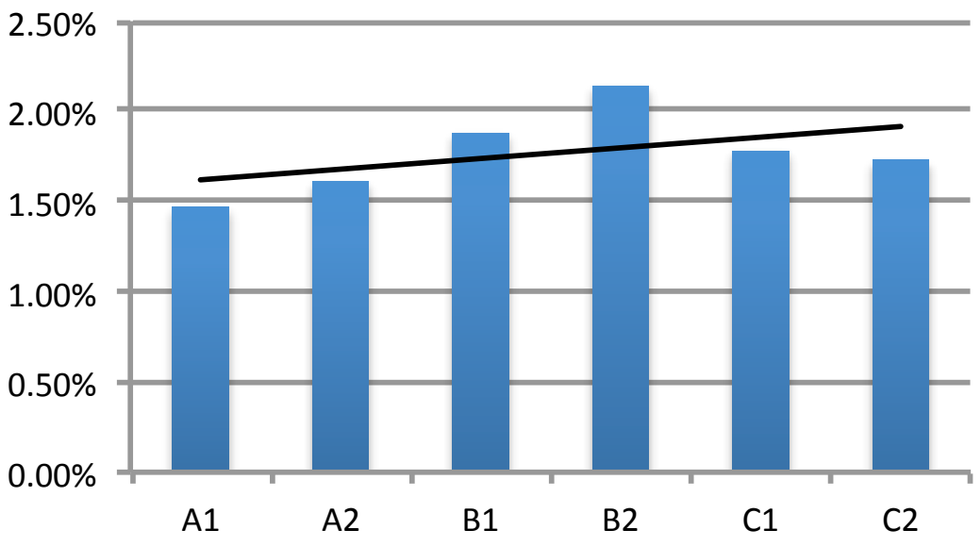


Use of past-progressive aspect

- 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.

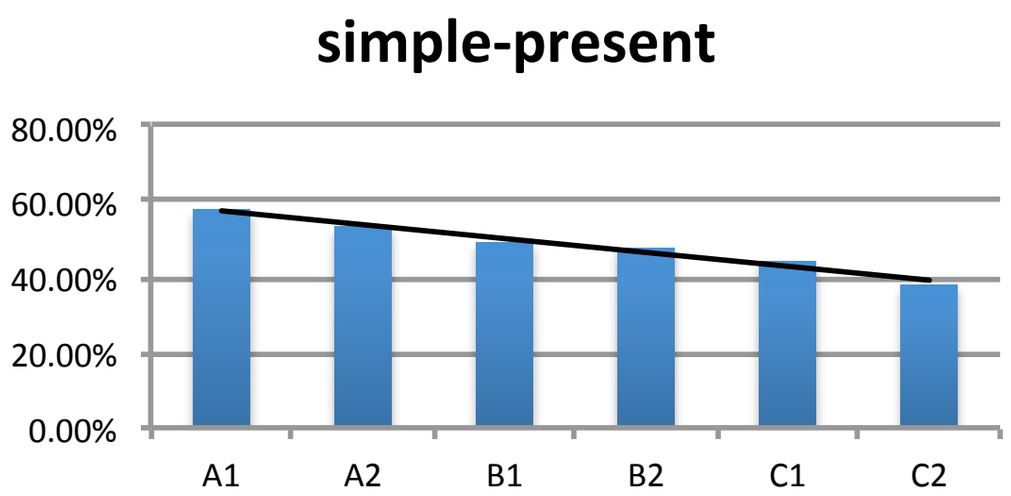
3.2 Using the corpora to sequence concepts (iii)

Ordering structures in difficulty

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



Transitivity

- **Recognition of semantic roles**
  - Actor, Process, Goal, Sensor, Phenomenon, etc.

John	*gave*	*Mary*	*a book*	*to read*
Actor	*Process*	*Recipient*	*Goal*	*Circumstance*

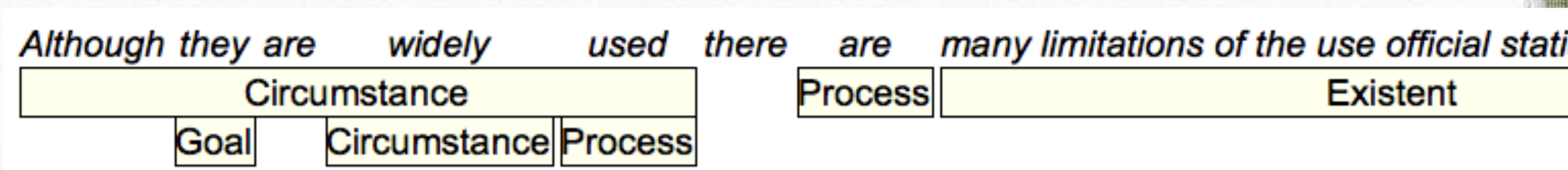
John	*hates*	*that Mary is a better player*
Senser	*Process*	*Phenomenon*

John	*told*	*Mary*	*to go*
Sayer	*Process*	*Addressee*	*Verbiage*

John	*is*	*a bad chess player*
Carrier	*Process*	*Attribute*

Transitivity

- **Each clause assigned a process type**
  - material, mental, verbal, relational, existential
- **Key patterns recognised:**
  - **verbal-passive** (*it has been said that…*)
  - **mental-passive** (*it is believed that…*)
  - **Say-type vs. tell-type,**
  - **please-type vs. like-type**



Automatic Transitivity Analysis

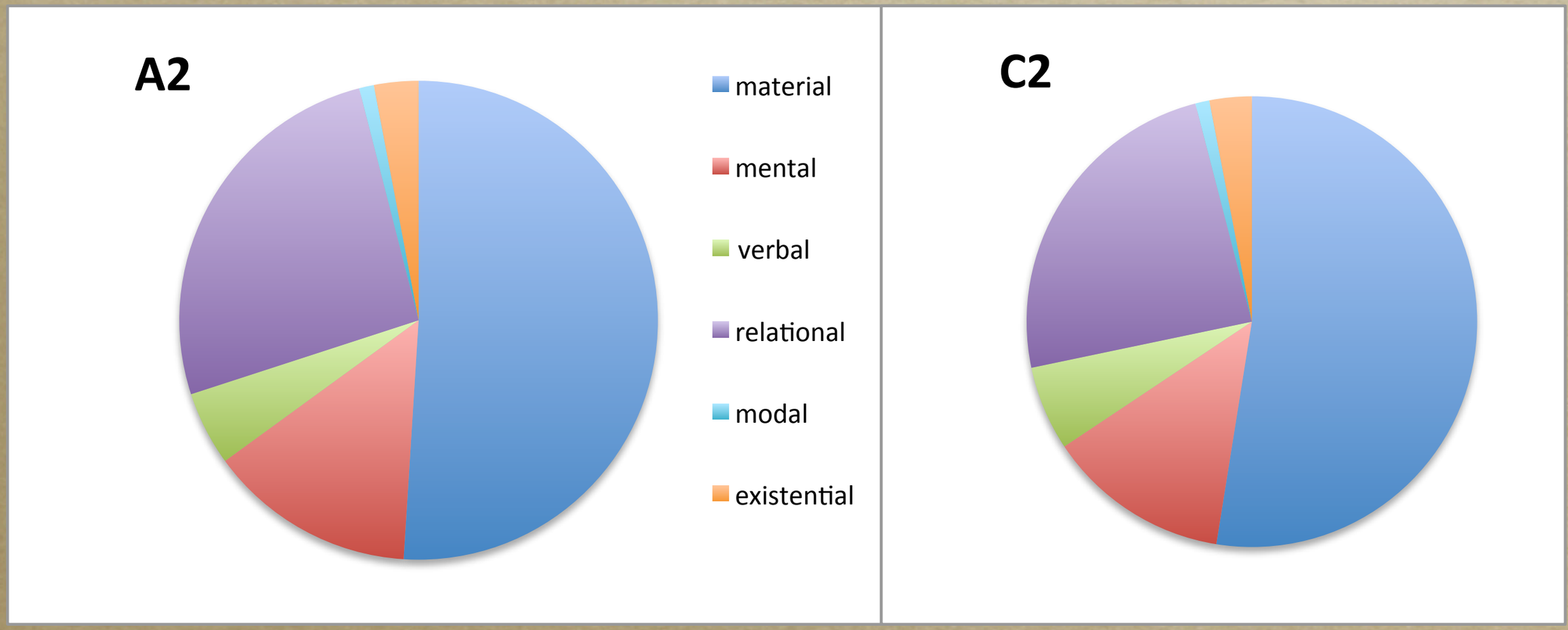
- Transitivity analysis derived from the Mood analysis of each clause unit:
  - Process type derived by:
    a. Looking up verb in process-type lexicon (9,300 verb senses)
    b. Where ambiguous, syntactic information used to disambiguate

Gingrich launched a blistering attack on Romney in what is essentially

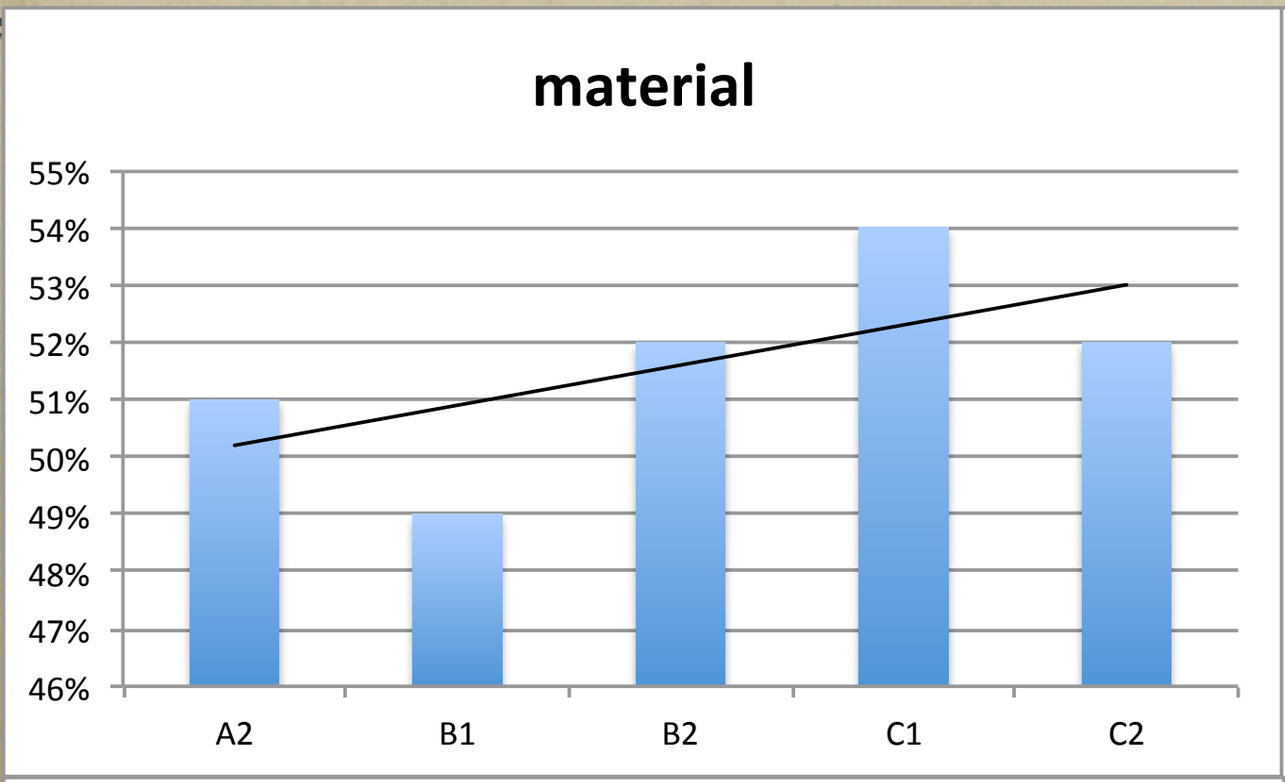
Actor	Process	Goal	Circumstance			
				Carrier	Process	Circumstance

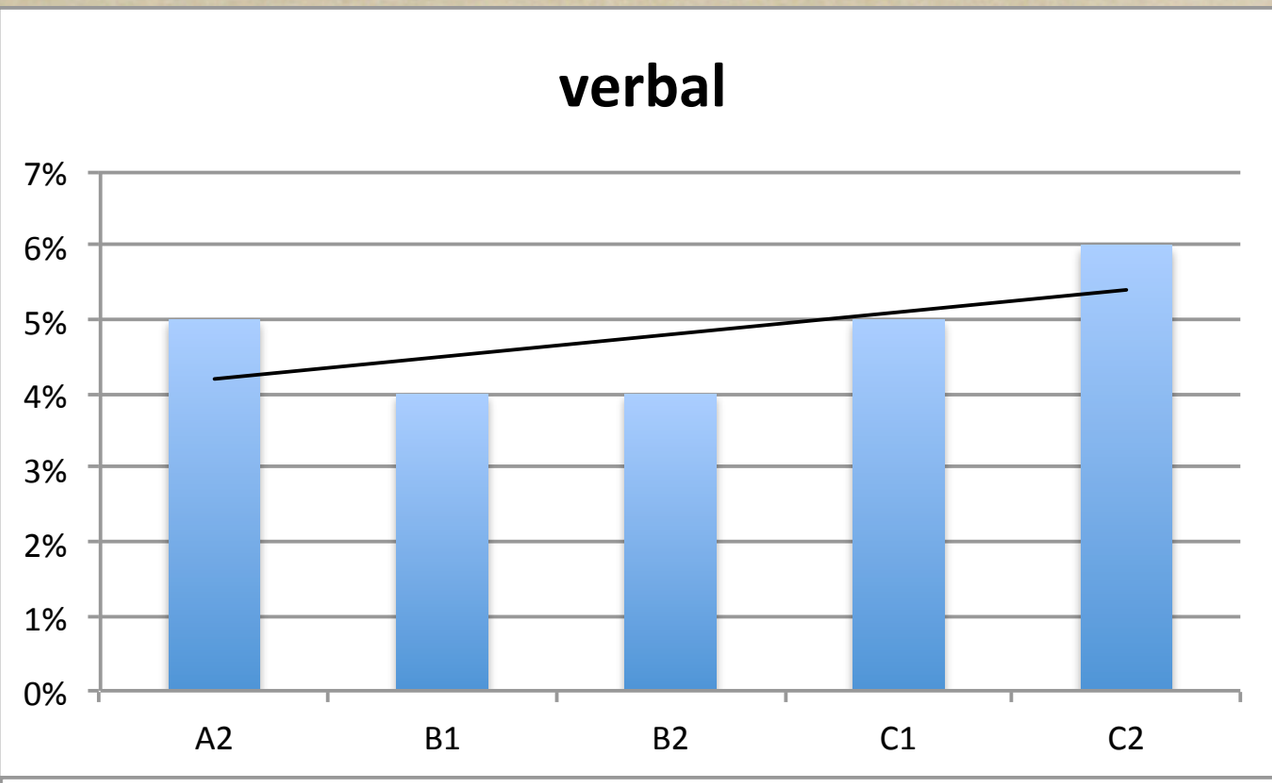
5. Results (i): General Process Type Usage

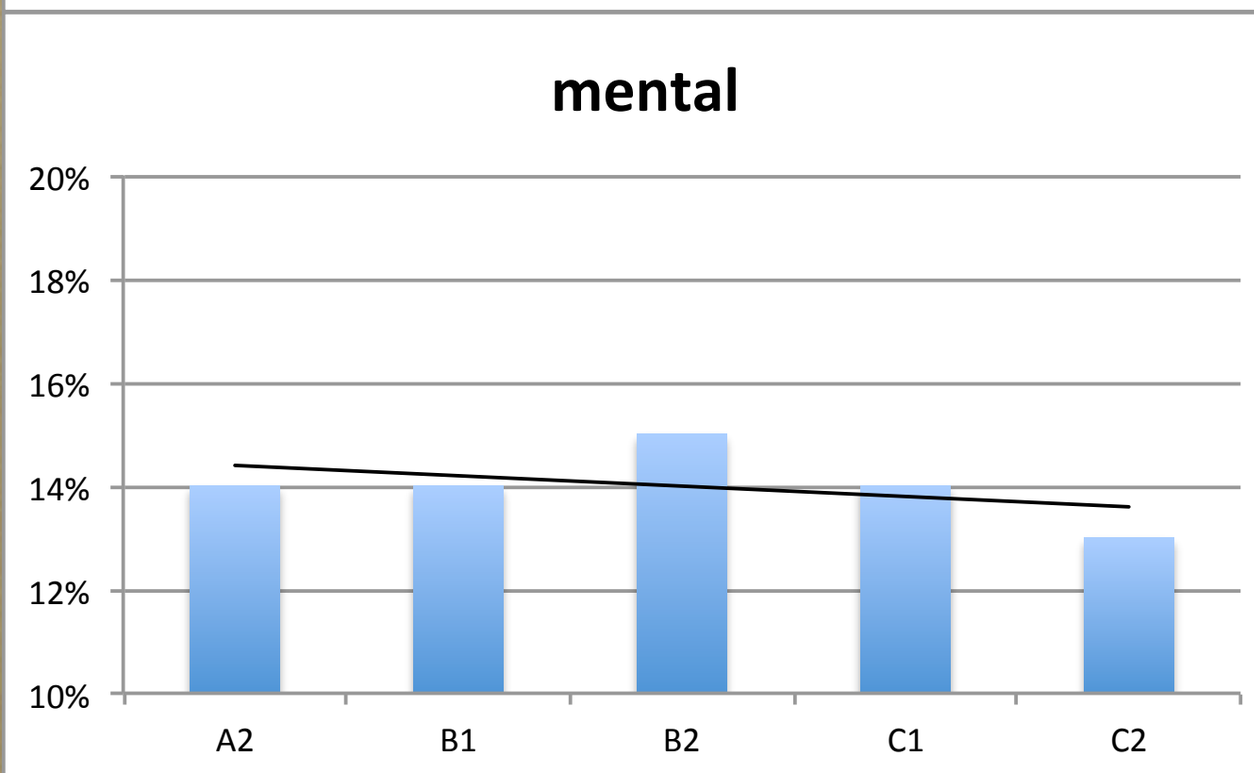
- Changing mix of process type usage with increasing proficiency: doesn't seem like much, but some shifts: fall in relational, increase in verbal

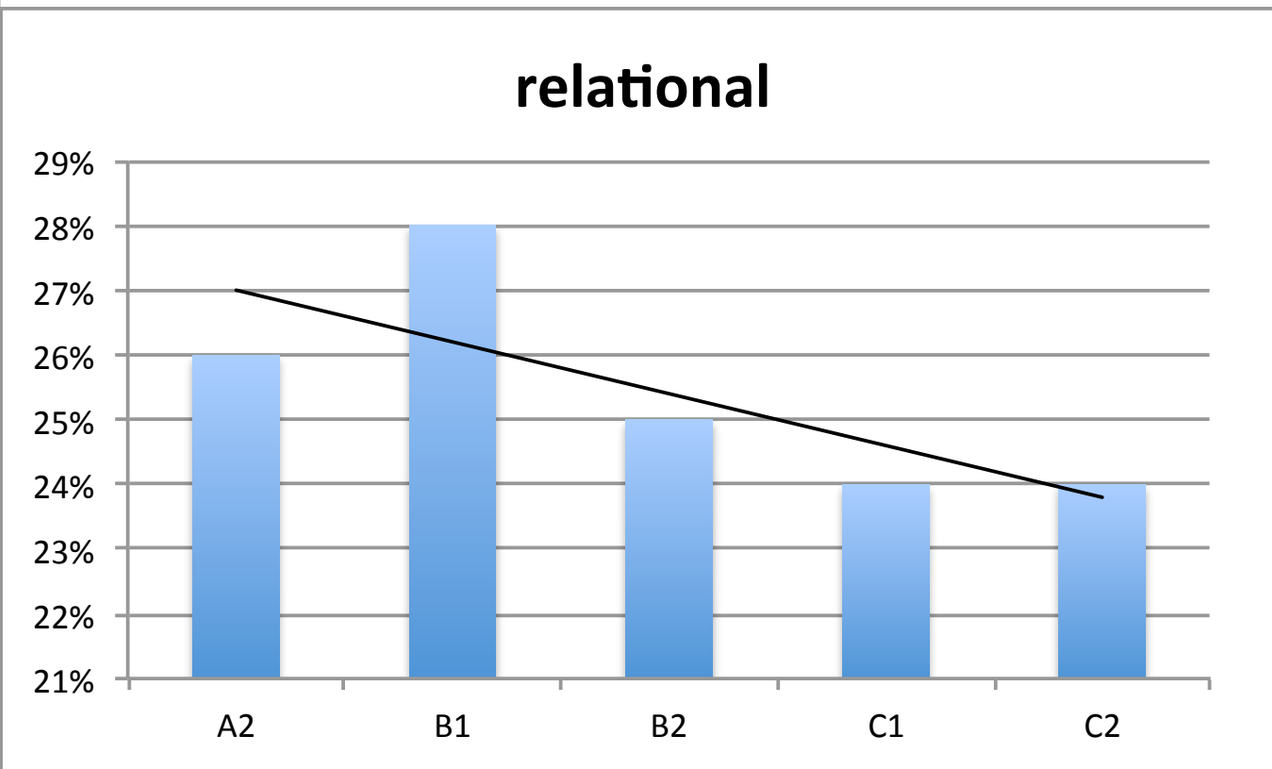


5. Results (i): General Process Type Usage



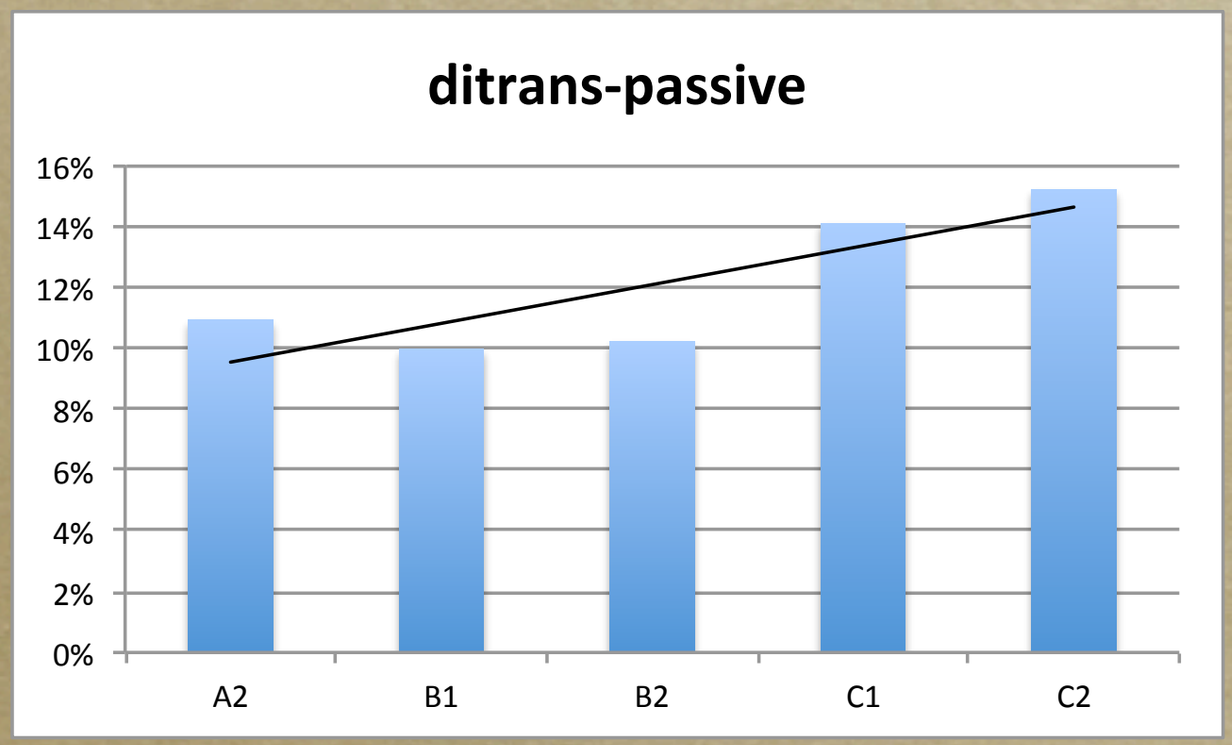






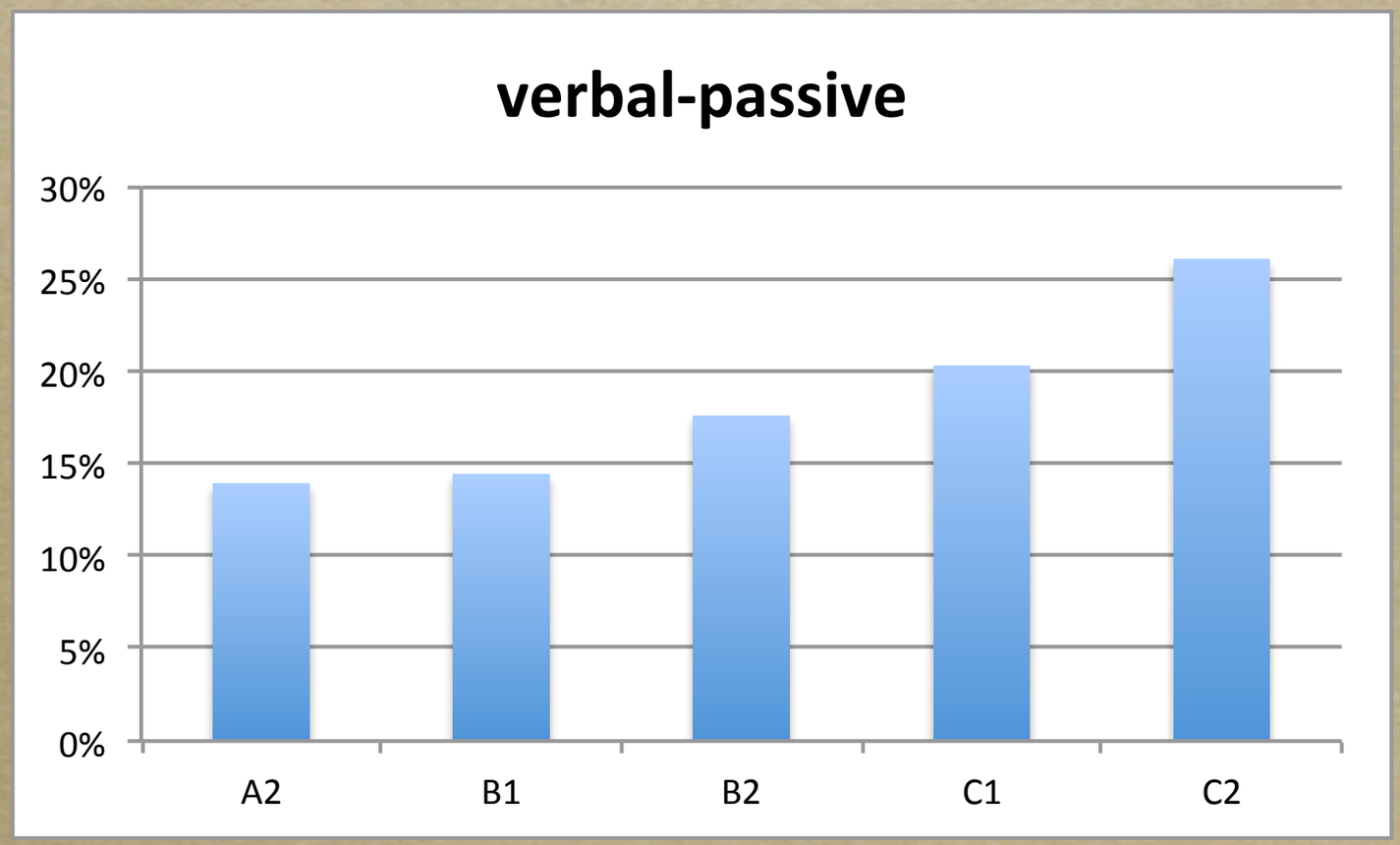
5. Results (ii): Material Processes

- Ditransitive verbs in Passive clauses:
- As with other process types, increased use of passive with ditransitive verbs
- Most of increase in Recipient^Process^Goal structures (*Mary was given an apple*)



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.



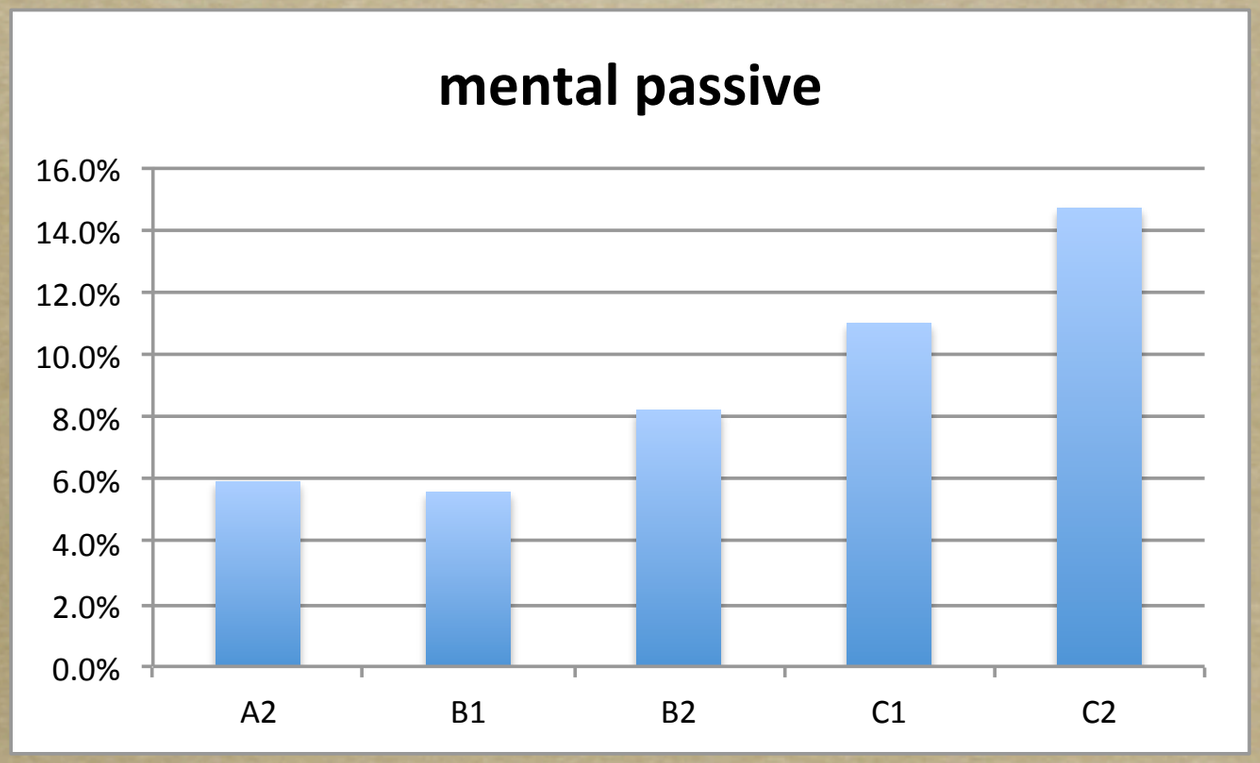
5. Results (iv): Mental Processes

Mental processes:

- As with other processes, clear increase in passive forms:
  - *It is considered/believed/expected/felt that …*

(postponed Subj=Phenom.)

- Again, students avoiding mention of the Senser!



Theme-Rheme

- I assume Halliday's model, as presented in "Introduction to Functional Grammar", 4th Edition (Halliday and Matthiessen).
- In declaratives, Theme includes all clausal elements up to and including the first experiential element (most typically the Subject).
- So, Textual and Interpersonal elements may precede:

Unfortunately	*however*	*the revolution*	*failed*
Theme			Rheme
Interpers.	Textual	Topical	

Theme-Rheme

- Recognition of Topical, Interpersonal and Textual Themes (Halliday)
  - **Textual**: link clause to previous clauses.
  - **Interpersonal**: Speaker comment or provision of probability etc. (*Luckily, apparently*, etc.)
  - **Topical**: The first ideational item in the clause

Linguistic Model
Topical choices

In declaratives:

- unmarked Topical theme is SUBJECT:
  -> ***John*** *likes coffee in the morning*.

- fronted Adjunct:
  -> ***In the morning*** *John likes coffee*.

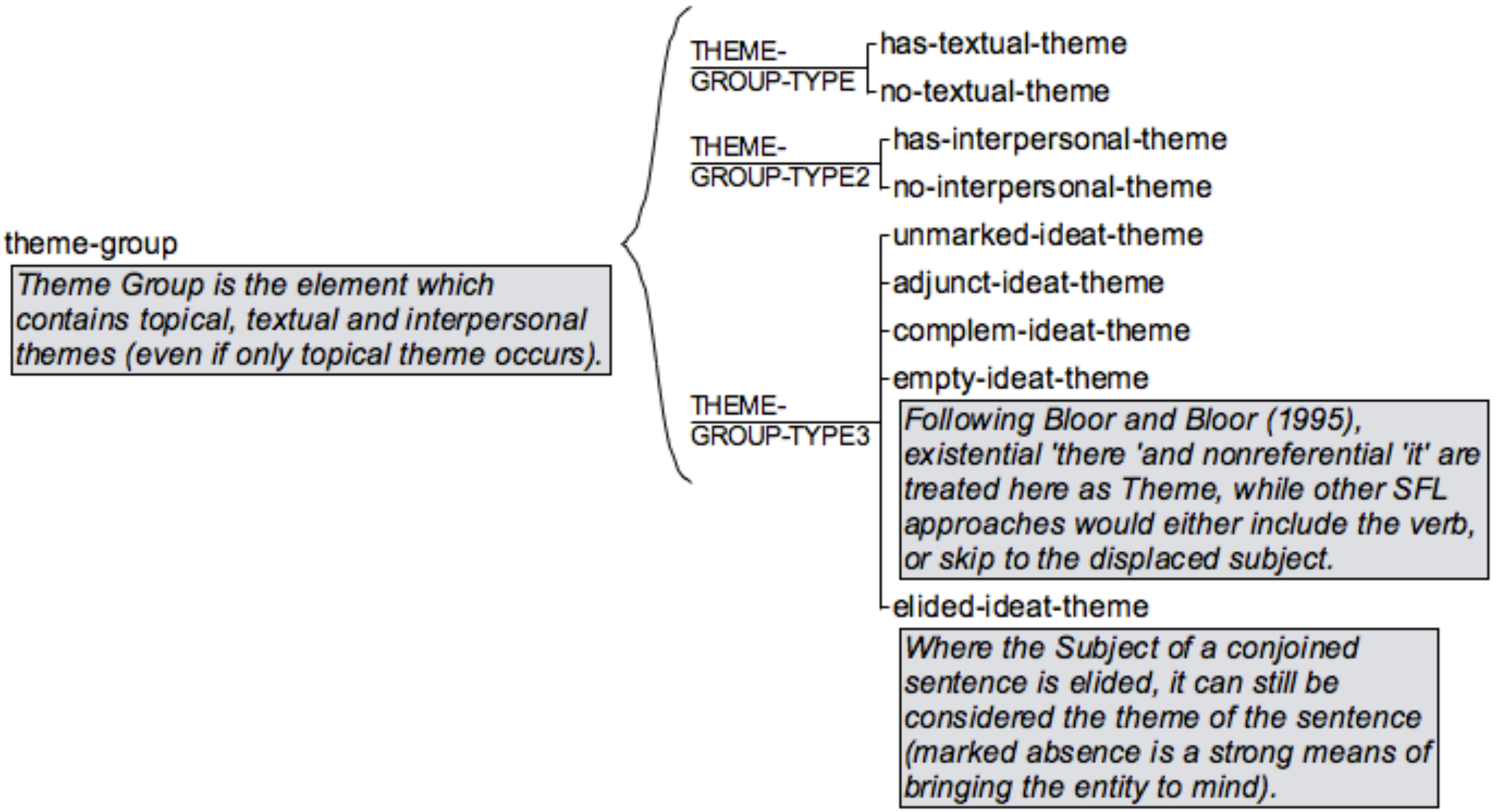
- fronted Complement:
  -> ***Coffee****, John likes*.

- fronted Dependent Clause:
  -> ***Because I drank too much coffee,*** *I cannot sleep*.

Theme-Rheme

- Featurised in terms of:
  - **presence** of textual, interpersonal themes
  - **marked** topical themes: *fronted-adjunct, elided-theme, dummy-theme*, etc.
  - **textual** semantic types: *structuring* (*firstly*), *arguing* (thus), *extending* (and)
  - **interpersonal** semantic types: *evidence (probably), evaluation (happily), admission (honestly),* etc.

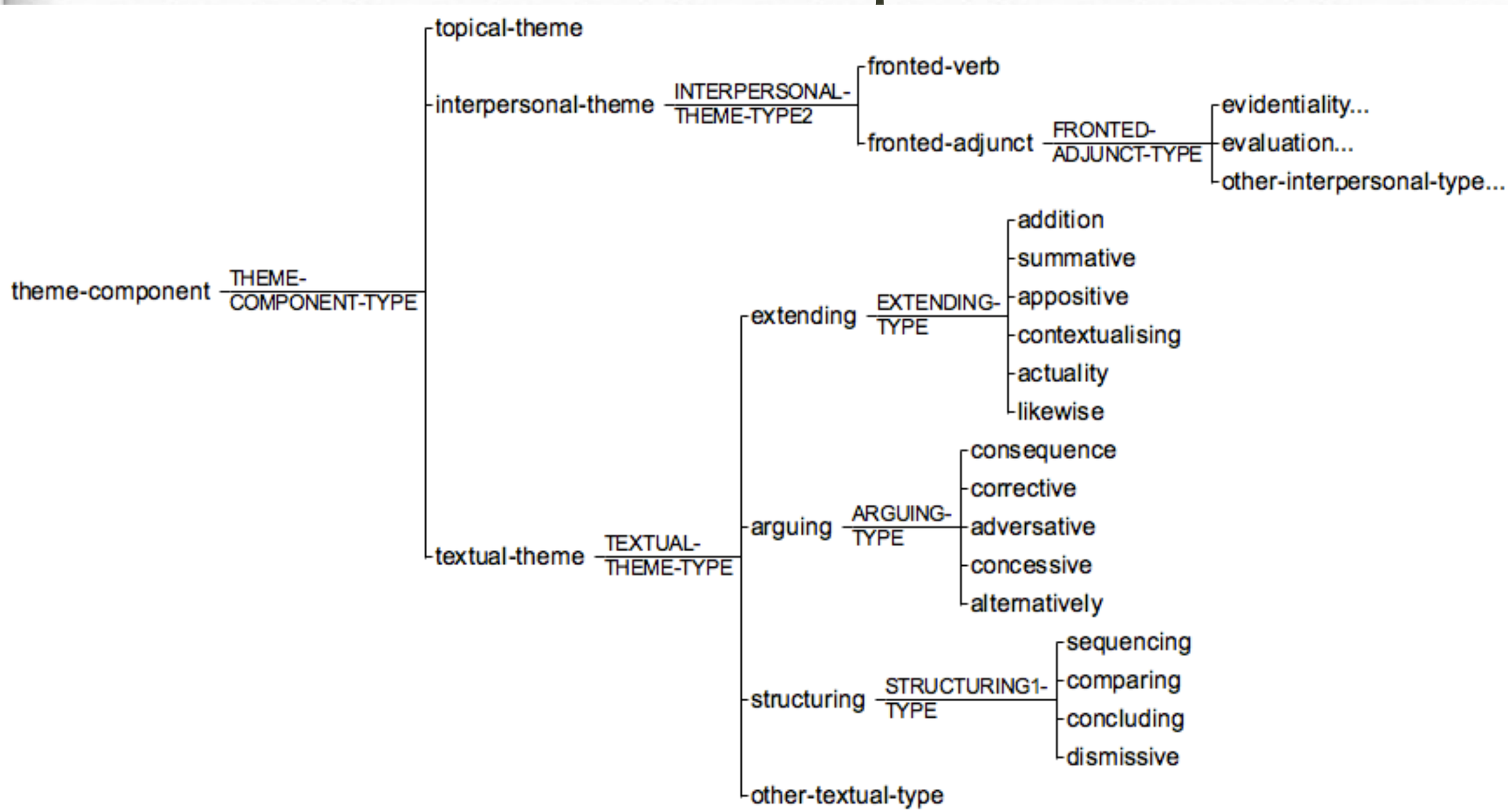
Theme Group



Theme group →

Unfortunately	*however*	*the revolution*	*failed*
Theme			Rheme
Interpers.	Textual	Topical	

Theme Component



	Unfortunately	*however*	*the revolution*	*failed*
Theme Component →	Theme			Rheme
	Interpers.	Textual	Topical	

Modality

- **Syntactic types**
  - modal auxiliary, (*should*)
  - semi-lexical (*have to, ought to*),
  - verb (*require*),
  - adverb (*possibly*)
  - adjective (*it is possible*)

- **Semantic types (of lexical modals)**
  - possibility, necessity, obligation, etc.

(based on work with Rebeca Garcia)