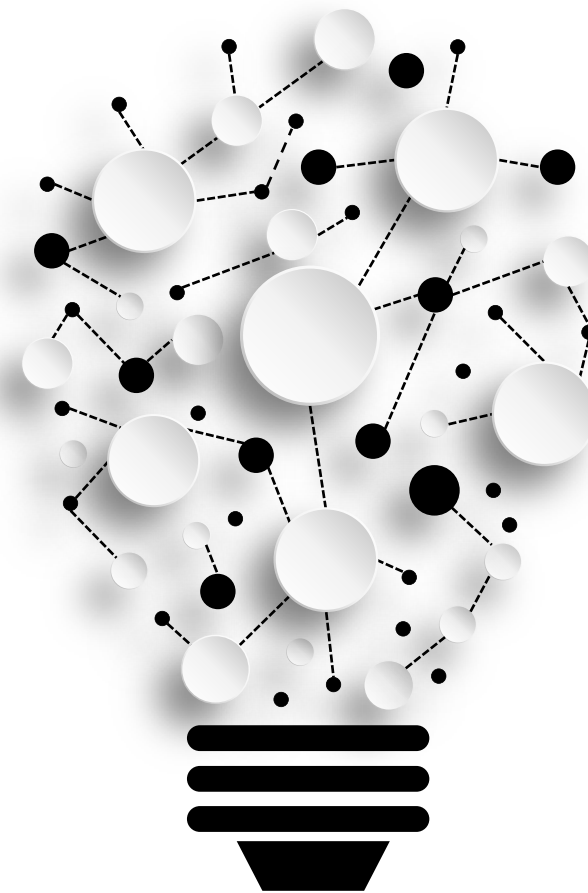


Applying an error corpus to the classroom: from coarse-grained errors to teachable concepts

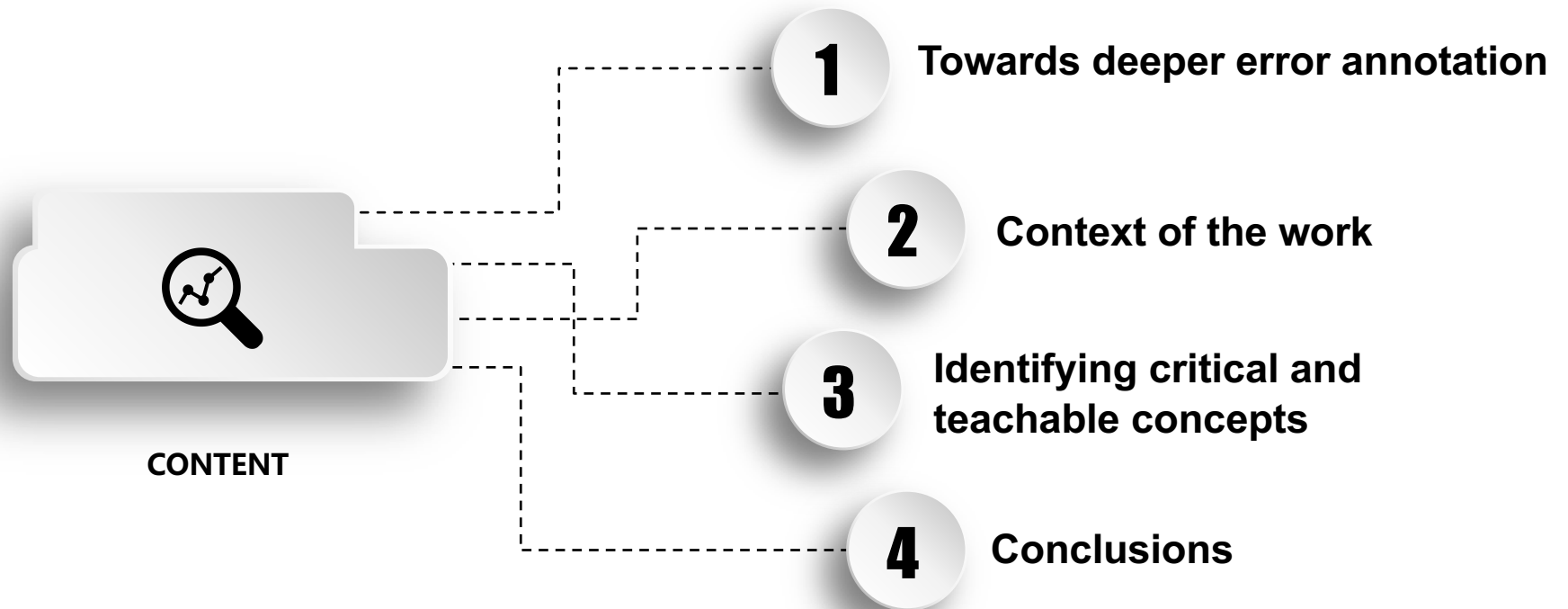
Mick O'Donnell (Universidad Autónoma de Madrid)

Oksana Polyakova (Universidad Católica de València)

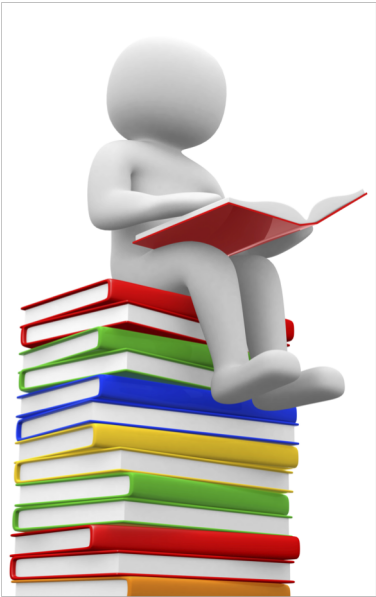
Penny MacDonald (Universitat Politècnica de València)



ALEGRO



Main point:



In the context of exploring grammatical errors in learner corpora to discover what students need to be taught, and at what level:

- Most error annotation projects do not annotate in sufficient detail to properly inform as to the needs of the language learner.

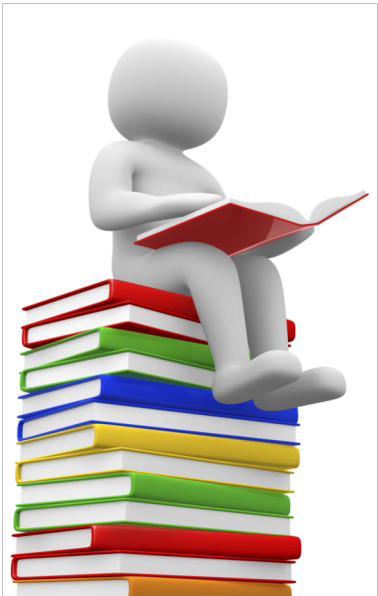
Argument

1. Most error annotation projects annotate:

- The **locale** of the error (e.g., article errors, preposition errors, noun errors, etc.)
- A general **type** of error (insertion, omission, selection, order, spelling, etc.)

→The most frequent errors of a given locale and type reveal general areas that need to be addressed in the learner population

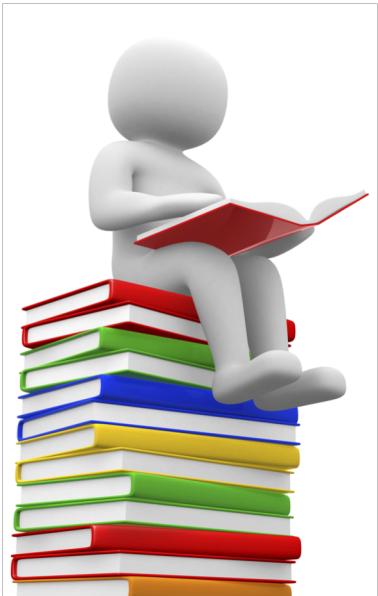
e.g., `article-wrongly-inserted` errors account for around 15% of grammar errors in Spanish learners of English.



Tagging system developed by:	Codes:	Learner corpus:
-Hutchinson, 1996 -Dagneaux, Denness, Granger, & Meunier, 1996	40 codes. Main categories: Form, Grammar, Lexis, Register, Style, Word Missing, Word Order, Word Redundant, Lexico-Grammar.	ICLE – International Corpus of Learner English
Nicholls (2003)	Learner Error Coding System. Two-letter system: General type of error (e.g. wrong form, omission) + word class of word. 88 codes, e.g. F wrong Form + C conjunction M something Missing + N noun	Cambridge Learners Corpus (Cambridge University Press)
Izumi et al. 2005	Morphological, grammatical, lexical errors + naturalness, intelligibility	NICT JLE (Japanese Learner English)
Díaz-Negrillo & García-Cumbreras 2007	EARS annotation system Six linguistic levels (spelling, punctuation, word grammar, clause grammar, phrase grammar or lexis) + -unit involved in the error -category associated with it -distinction between usage and realization errors -surface structure modification classification	Spanish university students doing English Studies at University of Granada
O'Donnell 2008	UAM Corpus Tool. 6 main categories: Lexical, Grammatical, Phrasing, Pragmatic, Punctuation and Uncodable + 130 leaf features	Wricle Corpus (Rollinson and Mendikoetxea 2010) + UPV learner corpus (Andreu-Andrés et al. 2010)

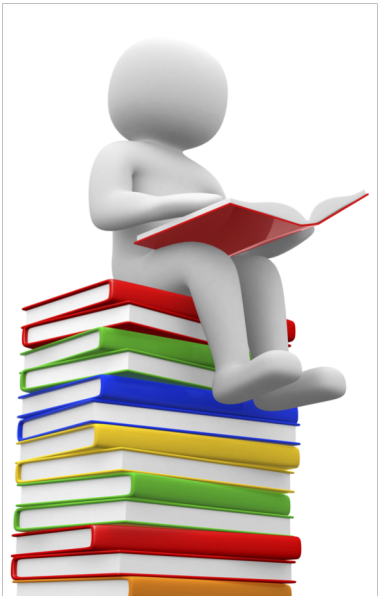
Argument

2. But one cannot teach “Use less articles”!
 - Learners need clear directions as to when the structure should be used or not used.
3. The error annotation at this level of delicacy does not reveal which factors condition the appropriateness of the structure.



Towards teachable concepts.

3. To properly correct learner's errors, one must understand in which contexts of use the structure is most problematic for the students.
 4. One thus needs to extend the coding on these critical error types with more delicate features identifying the underlying cause of the error.
- We need to know not only WHERE the errors occur, but WHY they are occurring.



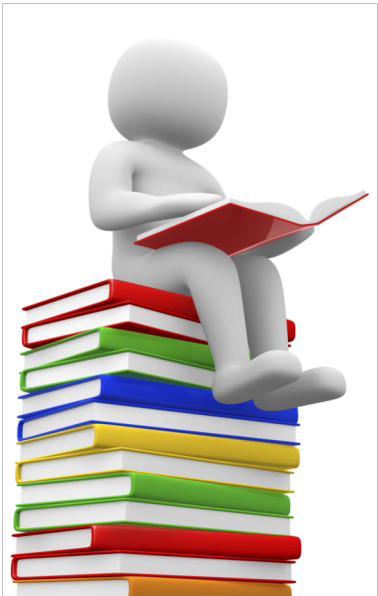
Towards teachable concepts

We need to use the corpus to identify not only general areas which are critical for the student, but also to help us identify concepts that are teachable:

Teachable concept: A rule or concept that the student can apply in their own practice to resolve whether a language instance is valid or not.


- **Unteachable:** Use less articles!
- **Teachable:** When a plural noun is used to refer to a class of entities (generic reference) do not use an article.

The drugs are a problem for some students.





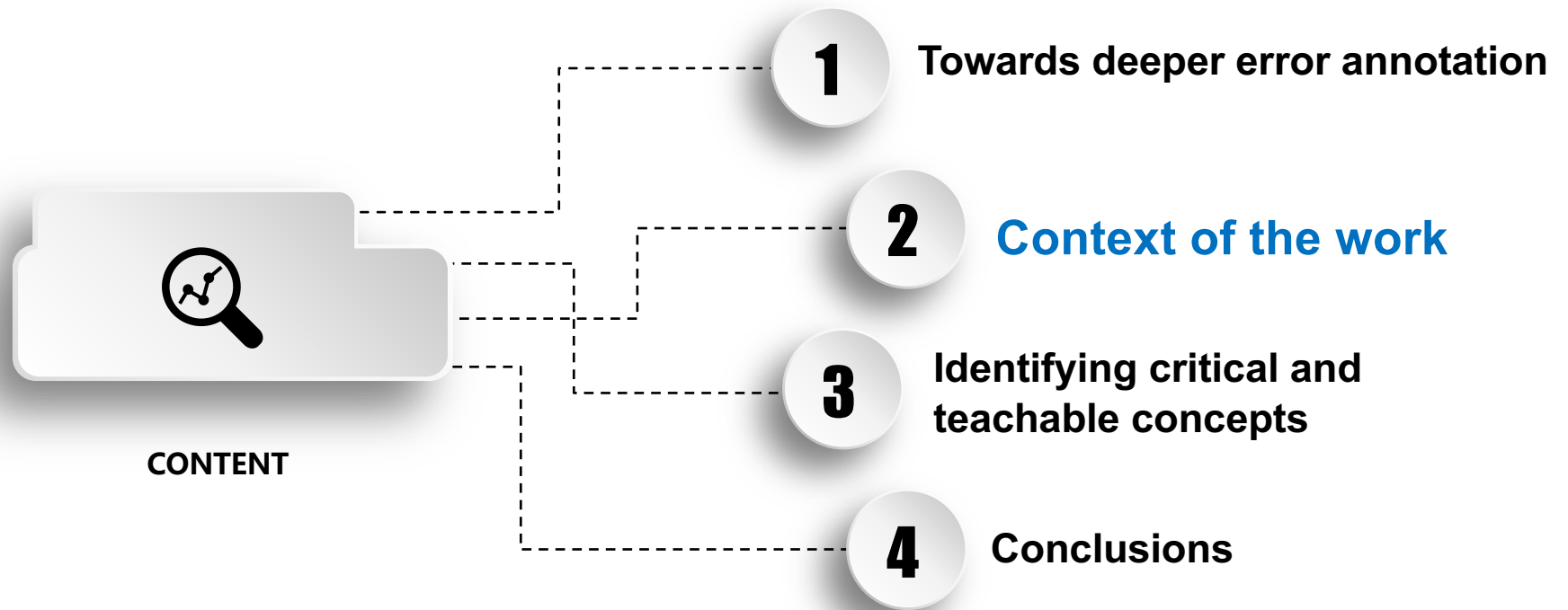
By “language concept” we mean:

- A packet of information a speaker needs to produce the language well.
 - (mostly equivalent to “rule”)
- 

E.g.

- When referring to two items, use a dual determiner.
- “both” is a dual determiner.
- “both”+NOUN is plural

ALEGRO



The ALEGRO Project



Adaptive
Learning of
English
GRammar
Online

A cooperation between:

- Universidad Autónoma de Madrid,
- Universitat Politècnica de València
- Universitat de València

Funded by the Ministerio de Economía y Competitividad 2016-2018
(FFI2015-67992-R)

The Alegro System: Goals



- The goal of our project is to develop an **online learning system** to assist our Spanish University students in the acquisition of important grammatical concepts.
 - Learners **choose** grammatical **themes** to study (article usage, quantifiers, etc.) or certain topics are suggested for practice after doing an initial test
 - Presented with **explanations** of concepts
 - Can take **quizzes** on the concepts.
- **Key element:** system is **adaptive**: it tracks learner assimilation of concepts via the quizzes and tailors the learner experience on that basis.

Indicate which sentence is correct:

☐ A Both my twin daughters are coming home for Christmas.

☐ B Each my twin daughters are coming home for Christmas.

☐ C Either my twin daughters are coming home for Christmas.

INTRODUCTION: The ALEGRO System

- The system is **adaptive** in two ways:



1. **Critical concepts**: Only addresses the 1000 or so grammatical concepts which are most critical for this learner group.
2. **Timely concepts**: Focus student attention on exactly those grammatical concepts which are within their Zone of Proximal Development (Vygotsky 1978).



INTRODUCTION: The ALEGRO System

- The system is **adaptive** in two ways:



1. **Critical concepts**: Only addresses the 1000 or so grammatical concepts which are most critical for this learner group.

This talk

2. **Timely concepts**: Focus student attention on exactly those grammatical concepts which are within their Zone of Proximal Development (Vygotsky 1978).



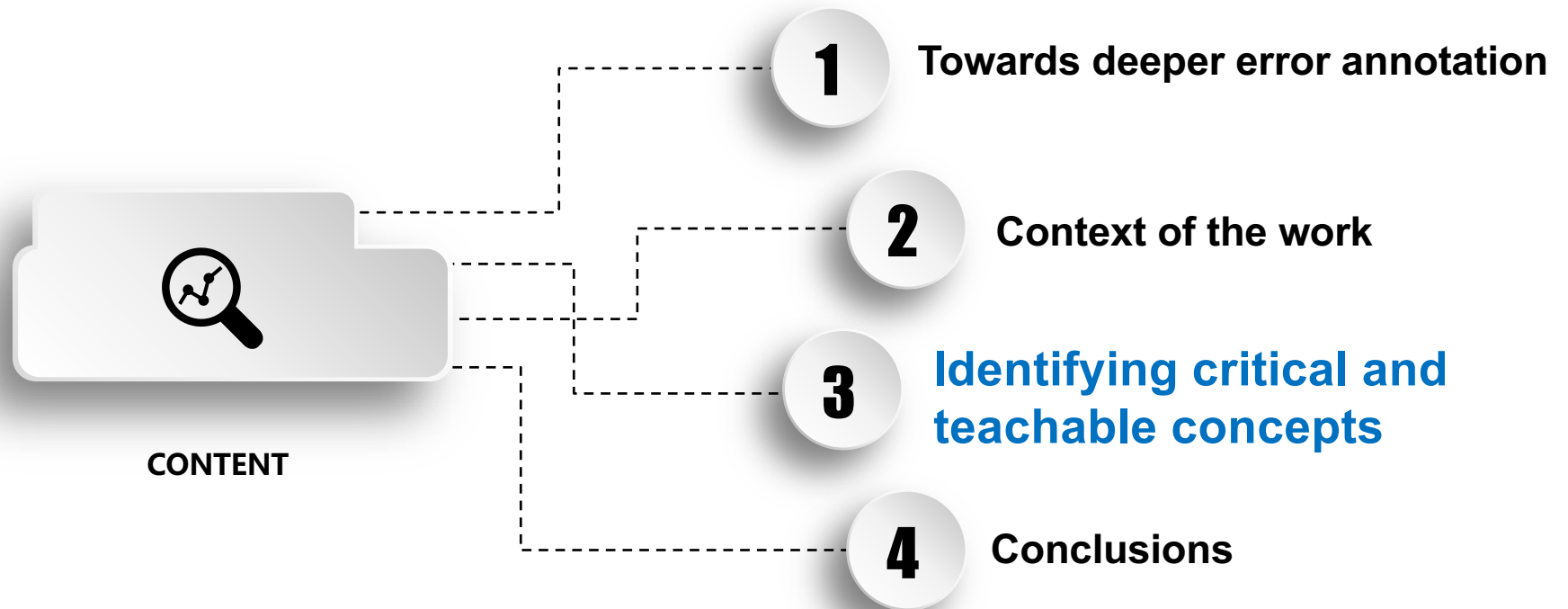
The ALEGRO System: Approach

Critical language concepts:

- Not all concepts involved in a language are equally important for a given L1/L2 combination:
- Some can be transferred easily from the L1.
- Other concepts difficult to learn but infrequently called upon.
- THUS: only teach concepts which frequently lead to errors in learner productions.



ALEGRO



The ALEGRO System: Approach

Critical concepts from error-coded learner corpus

- We want the system to focus on the **most critical grammar concepts** for a specific user group: Spanish University learners of English.
- We can study grammatical **errors** made by this group to identify their critical grammar concepts.
- Thus, we turn to our error-annotated learner corpus.



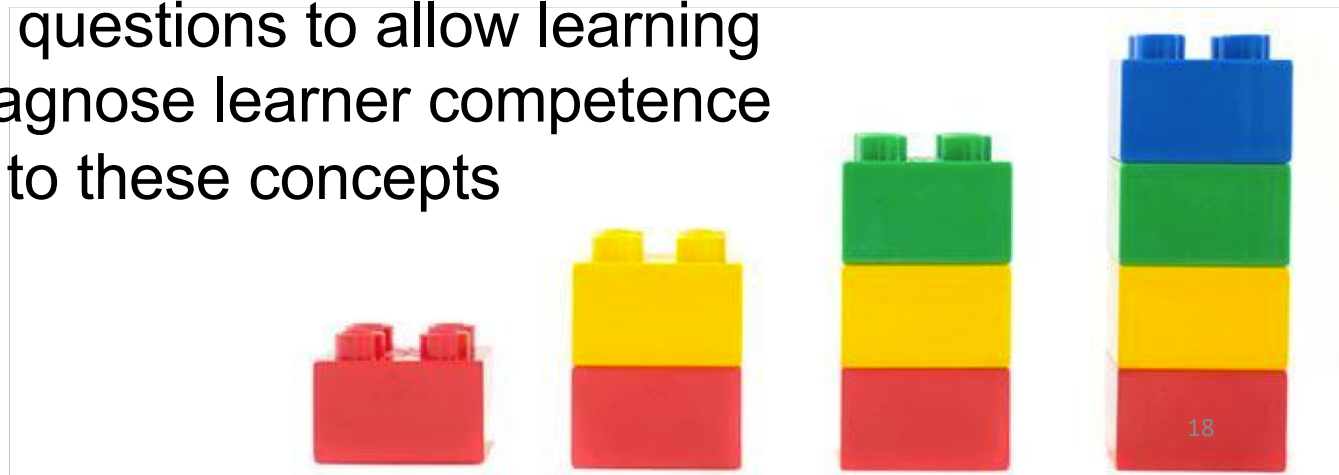
Stages of Alegro Data Implementation

Stage 1 : Identify Critical Language Areas

Stage 2 : Identify Critical Language Concepts

Stage 3 : Provide teaching material for critical concepts

Stage 4 : Provide quiz questions to allow learning system to diagnose learner competence with respect to these concepts



Stages of Alegro Data Implementation

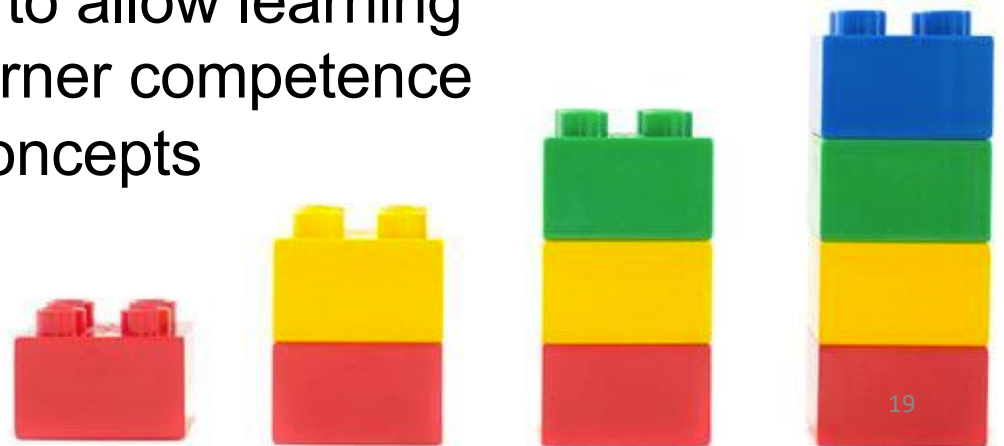
Stage 1 : Identify Critical Language Areas

Stage 2 : Identify Critical Language Concepts

Covered in
this talk

Stage 3 : Provide teaching material for critical
concepts

Stage 4 : Provide quiz questions to allow learning
system to diagnose learner competence
with respect to these concepts

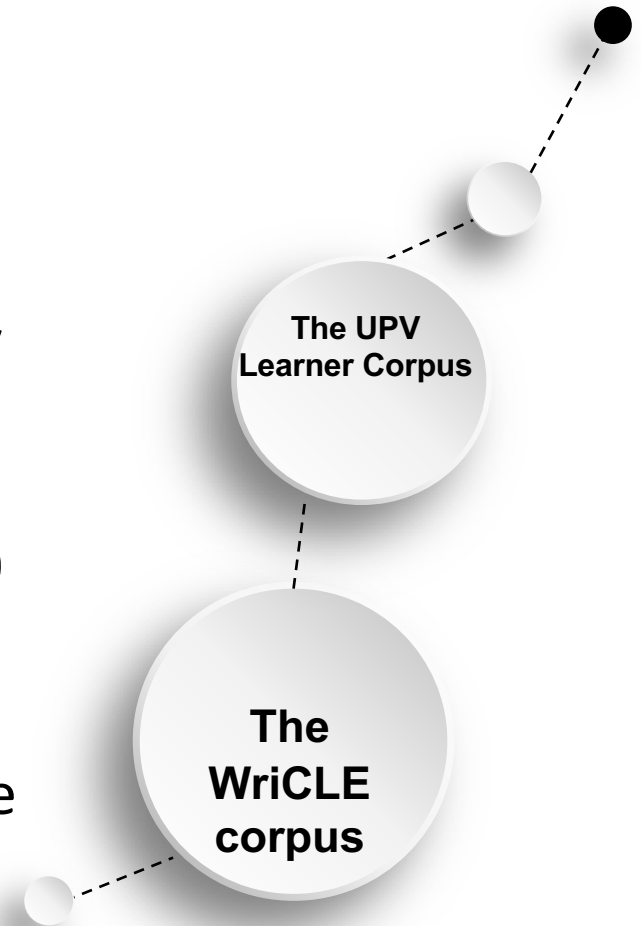


Stage 1: From Learner Corpus to Critical Language Areas

The project makes use of two corpora:

- The **WriCLE** corpus (UAM) - *Written Corpus of Learner English*. 521 essays of ~1000 words each, written by Spanish learners of English at University level (about 500,000 words) (Rollinson and Mendikoetxea 2008)
- The **UPV Learner Corpus** (UPV) containing 150,000 words of shorter texts by ESP students (Andreu Andrés et al. 2010)

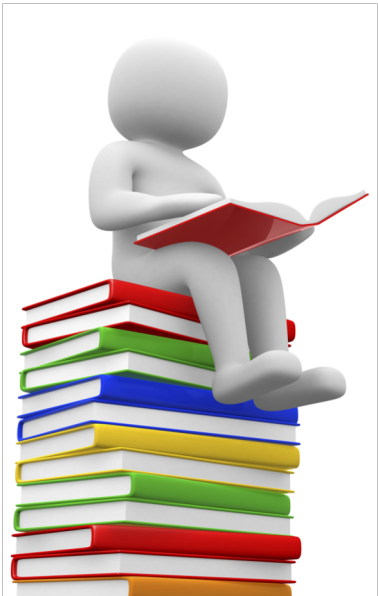
Oxford Placement test given at same time, to measure proficiency



Stage 1: From Learner Corpus to Critical Language Areas

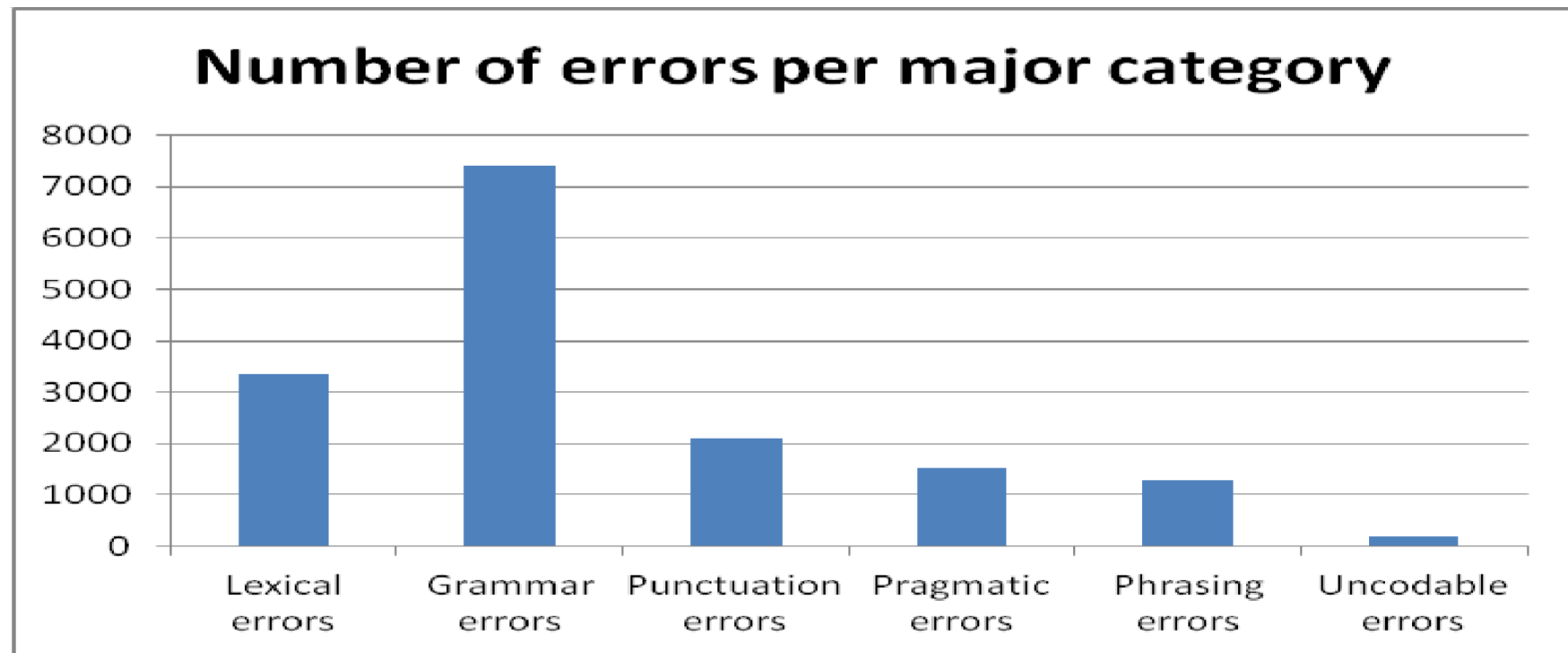
The Error corpus uses a sub-section:

- 307 essays (109,974 words) analysed
- Annotated using UAMCorpus Tool within the TREACLE project (2009-12)
- 16,200 errors identified
- 7,400 errors are grammar related



Stage 1: From Learner Corpus to Critical Language Areas

- The error coding scheme contains six main categories of errors with **138** error features in total



Stage 1: From Learner Corpus to Critical Language Areas

<u>Linguistic feature</u>	<u>MAIN ERROR TYPE</u>	<u>Subcategories Grammar error</u>		<u>Most delicate level</u>
<u>ERROR</u>	Punctuation	NOUN PHRASE	DETERMINER	determiner-order
	Lexical	Adjectival-phrase-error	Premodifier	determiner-present-not-required
	GRAMMAR	Adverb-phrase-error	Head	determiner-absent-required
	Pragmatic	Prep-phrase-error	Postmodifier	determiner-choice
	Phrasing	Verb-phrase-error	NP-complex error	determiner-agreement
	Uncodable	Clause-error	Proper name error	
		Clause-complex-error	Pronoun error	
		Special-structure-error		
		Other-grammatical-error		

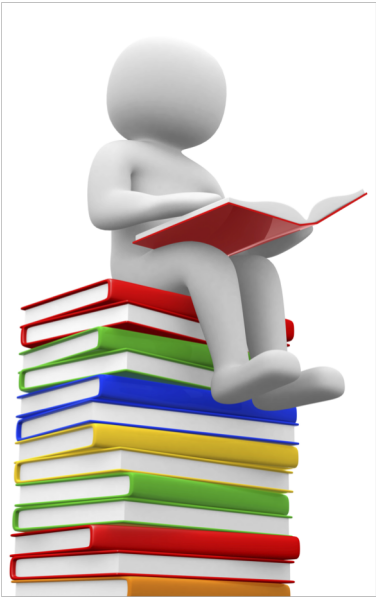
ALEGRO Methodology

- We decided to focus on the 20 most frequent grammatical errors, as these account for 73% of all of our learner's grammar errors. Top 12 shown below:

Topic	Error	Count	% (of Gramm. Errors)
Determiner	det-present-not-required	1087	14.7%
	det-absent-required	439	5.9%
	determiner-choice	250	3.4%
	determiner-agreement	231	3.1%
Head	wrong-number	408	5.5%
	pronoun-choice-error	134	1.8%
	wrong-category	122	1.6%
Preposition	preposition-choice	823	11.1%
	unnecessary-preposition	205	2.8%
Clause	subject-finite-agreement	536	7.2%
	obligatory-subject-absent	227	3.1%
	adjunct-order	179	2.4%

Stages of Alegro

Stage 2 : Identify Critical Language Concepts

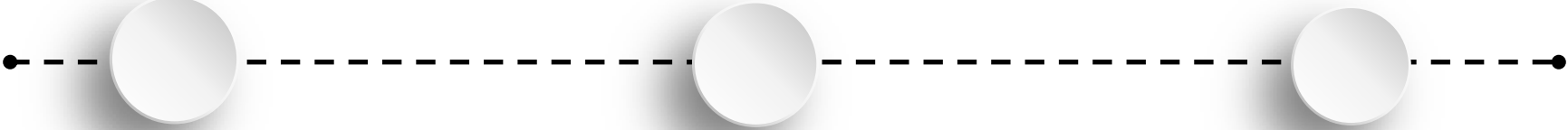


- Our error categories are not by themselves ‘concepts’ that can be taught/learnt.
- Each structural error can result from a range of misunderstood concepts
- We are coding the most critical error types more finely in terms of the grammatical rule/concept broken.
- This process reveals which language concepts are most critical to the learner.

ALEGRO

ALEGRO Methodology

Critical concepts from error-coded learner corpus



01. Identify the 20 most frequent grammatical errors.

02. For each error category,

- Examine each error instance in turn
- Identify the grammatical concept(s) that were not understood to produce that error.
- Tag the error with that grammatical concept.

03. Over the corpus, identify the grammatical concepts that most often caused the errors.

Stage 2: Identifying critical concepts

The use of the article by Spanish learners of English breaks down into several component uses:

- Referring to **specific** entities
 - Normal “the president” / “el presidente”)
 - Percentages: “10 percent” / “**el** 10 por ciento”
 - Places of work etc.: “go to university” / “ir a **la** universidad”
 - Meals: “after breakfast” / “después **del** desayuno”
- Referring to **generic** entities:
 - Count: singular - “the cat” / “el gato”
 - Count: plural - “Cats” / “**los** gatos”
 - Noncount: - “Love” / “**el** amor”

Stage 2: Identifying critical concepts

The use of the article by Spanish learners of English breaks down into several component uses:

- Referring to **specific** entities

- Normal “the president” / “el presidente”
- Percentages: “10 percent” / “**el** 10 por ciento”
- Places of work etc.: “go to university” / “ir a **la** universidad”
- Meals: “after breakfast” / “después **del** desayuno”

- Referring to **generic** entities:

- Count: singular - “the cat” / “el gato”
- Count: plural - “Cats” / “**los** gatos”
- Noncount: - “Love” / “**el** amor”

Cases that cause nearly all errors

Stage 2: Identifying critical concepts

- Finer coding of the errors of this category revealed 6 distinct language concepts students need to acquire:

Error	Broken concept
<i>The terrorism is bad.</i>	Generic noncount don't take article
<i>The cats are mammals.</i>	Generic plurals don't take articles
<i>The seventy percent of...</i>	Percentages don't take article
<i>I study in the university</i>	Places of work/internment don't take article
<i>See you after the coffee</i>	Mealnames don't take article
<i>The most of my friends...</i>	<i>Most</i> as predeterminer doesn't take article

Stage 2: Identifying critical concepts

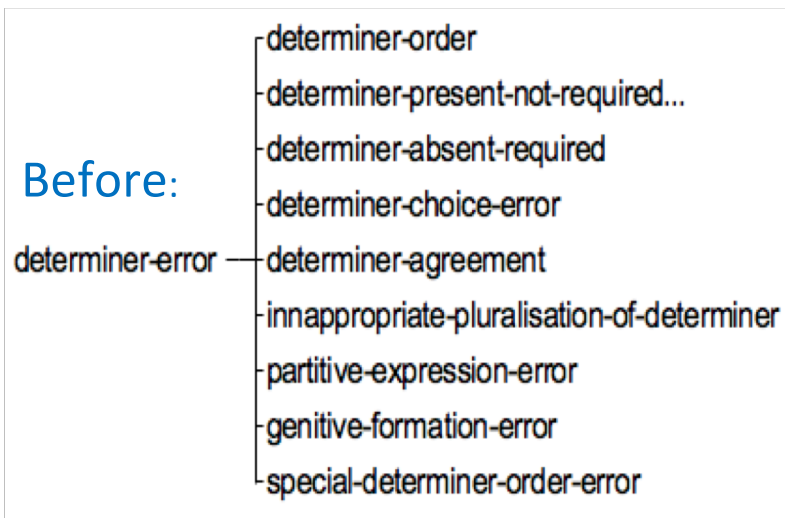
- But these errors are critical to different degrees:

Error	Broken concept	% errors
<i>The terrorism is bad.</i>	Generic noncount don't take article	57%
<i>The cats are mammals.</i>	Generic plurals don't take articles	33%
<i>The seventy percent of...</i>	Percentages don't take article	3%
<i>I study in the university</i>	Places of work/internment don't take article	1%
<i>See you after the coffee</i>	Meal-names don't take article	1%
<i>The most of my friends...</i>	<i>Most</i> as predeterminer doesn't take article	1%

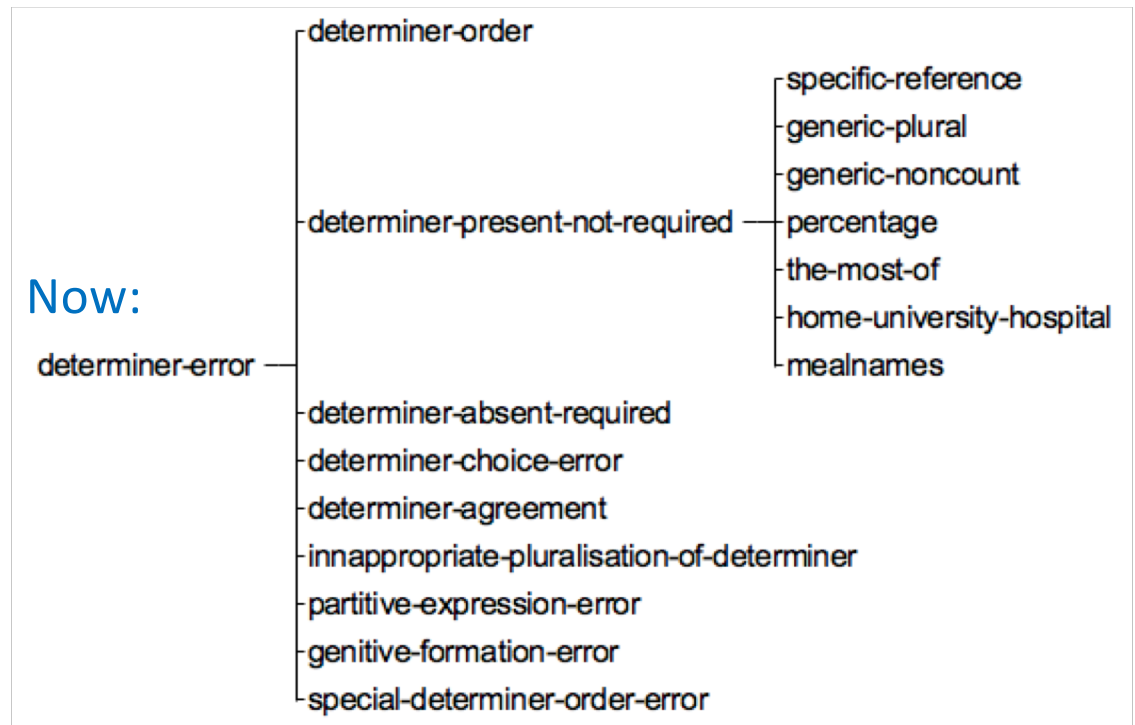
Stage 2: Identifying critical concepts

- The approach is data driven, not theory driven:
 - we examine each error in turn,
 - determining the underlying broken rule/concept, and tag the instance.
- The list of explanations evolves through the examination.

Before:



Now:



Stage 2: Identifying critical concepts

- We are not talking about just more fine-coding of error categories by itself.
- The finer coding must be in terms of **teachable concepts**.
- Example:
 - a student was assigned the task of exploring: preposition-insertion errors (e.g., *He gave **to** me a book*)
 - Initial coding in terms of which prep was inserted ('to', 'of', etc.)
 - But this did not identify the underlying cause of the error.

Stage 2: Identifying critical concepts

- Preposition-insertion errors: discovered critical and teachable concepts:
 - Don't use 'to recipient' after verb: *He gave ~~to~~ John a book.*
 - Some locations don't require 'to': *I will go ~~to~~ anywhere.*
 - Don't use 'of' after cardinal pre-modifier: *6 million ~~of~~ jobs.*
 - Don't use 'to' preposition with some verbs: *He faced ~~to~~ his problem*
(*join, see, affect, force, kill, hit, reach, concern, consider, despise, attend, survive, avoid, exploit, help, interest, prejudice, enter, visit*)
[The Spanish equivalent DOES require the preposition]
 - Etc.

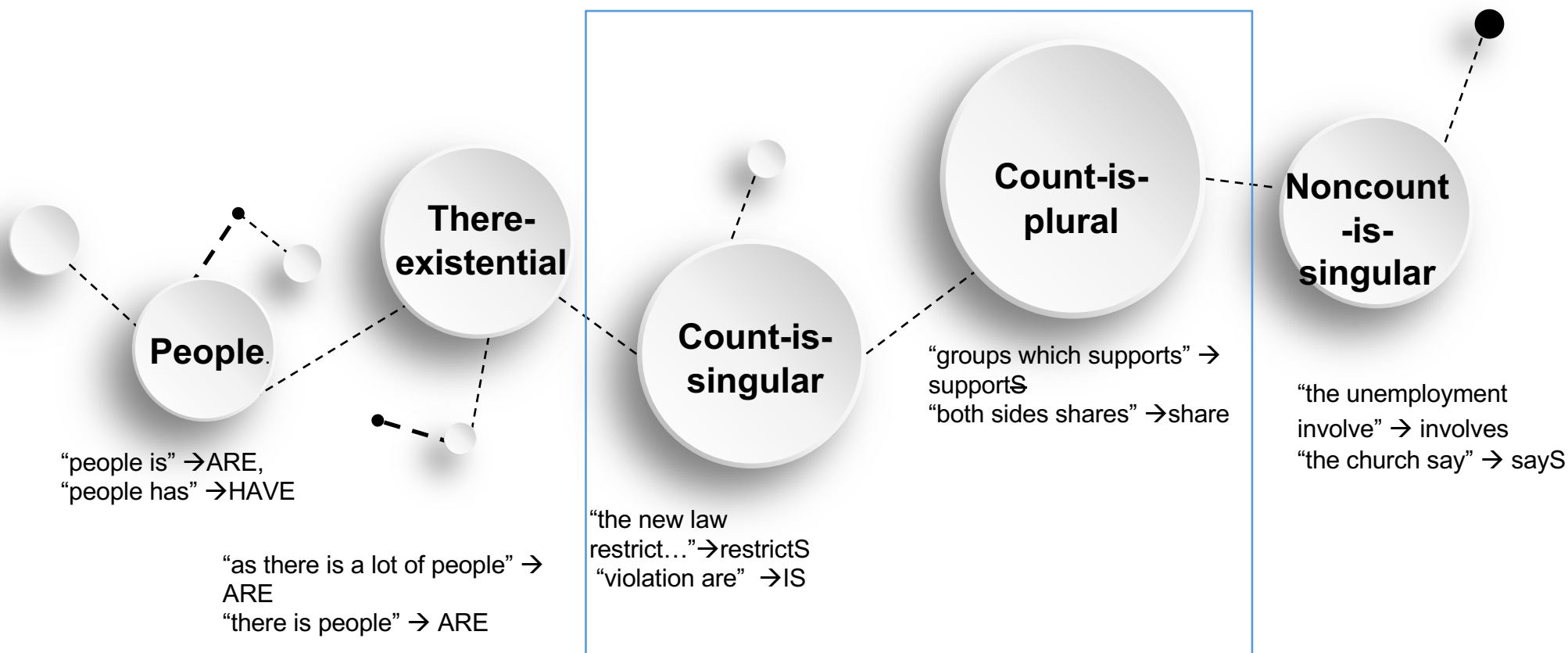
Stage 2: Identifying critical concepts

All errors involve multiple language concepts

- Even an apparently single-concept error really requires multiple concepts to explain the errors: e.g., **subject-finite-agreement**
 - **Core rule:** The Subject and the Finite should agree in number.
- But some cases of errors explained by other missing concepts:
 - ‘people’ is plural: *The people is not happy.* (Spanish ‘gente’ is singular)
 - ‘everybody/everyone’ is singular.
 - Subject in existential follows verb: *There is many reasons...*
 - ‘you’ (singular) has plural finite: *You is ...*
 - Conjunction of singulars is plural: *John and Mary is ...*

ALEGRO Methodology

Explanations not always clear - subject-finite-agreement



Some concepts cause distinct error types

- Some language concepts can affect multiple points of structure
 - Error-type: **Determiner-head-agreement:**
 - Example: *a people*
 - Underlying concept broken: **'people is plural'**
- Error-type: **Subject-Finite-agreement:**
- Example: *the people is unhappy*
- Underlying concept broken: **'people is plural'**

Shows that language concepts are not tied exclusively to error-type (not a hierarchical relation)

State of the Work

Some topics covered

- ✓ det-absent-required
- ✓ subject-finite-agreement
- ✓ unnecessary-preposition

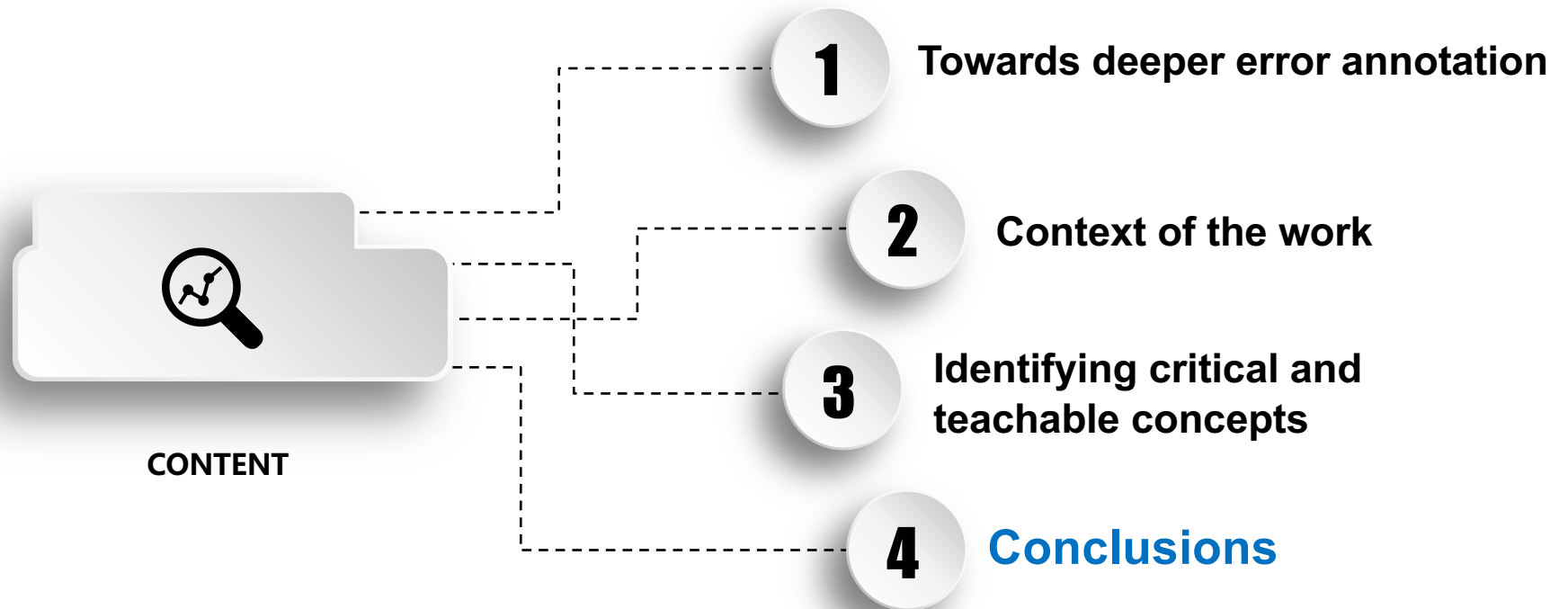
Others currently in progress

- obligatory-subject-absent
- subject-doubling
- preposition-choice
- det-present-not-required
- Quantification-errors

Others still in progress

Some topics covered

ALEGRO



Summary of Talk

- Errors usually defined by locale+operation
- These tags not by themselves detailed enough to support language teaching.
(X Use less articles!)
- We have thus developed a corpus-methodology to identify the critical language concepts behind these errors.
 - More delicate coding of most frequent errors in terms of the underlying rule/concept broken.
- We do not start off with a pre-conceived idea of causes of each error.
- Rather, we advance through the cases one by one, re-using an existing concept, or adding a new one where it is not yet included.

Summary (ii)

- These concepts need to be at a level of specification that can be used by a learner to identify correct/incorrect use.
-> **teachable concepts**
- We then produce frequency lists of these critical concepts to see which of these concepts are most critical for the learner
-> **critical concept**
- These are the concepts given to the learning system to teach to the learner
- Equally, could be used to inform traditional classroom teaching.

Summary (iii)

Downside:

- Error coding is already labor intensive and time costly.
- We are proposing doing even more work.
- An important principle “You get what you pay for”
- We think that the extra work needed to error code down to broken concept level is worth the effort, in terms of what one can do with the results in regards to language teaching.

END