Exploring developing Quantification skill in Spanish Learners of English

Mick O'Donnell
Universidad Autónoma de Madrid

Map

1. Goals of Study
2. Linguistic Model (Nominal Quantification)
3. The Corpus
4. Results
5. Conclusions

Part 1: Goals of Study

- To understand how Spanish University learners of English develop their use of quantification resources in the noun phrase as they develop towards native competence
  - Uses a corpus of learner essays each associated with CEFR proficiency levels (A1, A2, B1, etc.)
  - Compared to a comparable native corpus (from BAWE)
- To discover indications as to where we need to focus our teaching of this area of grammar.

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)

Part 2: Linguistic Model

- I am studying quantification in the nominal group (noun phrase):
  - Thus exclude quantification in clause
    My friends have mostly gone home.

- *Two locations studied:*
  - Predeterminer*:* *all my friends; some of their friends*
  - Determiner*:* *all people; many cars; a few problems*

Part 2: Linguistic Model

- I am studying quantification in the nominal group (noun phrase):
  - Thus exclude quantification in clause
    My friends have mostly gone home.

- *Two locations studied:*
  - Predeterminer: *all my friends; some of their friends*
  - Determiner: *all people; many cars; a few problems*

Case like "the many shades of grey" treated as distinct phenomena, ignored.

Part 2: Linguistic Model

- I am studying quantification in the nominal group (noun phrase):
  - Thus exclude quantification in clause
    My friends have mostly gone home.

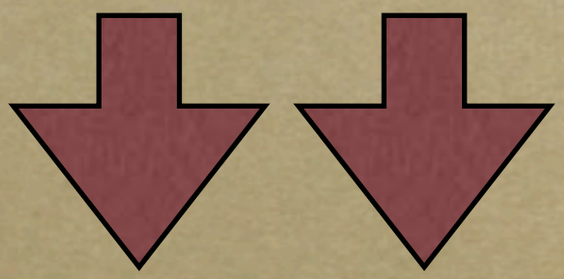
- *Two locations studied:*
  - Predeterminer: *all my friends; some of their friends*
  - Determiner: *all people; many cars; a few problems*

'a few' and 'a little' considered as one unit (note lack of agreement between 'a' and 'problems')

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 2: Linguistic Model

Tokens of interest:

- Dual: neither, either, both
- Mass: little, a little, much,
- Count: few, a few, many, several, every, each
- Mixed: no, any, some, all, more, most, half, a lot, lots

Structures also recognised:

- a large part of, so many of, far fewer, etc.

Part 2: Linguistic Model: errors and context

- Interested in production errors:
  - *I have much water* -> `much/any-in-positive-statement`
  - *I don't have some water.* -> `such-in-negative-sentence`
  - *I don't have much dogs.*-> `mass-quantifier-with-count-noun`
- Thus, context of the noun phrase is also relevant:
  - positive or negative sentence
  - Interrogative, declarative, imperative
  - Presence of intensification/comparison:
    I have so much water.
    I have as much water as you have.

Part 3: The Data (i) Composition

- A Learner Corpus of short essays produced by Spanish university students:
  - 560,000 words, dedicated English Studies degree (Wricle Corpus, UAM)
  - 150,000 words, English in other degrees (UPV Learner Corpus)

  (All essays associated with CEFR level by Oxford Placement Test)
- A Native Corpus for comparison purposes:
  - 190,000 words, essays produced by British university Sociology students (BAWE Corpus)

900,000 words

BAWE TEXTS

Racism is still a problem within our society today, and many ethnic minorities face inequalities in many areas, including education, housing, and employment. Ethnic minorities are concentrated into certain areas of the job market, such as manufacture and communication (Brown, 1992), are most likely to be the victims of assault (Abercrombie et al, 1994), and recent surveys have shown that racist ideas still exist in society. This can be seen in a survey, carried out in 1993, that asked a white sample whether they agreed or disagreed with the statement: 'Immigration has enriched the quality of life in Britain' and nearly half of the sample disagreed (Abercrombie et al, 1994, p255). In this essay I will look at what racism is, and how it is defined in contemporary society, and I will then explore why it still persists. In this section I will cover three areas that I think have contributed

LEARNER TEXTS

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.

Part 3: The Data (ii) Distribution

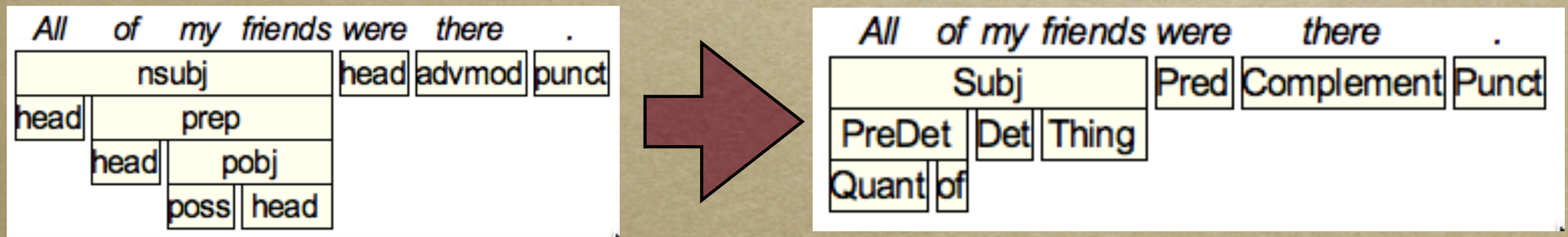
	A1	A2	B1	B2	C1	C2	Native
Words	16,000	90,000	250,000	200,000	116,000	20,000	190,000
Common NPs	3,144	19,500	51,663	40,200	24,067	3,921	42,328
Quantified Determiners	137	893	2764	2091	1192	189	1043
Quantified Predet.	69	453	943	594	230	43	146

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.
- Lemmas of each word provided by TreeTagger (Schmid, 1995) and merged into the syntax trees.
- Stanford parse is transformed into a richer corpus annotation:
  - Transformation towards more appropriate tree structure.
  - Featurisation of linguistic aspects of interest.

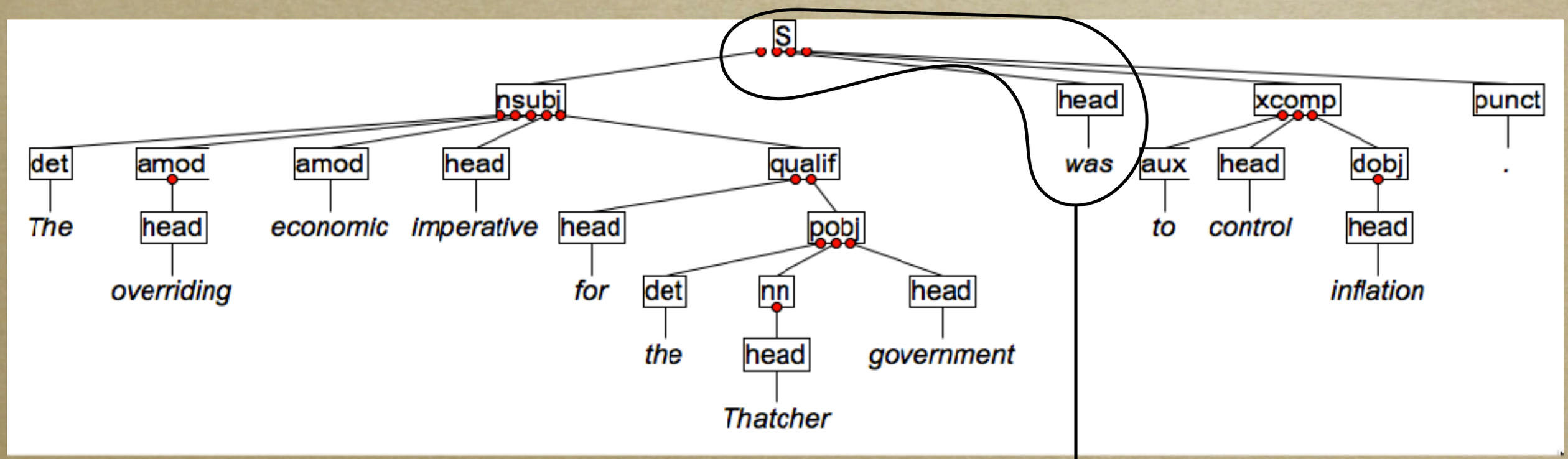
3. The Data: Tree Transformation

- 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.



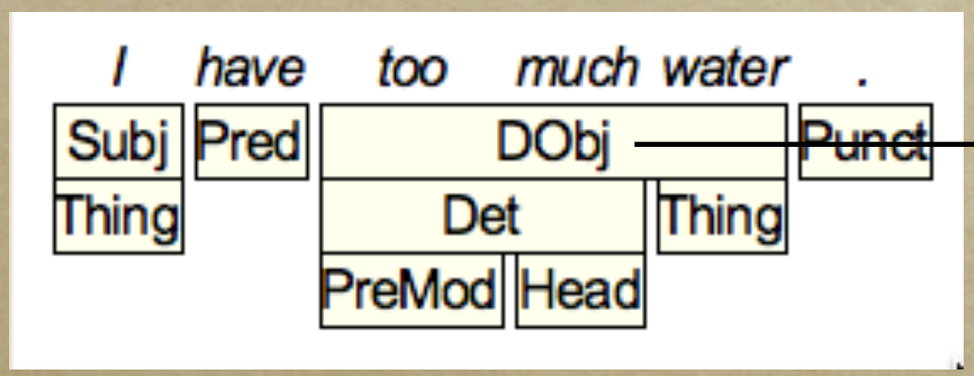
3. The Data: Featurisation

- 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:



3. The Data: Featurisation

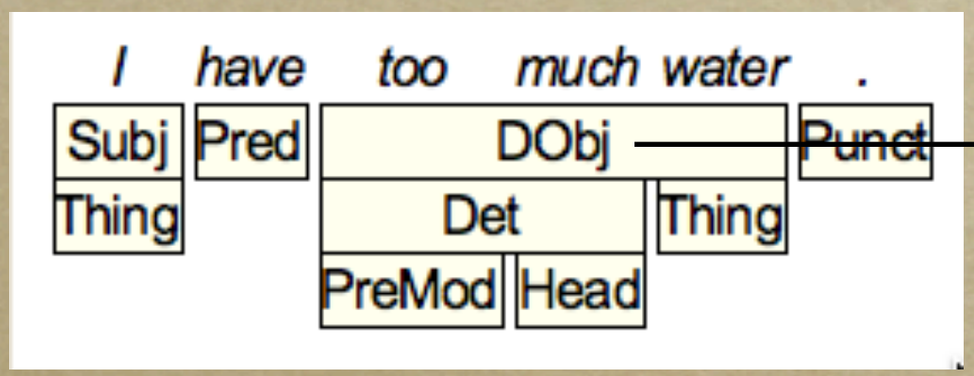
- UAM CorpusTool's internal code supplies features to each nominal group (noun phrase):



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):



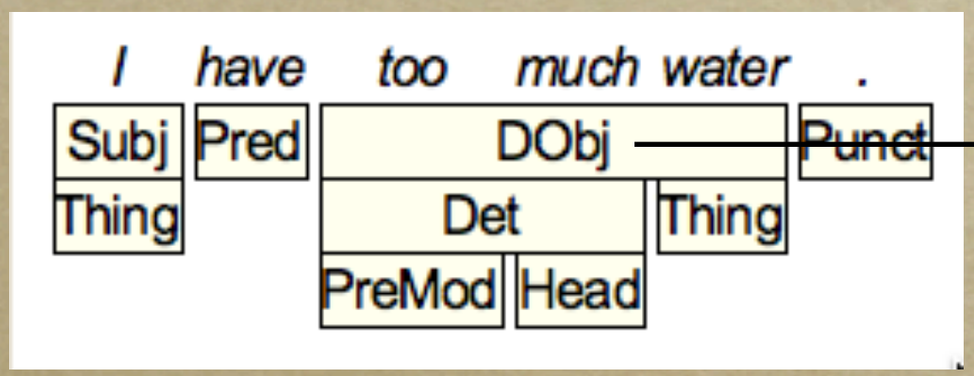
group
nominal-group
common-group
not-predetermined-group
determined-group
quantifier-determined
much-determined
not-premodified-group
not-postmodified-group

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:

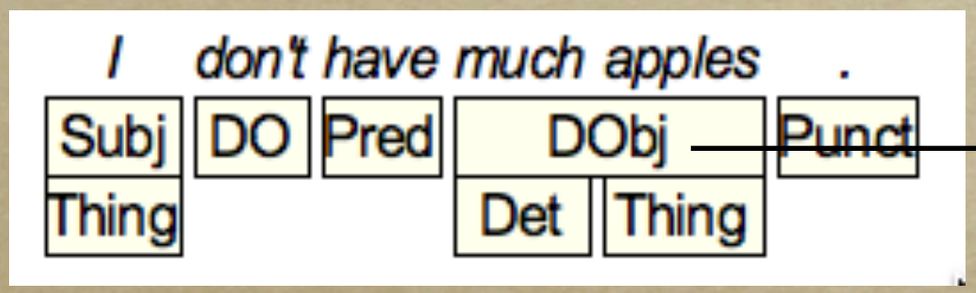
intensified-quantification
noncount-group
singular-group
concrete-group

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:

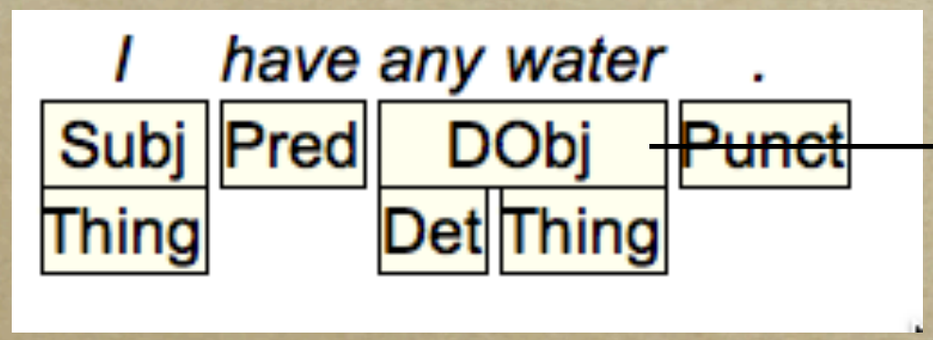
not-intensified-quantif
count-group
plural-group
concrete-group

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: Classifying count/noncount

- Count vs. noncount important to determine correct usage:
  - *few water // much people // many trouble*
- Not provided by Stanford Tagger or Parser

- Constructed index of countability of nouns using BAWE corpus:
  - Count times the noun appears in noncount contexts:
    - "much X" (strong evidence),
    - singular-noun with no determiner "Love", "water"
  - Count times the noun appears in count contexts:
    - "a X" etc.
    - plural form ("people")
  - Ratio of these evidences

3. The Data: Classifying count/noncount

- => Automatically produced index of countability, ordering nouns in degree of ‘countiness'
- List processed by hand to produce lists of noncount nouns, and mixed categories nouns (e.g., Activity is good/Here are two activities)
  - 600 noncount nouns
  - 150 mixed category
  - (THE LISTS NEEDS TO BE EXTENDED, BUT ARE FINE FOR THIS STUDY)
- After tagging the corpus of tagged NPs, examined most frequent count or noncount nouns to find tagging errors.
- Added these nouns to the appropriate list.

3. The Data: Identifying negative contexts

- Quantifiers like “much” and “any” are possible in negative contexts but not always in simple positive statements:

  ✗ *I have much money*

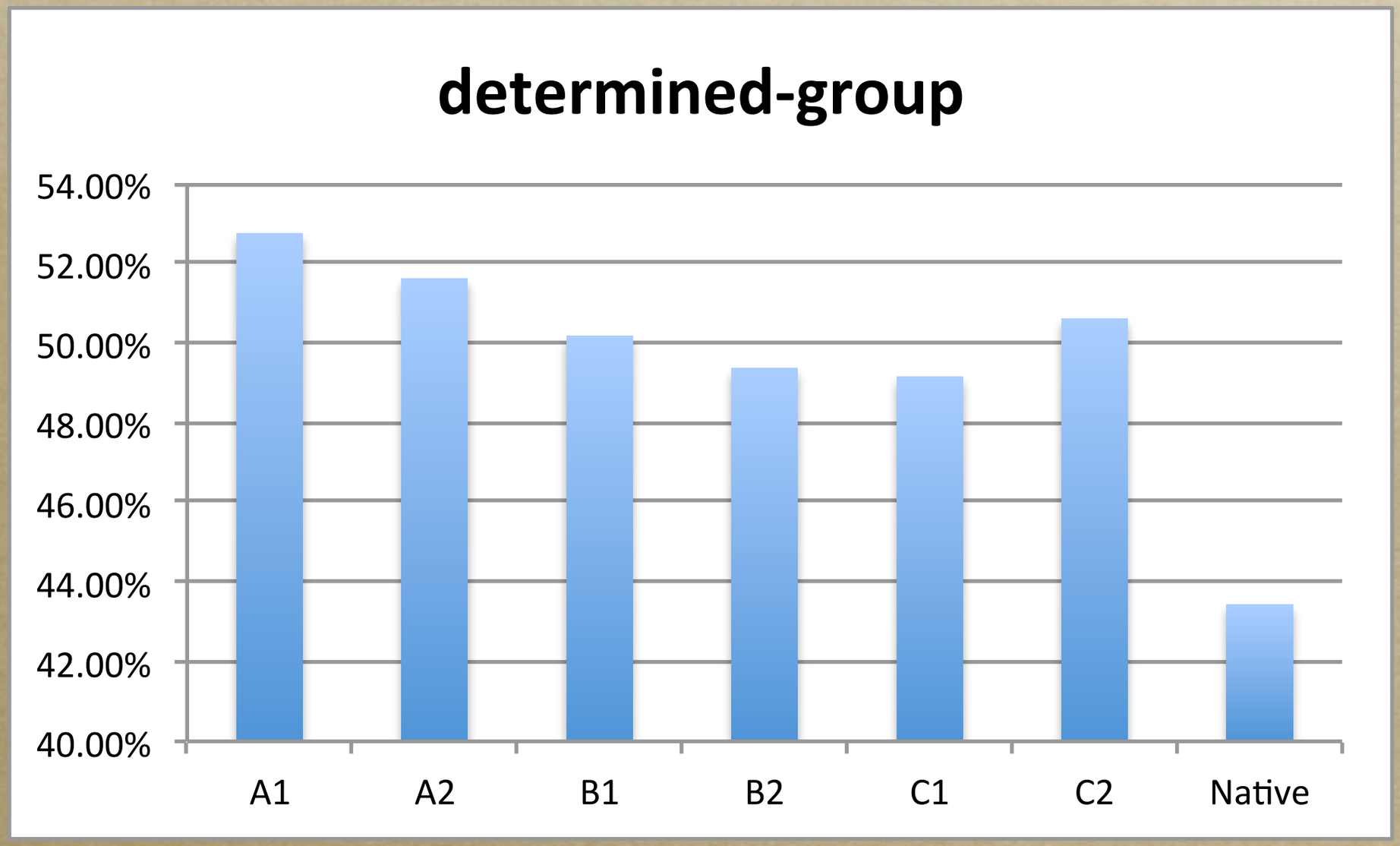
  ✓ *I don’t have much money.*

- 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.
  - Negative Adjunct: *I rarely have any money.*

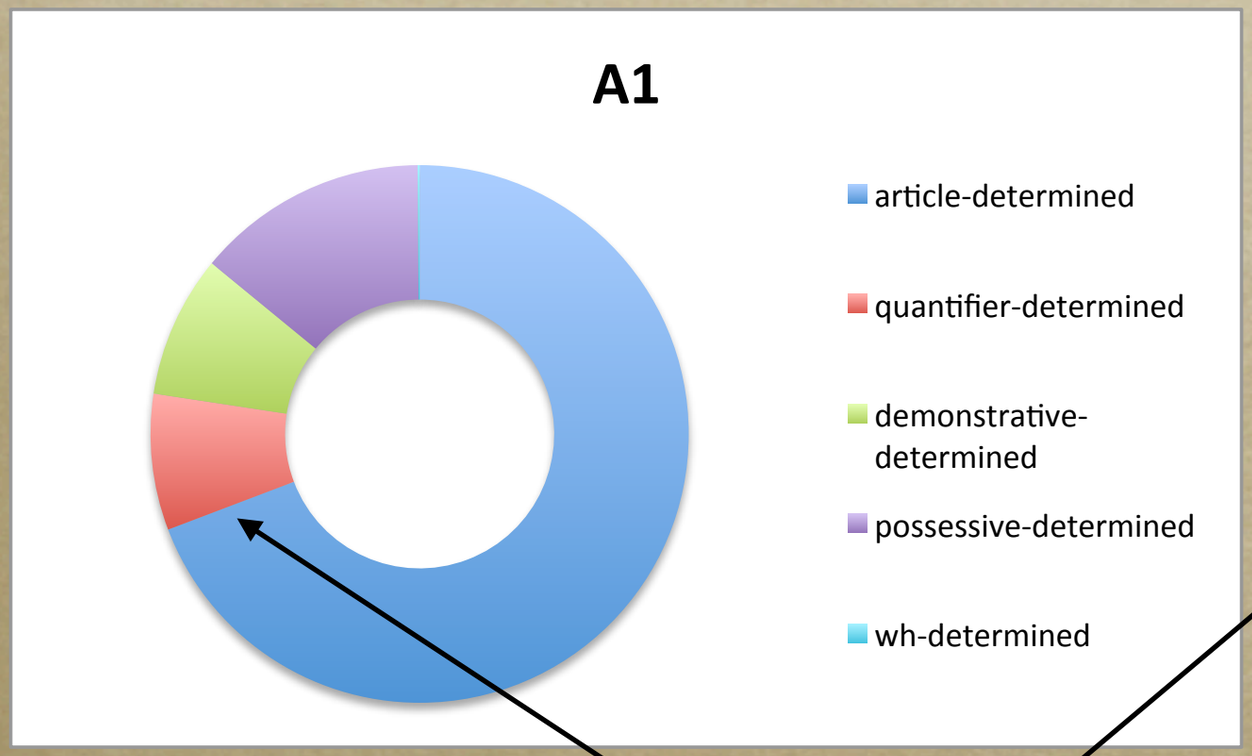
PART 4: Results
(i) Use of Quantifiers in Determiner slot

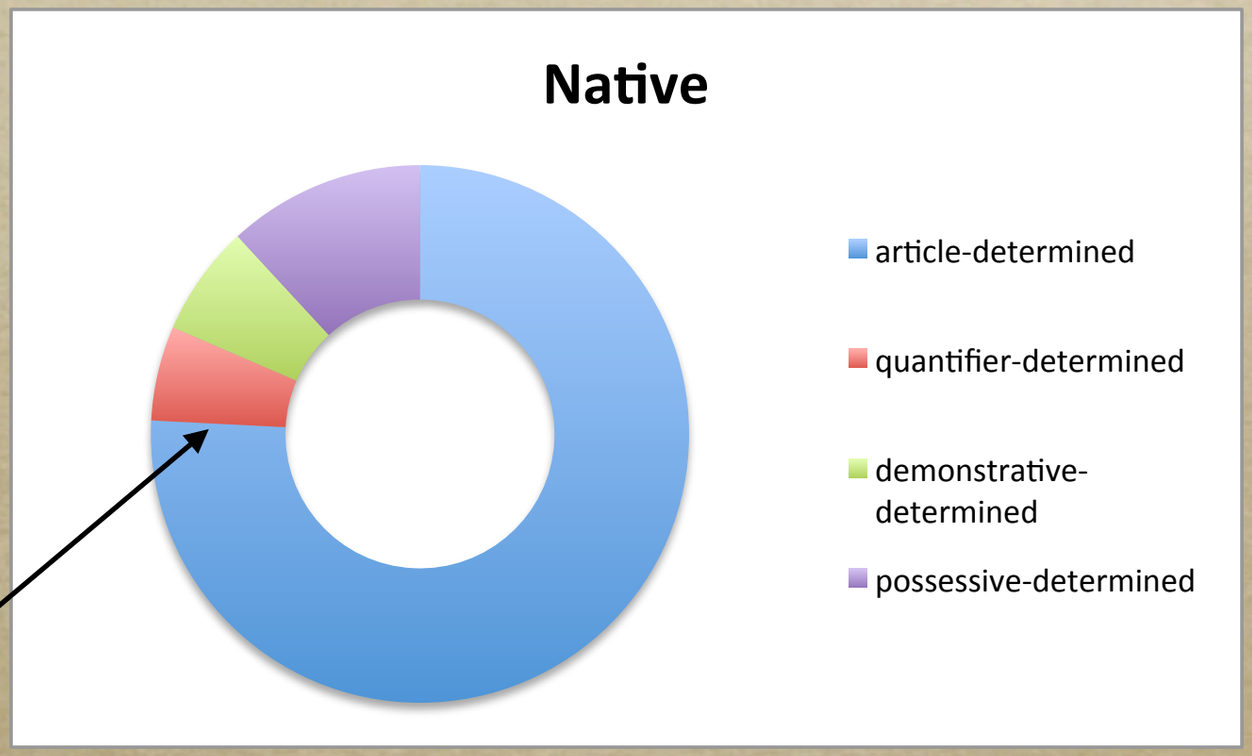
- Spanish learners over-produce determined common phrases
- (Graph: % of common noun phrases with determiner slot)



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

- Spanish learners over-produce quantified common phrases

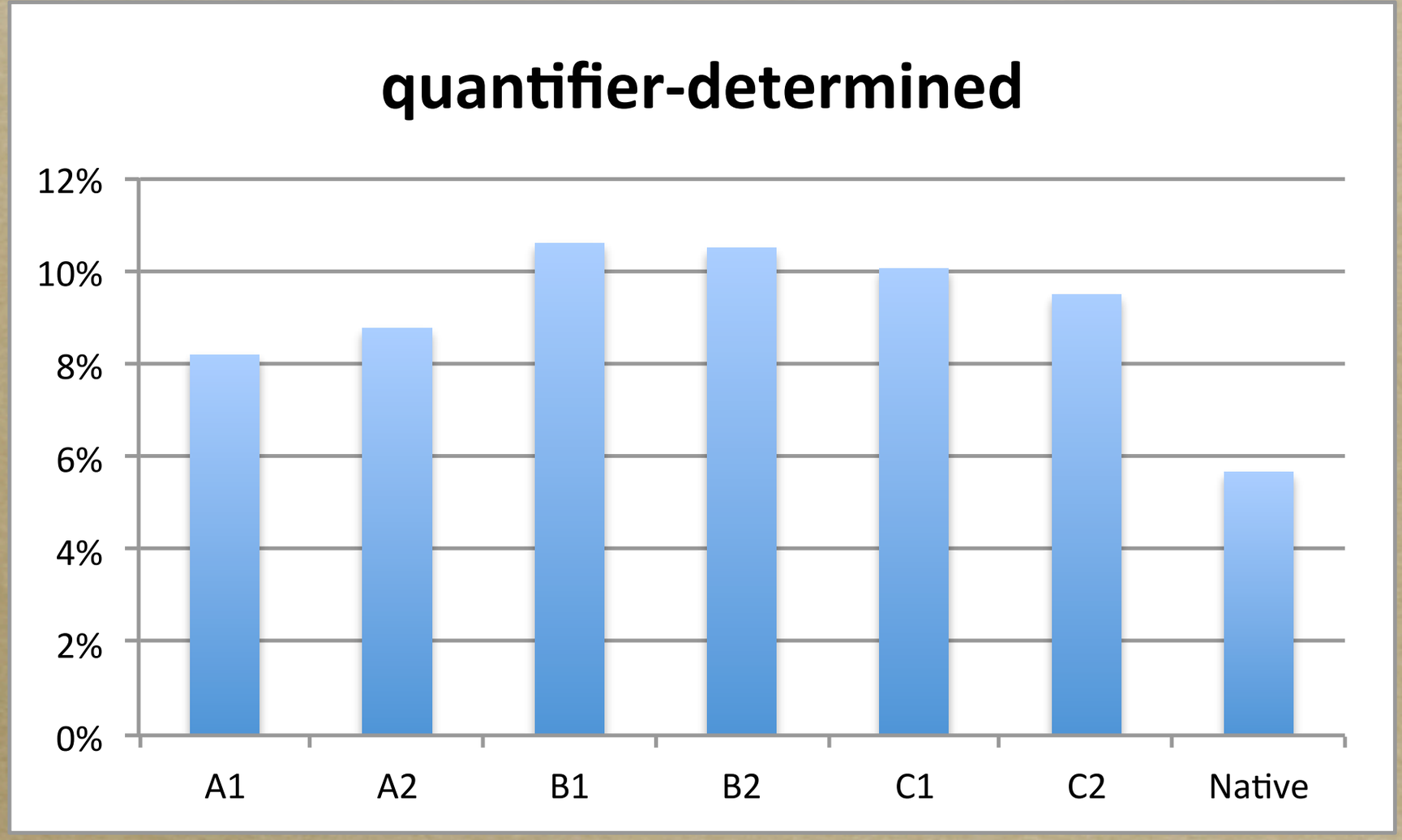




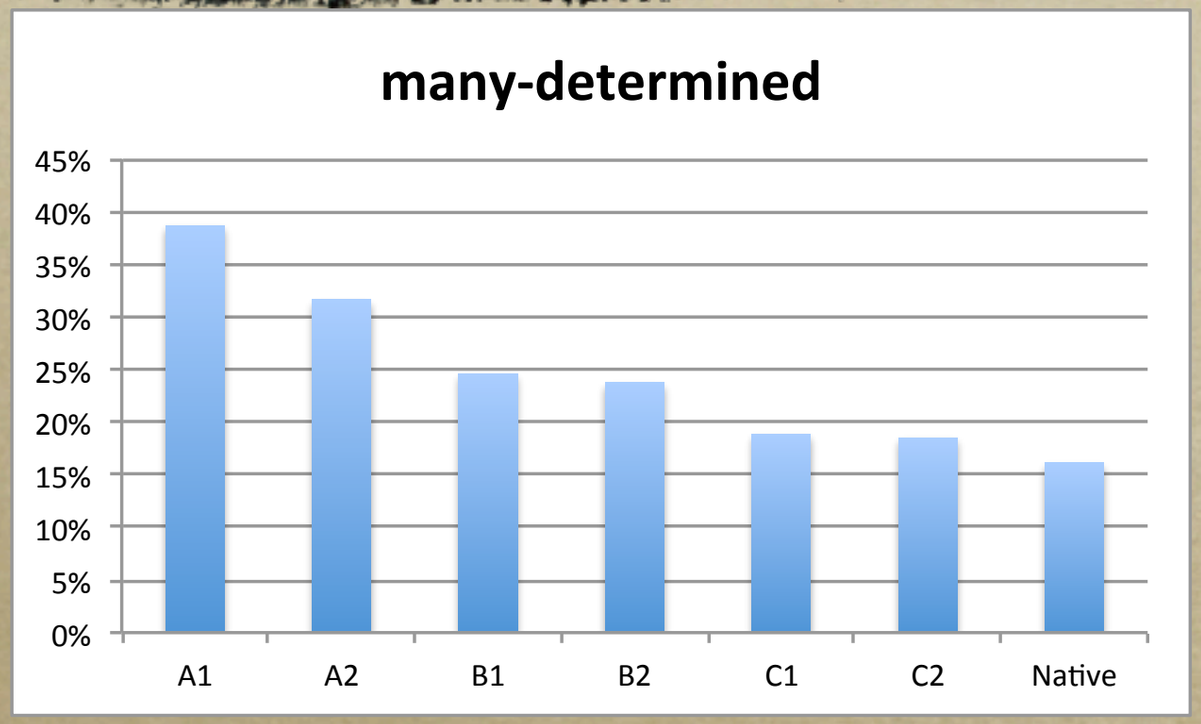
% of determiners which are quantifiers

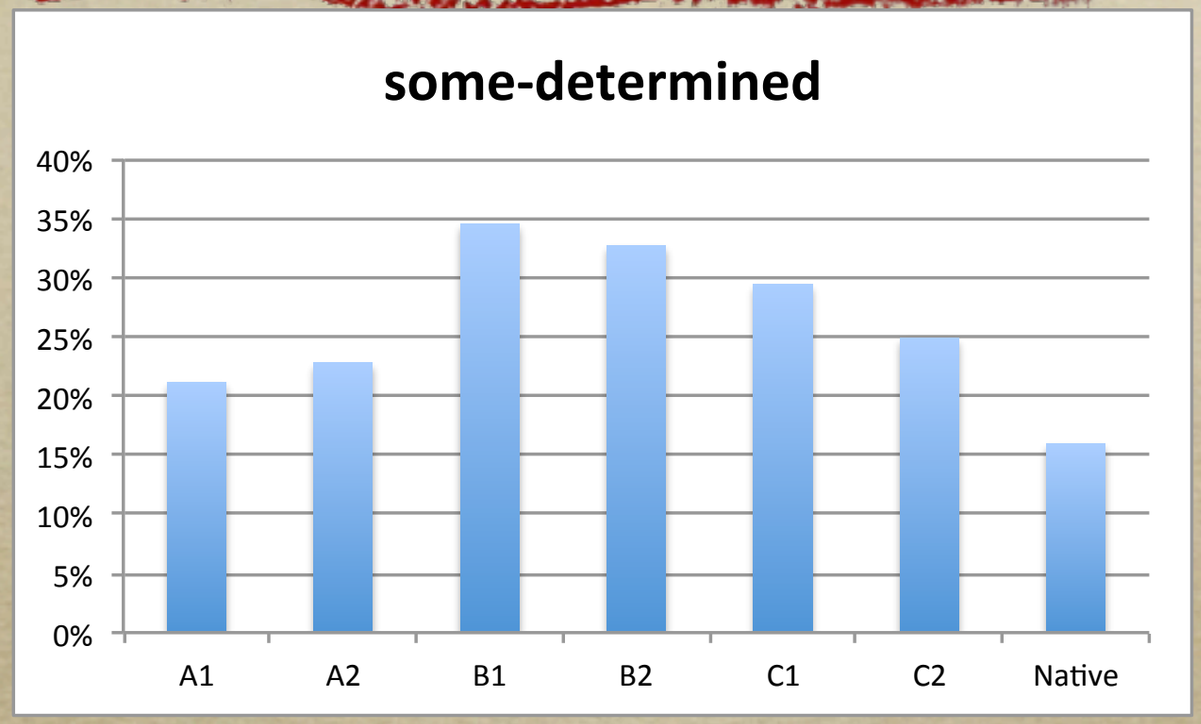
PART 4: Results
(i) Use of Quantifiers in Determiner slot

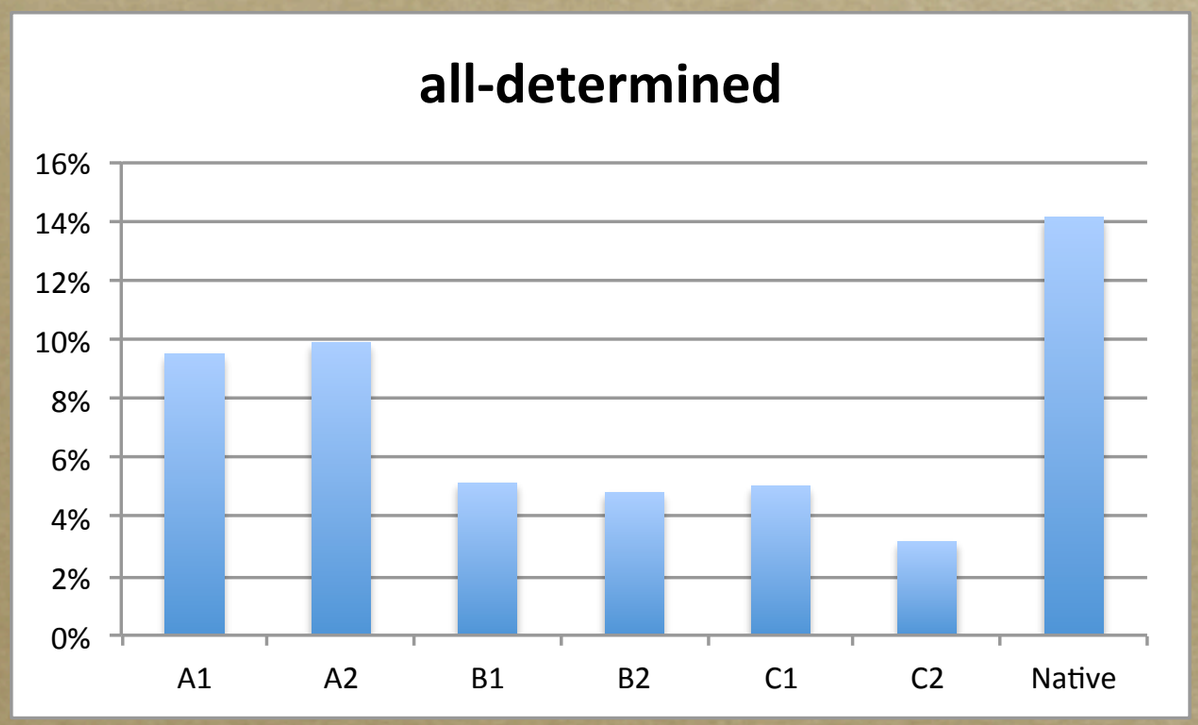
- Of the determined NPs, our learners use more quantifier determination than natives.
- E.g., “both reasons”, “no profit”, “many people”.

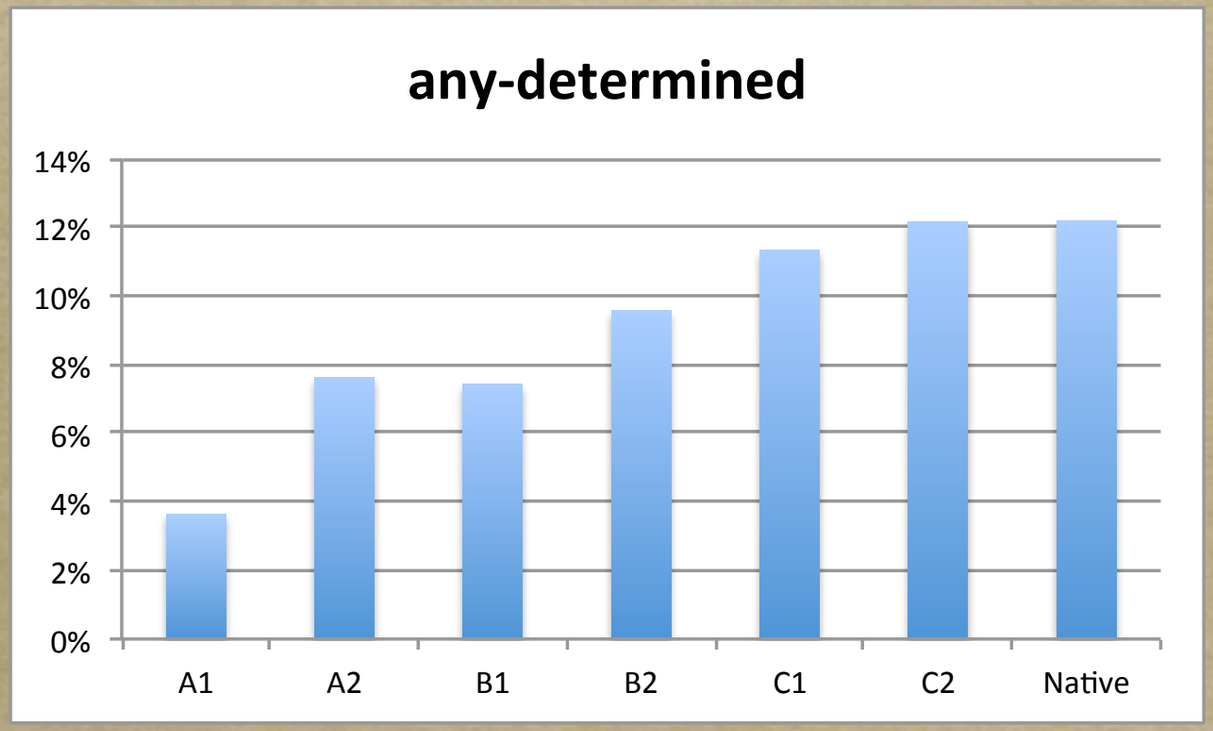


PART 4: Results
(i) Types of quantifier in Det slot

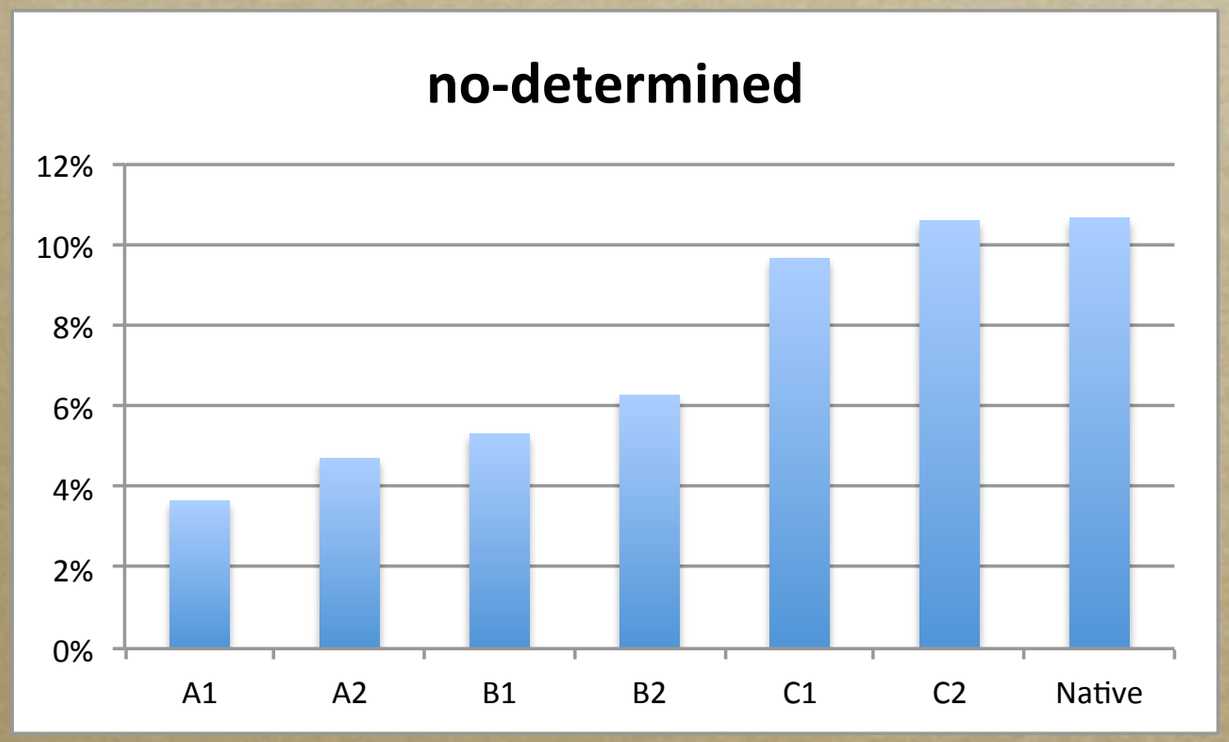


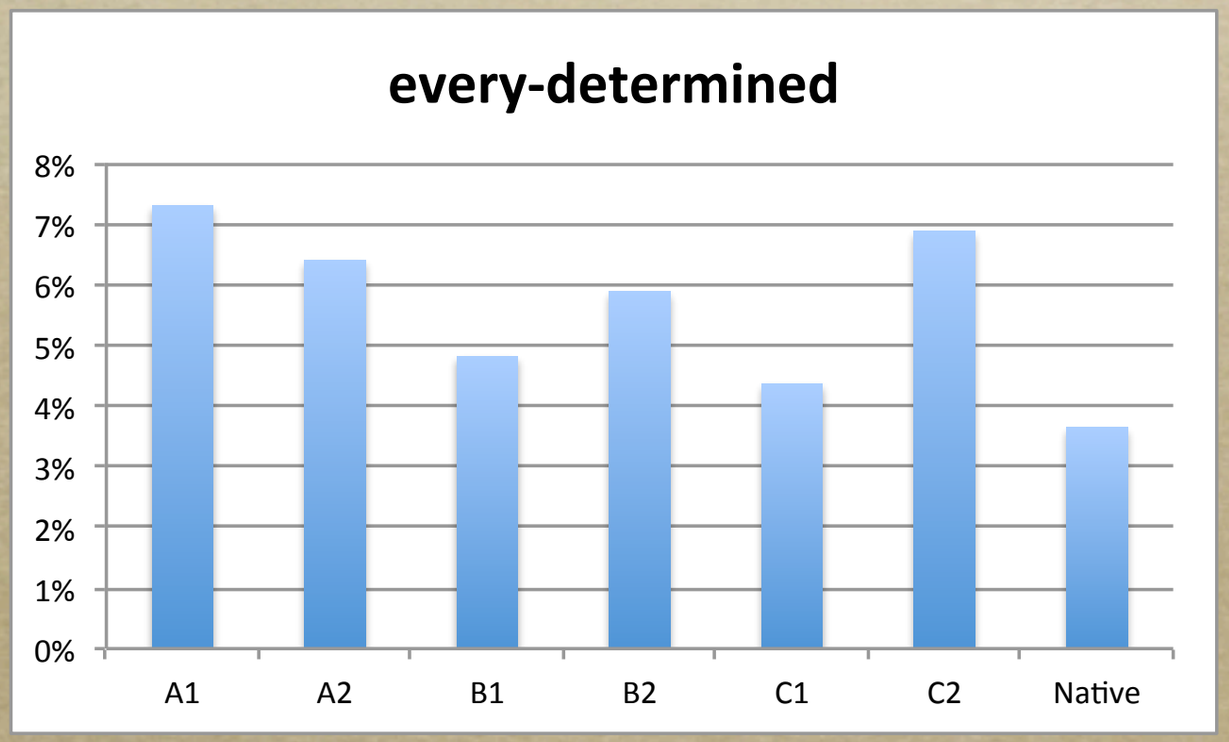






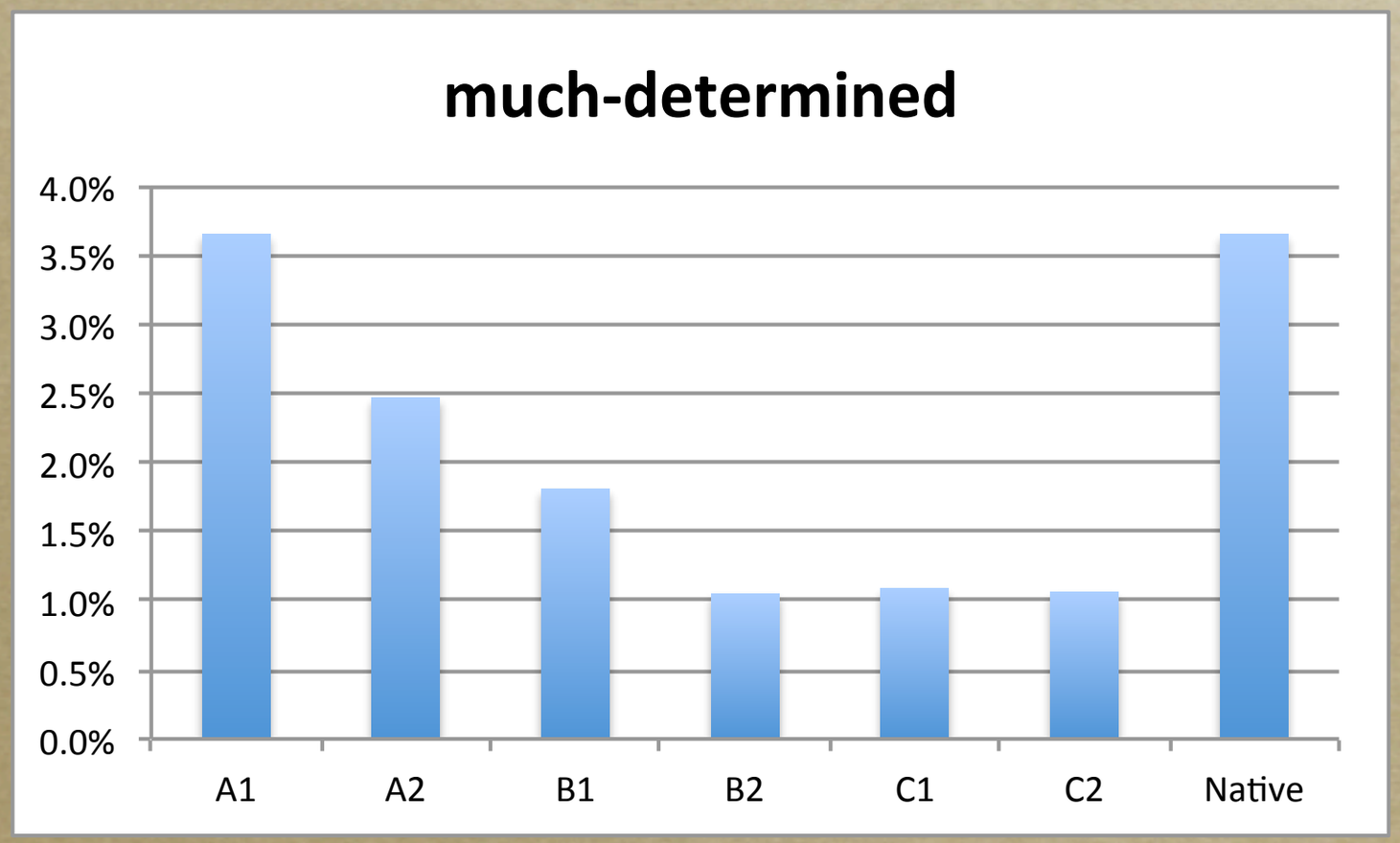
PART 4: Results
(i) Types of quantifier in Det slot





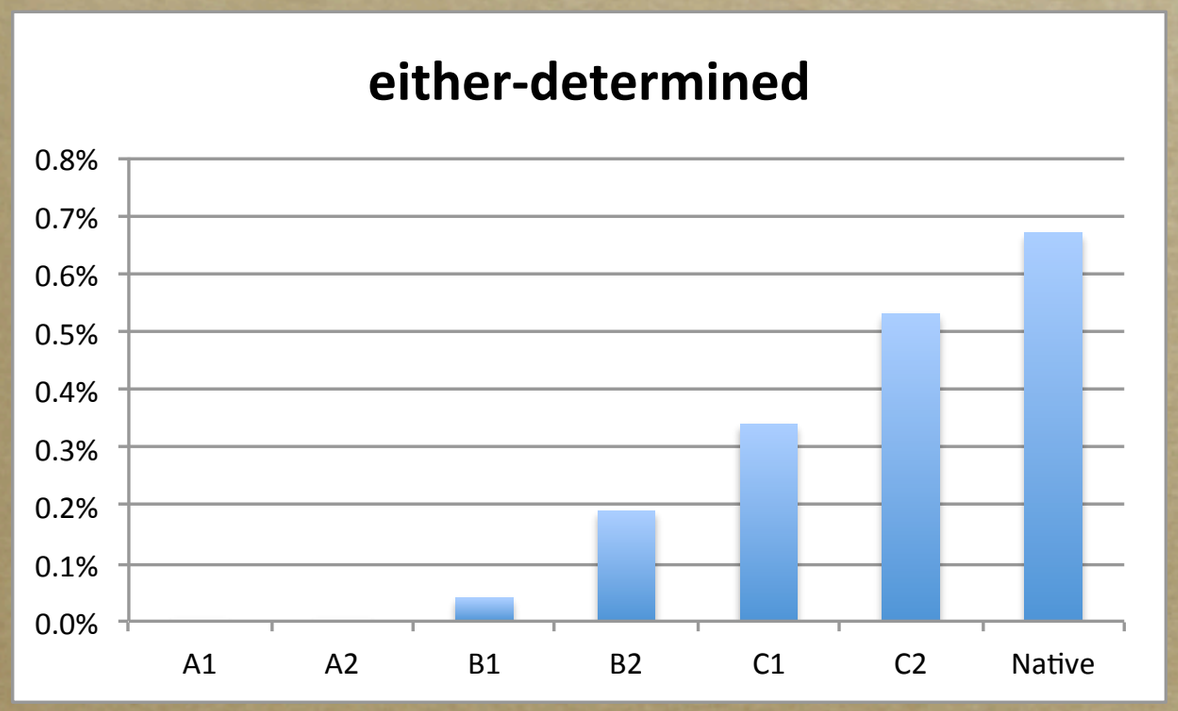
PART 4: Results
(i) Types of quantifier in Det slot

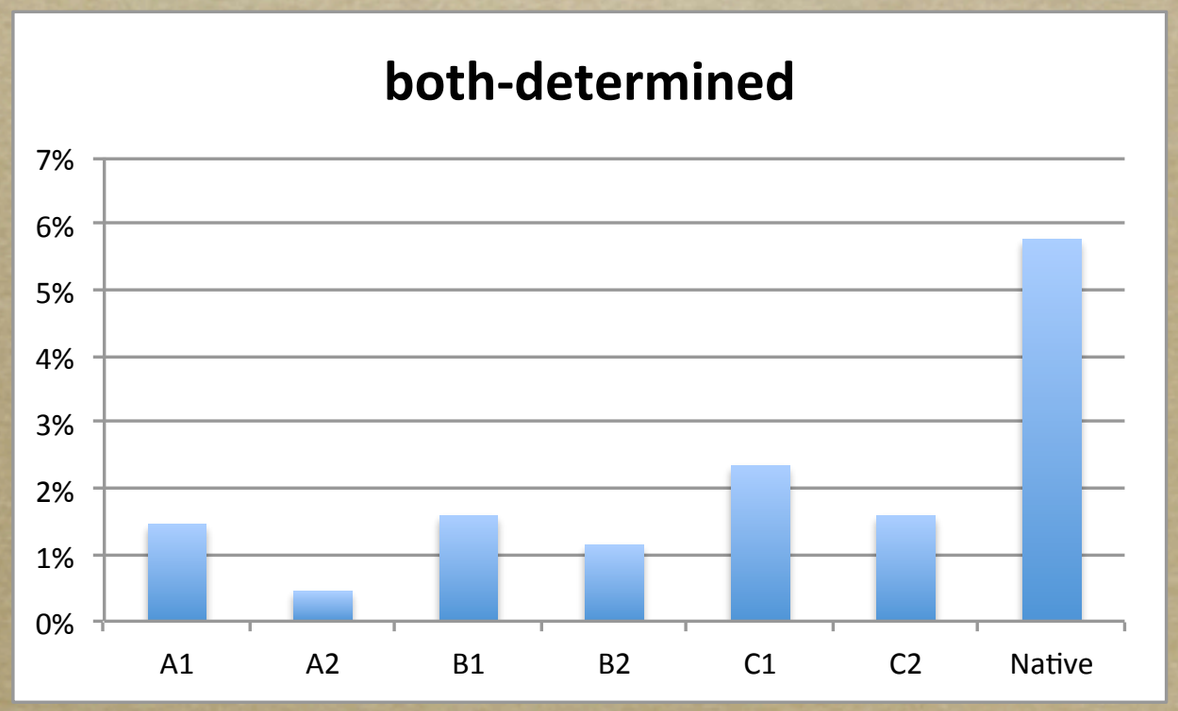
- Special Case: “Much” is wrongly used by many Spanish learners of English, since “mucho” is used for both count and mass (where English uses “much” for mass, “many” for count)
- Advancing learners appear to learn to avoid using it, to avoid errors.



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

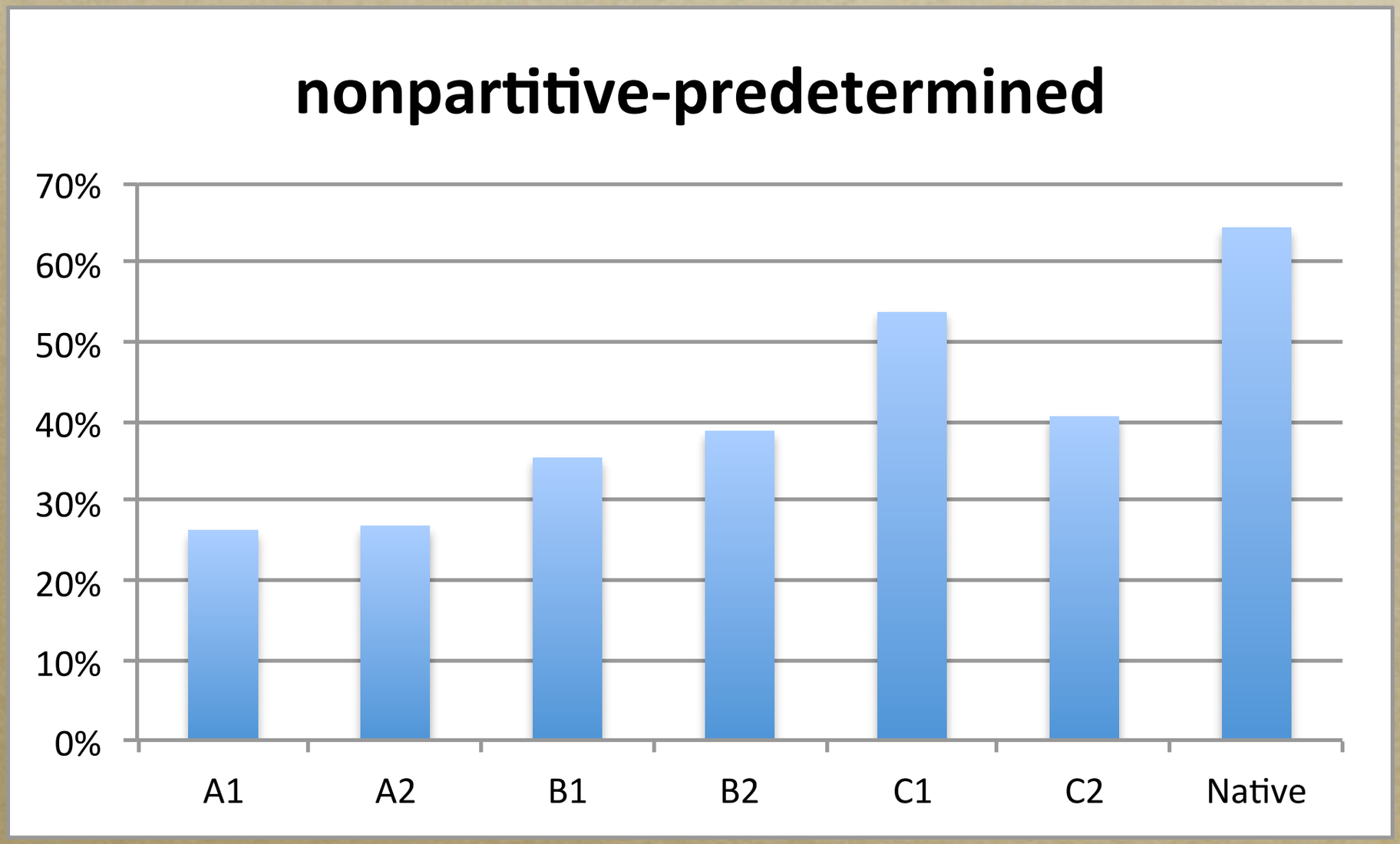
- Dual Determiners: clear that Spanish learners don't use these appropriately.
- While 'either' seems to be acquired with proficiency, both seems not to be properly acquired.



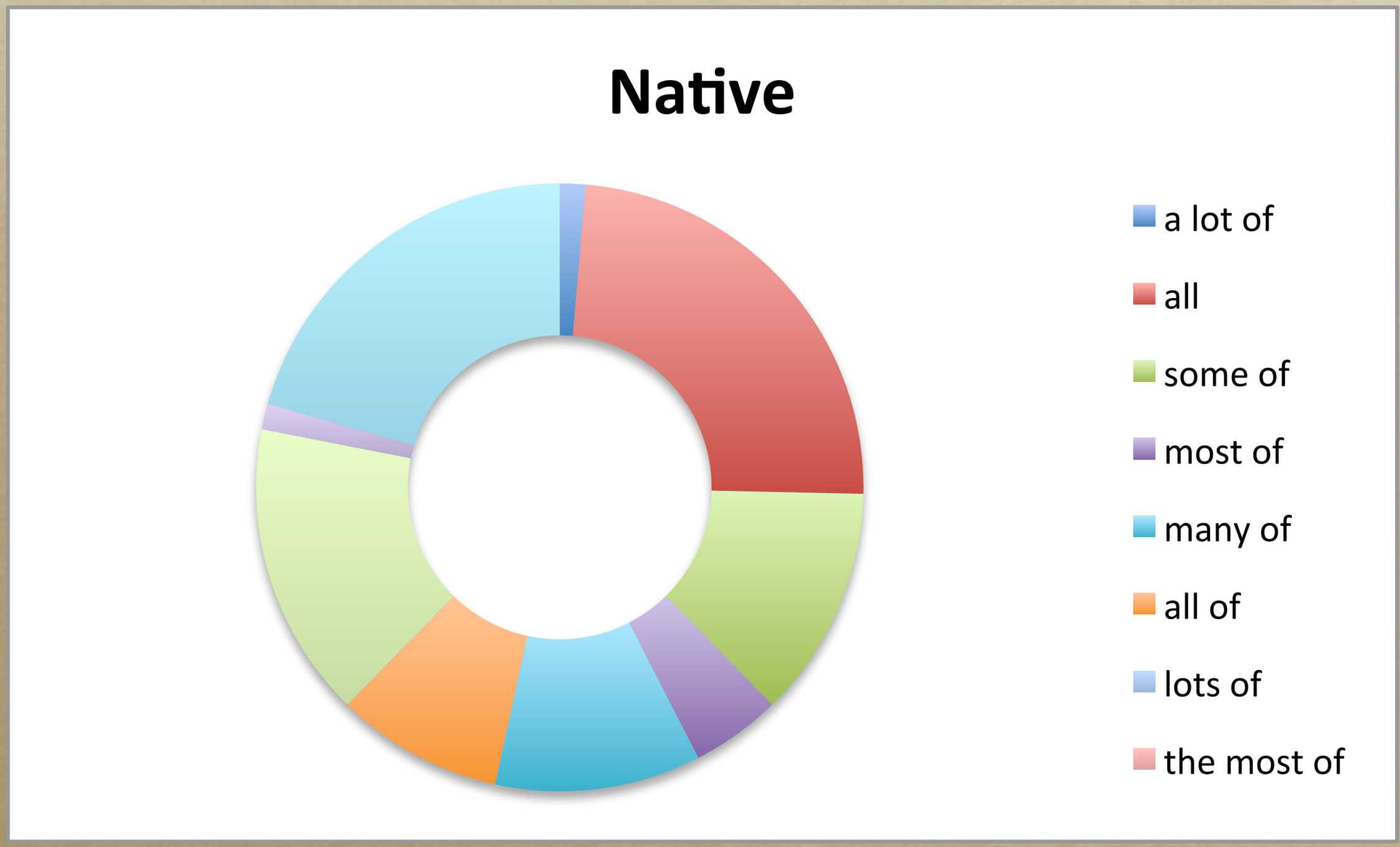


PART 4: Results
(ii) Types of quantifier in PreDet slot

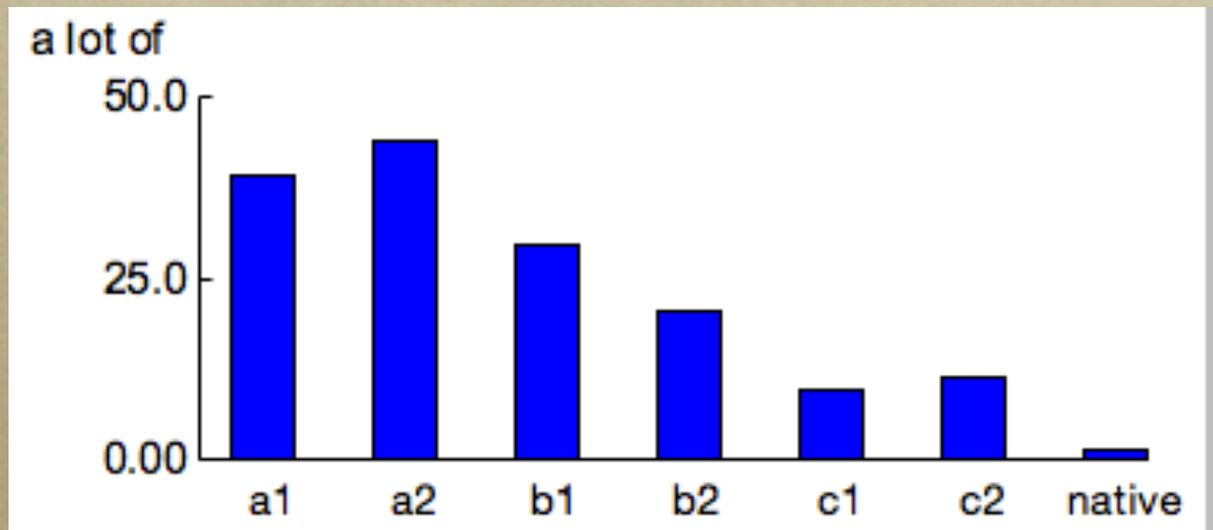
- % of predetermined without "of": "all the people", etc.

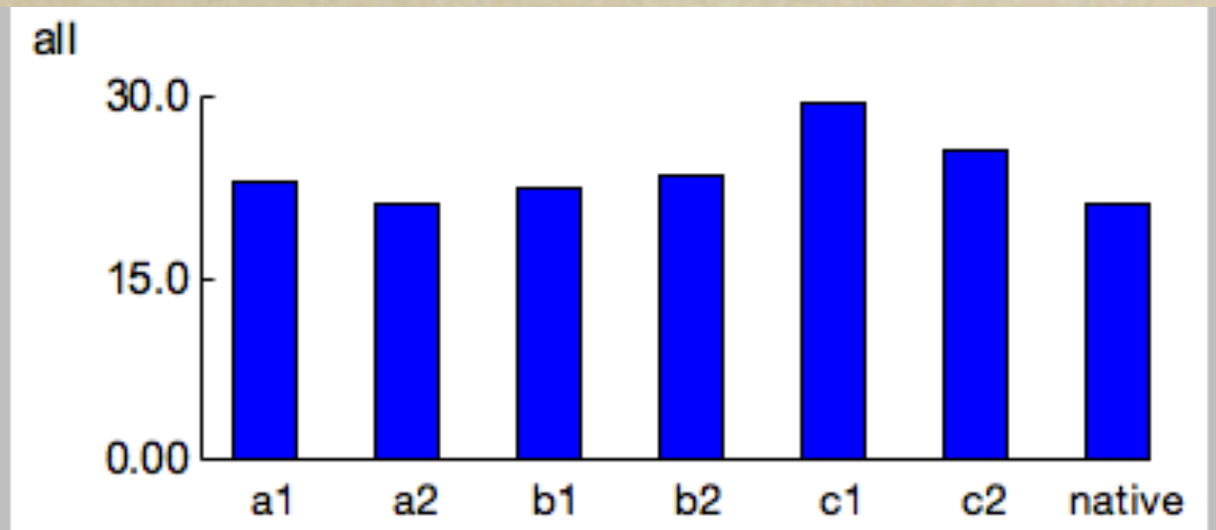


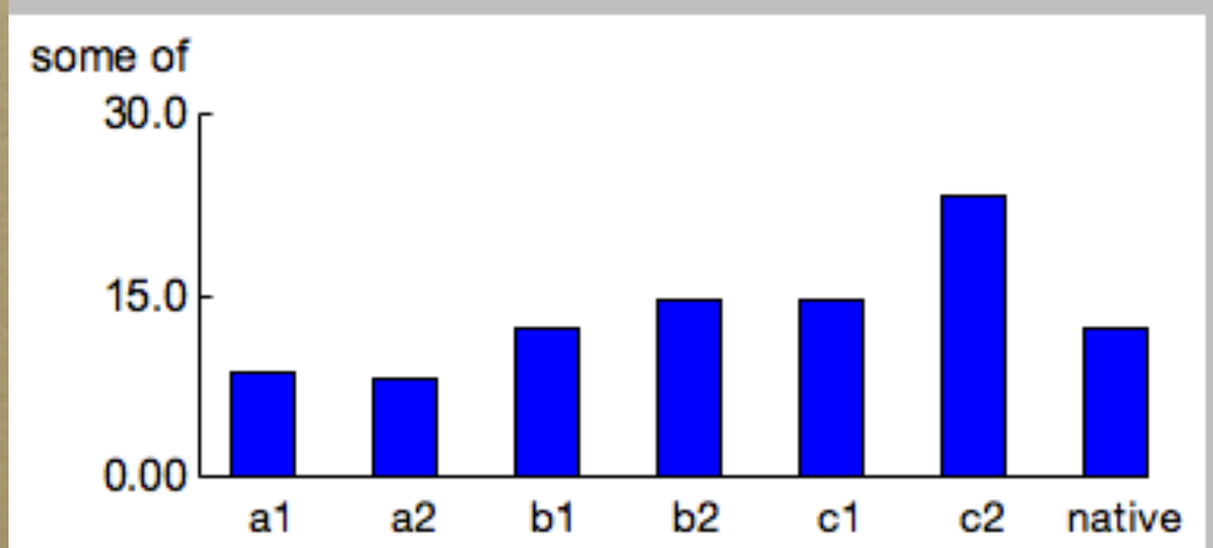
PART 4: Results
(ii) Types of quantifier in PreDet slot

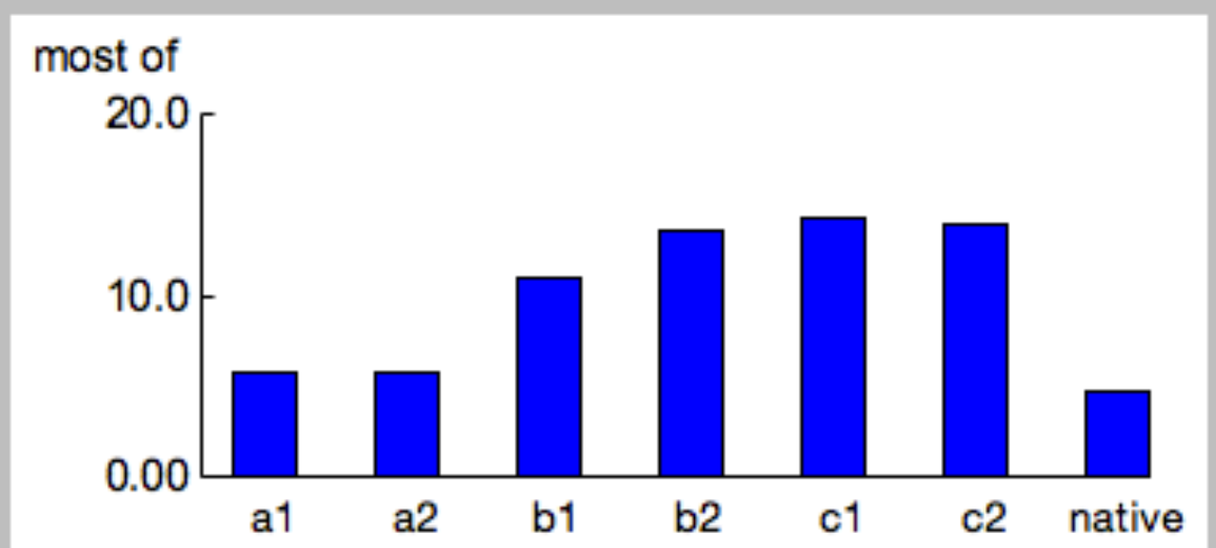


PART 4: Results
(ii) Types of quantifier in PreDet slot



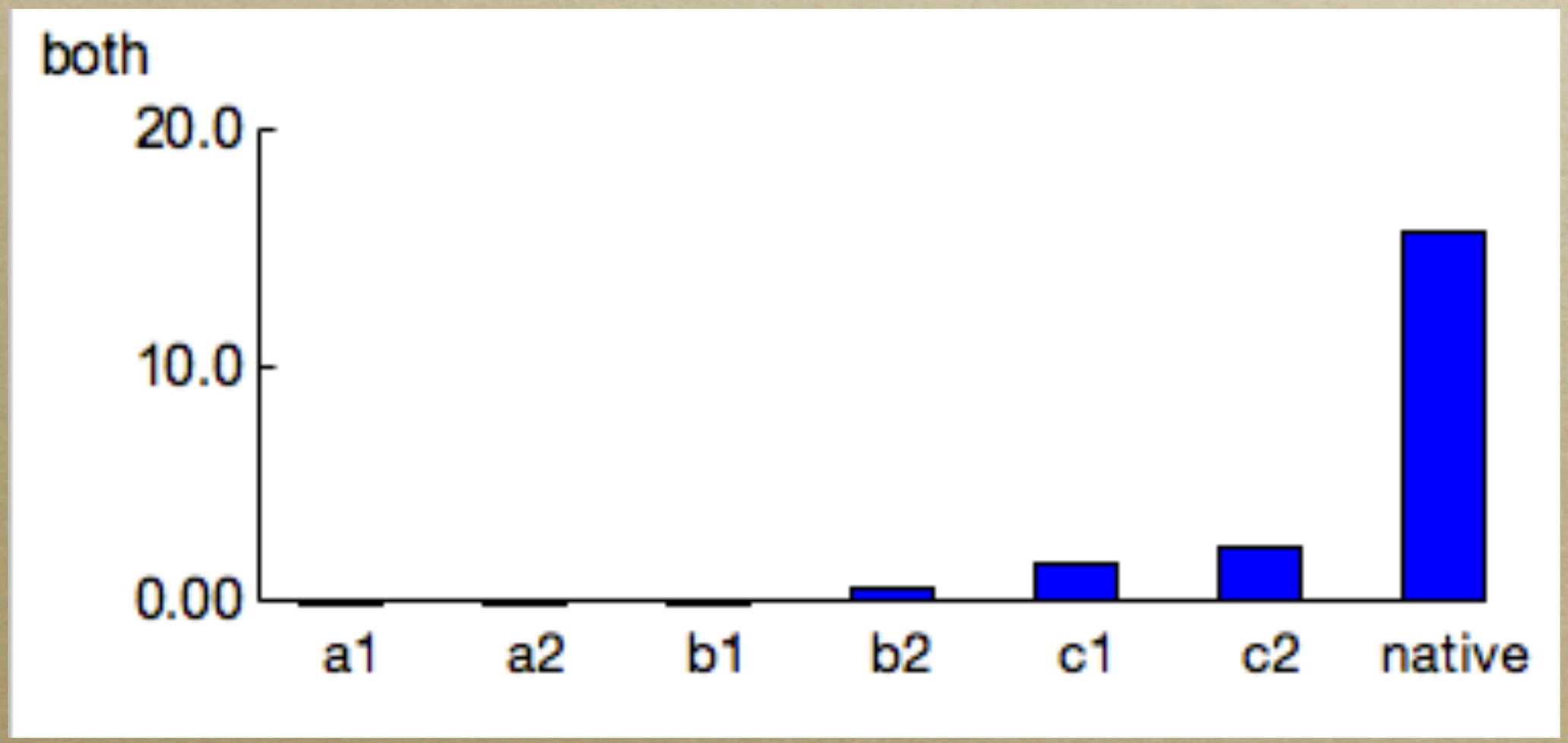






PART 4: Results
(ii) Types of quantifier in PreDet slot

- USE of Dual non partitive: “both my friends”



PART 4: Results
(iii) Common Errors

- Count nouns with mass determiner: “much apples”
- (NO TIME TO FINISH THIS STUDY)

- much people
- so much people
- much conflicts as in traditional families
- very much jobs,1
- very much people 1
- very much families 1
- not so much years 1
- very much traffic jams 1
- much problem 1

PART 4: Results
(iii) Common Errors

Singular nouns with plural determiner:

- (NO TIME TO FINISH THIS STUDY)

- *all type*
- *all kind of comforts*
- *all person*
- *few person*
- *several proyect*
- *all sort*
- *all sort of things*
- *all type of webs*
- *all kind of rehabilitation programmes*

PART 5: Conclusions

- The work has built software to allow automatic detection and tagging of nominal quantification phenomena in previously unseen text.
  - Identification of context of production of the quantification:
    - positive/negative context,
    - speech-act context,
    - count/noncount, etc.

PART 5: Conclusions

- The work has produced a quantification-annotated corpus of learner English.
- The patterns in this corpus will be used to inform our grammar teaching within our English Studies degree: focusing our teaching effort on those phenomena:
  - most over/under-used by students
  - most prone to errors