# Investigating Categories in a Syntactically-Annotated Corpus of Second Language Learners of English

Markus Dickinson & Marwa Ragheb

Dept. of Linguistics, Indiana University

Workshop on Categories & Categorization
in First & Second Language Acquisition
7 March, 2014





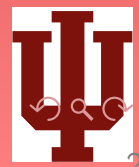
# Starting point

Learner corpora contain the output of second language learners at a certain point in their development

- Regardless of the state of the categories in their interlanguage, we only have the output to annotate







# Starting point

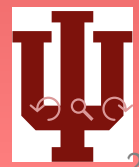
Learner corpora contain the output of second language learners at a certain point in their development

- Regardless of the state of the categories in their interlanguage, we only have the output to annotate

Goal (& challenge) of syntactic annotation:

- Provide linguistic representations of what learners do
  - Researchers can access representations & draw conclusions about the nature of learner categories





# Starting point

Learner corpora contain the output of second language learners at a certain point in their development

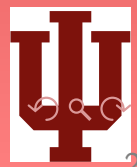
- Regardless of the state of the categories in their interlanguage, we only have the output to annotate

Goal (& challenge) of syntactic annotation:

- Provide linguistic representations of what learners do
  - Researchers can access representations & draw conclusions about the nature of learner categories
- Tension between:
  - Providing detailed representations for researchers
  - Specifying only what can be reliably (& "theory-neutrally") specified







# Starting point

Learner corpora contain the output of second language learners at a certain point in their development

- Regardless of the state of the categories in their interlanguage, we only have the output to annotate

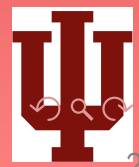
Goal (& challenge) of syntactic annotation:

- Provide linguistic representations of what learners do
  - Researchers can access representations & draw conclusions about the nature of learner categories
- Tension between:
  - Providing detailed representations for researchers
  - Specifying only what can be reliably (& "theory-neutrally") specified

**Today:** Describe our attempts to define categories in a way that leaves full interpretation to a researcher





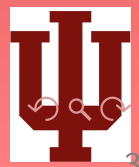


# How to linguistically annotate?

Approaches to annotating linguistic properties:

- In tandem with error annotation, i.e., annotating the category of the *correction*
  - Map to target hypotheses before annotating POS (e.g., Granger 2003) or syntax (e.g., Hirschmann et al. 2010)



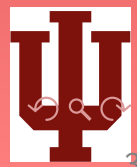


# How to linguistically annotate?

Approaches to annotating linguistic properties:

- In tandem with error annotation, i.e., annotating the category of the *correction*
  - Map to target hypotheses before annotating POS (e.g., Granger 2003) or syntax (e.g., Hirschmann et al. 2010)
- Directly annotate the text (Ragheb and Dickinson 2011; Díaz-Negrillo et al. 2010)





# How to linguistically annotate?

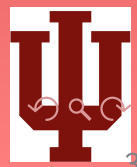
Approaches to annotating linguistic properties:

- In tandem with error annotation, i.e., annotating the category of the *correction*
  - Map to target hypotheses before annotating POS (e.g., Granger 2003) or syntax (e.g., Hirschmann et al. 2010)
- Directly annotate the text (Ragheb and Dickinson 2011; Díaz-Negrillo et al. 2010)

This latter approach allows for a thorough investigation of what learner categories are composed of & is useful for:

- Error detection (Tetreault et al. 2010)
- Learner profiling (Hawkins and Buttery 2010)
- Acquisition research (Ragheb and Dickinson 2011)





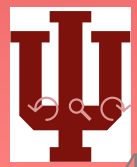
# Our goal

We discuss how our corpus annotation effort defines categories, surveying three aspects:

1. Ill-fitting categories: how do we precisely define types of categories?
2. Broadly-applicable annotation: how can the annotation be applied across various meta-variables?
3. Annotation utilization: how can the resulting annotation be used for research?







# Our goal

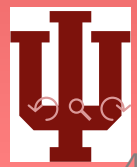
We discuss how our corpus annotation effort defines categories, surveying three aspects:

1. Ill-fitting categories: how do we precisely define types of categories?
2. Broadly-applicable annotation: how can the annotation be applied across various meta-variables?
3. Annotation utilization: how can the resulting annotation be used for research?

We are working to make these points practical, transparent, & able to be debated by annotating c. 10,000 tokens

- The SALLE (Syntactically Annotating Learner Language of English) Project at Indiana University (IU)
- Placement essays from IU, scored on a 1–7 scale



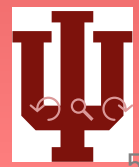


# Ill-fitting categories

We base our annotation on the syntactic categories of the L2 as this is the common goal among learners of that language.





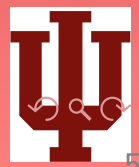


# Ill-fitting categories

We base our annotation on the syntactic categories of the L2 as this is the common goal among learners of that language.

- **Problem:** how does one handle cases that do not fit into a specific L2 category?



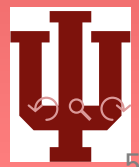


# Ill-fitting categories

We base our annotation on the syntactic categories of the L2 as this is the common goal among learners of that language.

- **Problem:** how does one handle cases that do not fit into a specific L2 category?
- **Solution:** break a category into multiple layers of annotation, based on different kinds of evidence
  - Díaz-Negrillo et al. (2010); Ragheb and Dickinson (2011, 2012)





# Ill-fitting categories

We base our annotation on the syntactic categories of the L2 as this is the common goal among learners of that language.

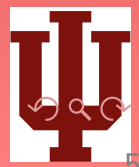
- **Problem:** how does one handle cases that do not fit into a specific L2 category?
- **Solution:** break a category into multiple layers of annotation, based on different kinds of evidence
  - Díaz-Negrillo et al. (2010); Ragheb and Dickinson (2011, 2012)

(1) I have **see** a movie

We annotate two POS tags for *see*:

- morphological properties of a baseform verb (VV0)
- distributional slot of a past participle (VVN)





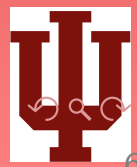
# Making the annotation concrete

This splitting of layers gets us quite far

- But we discovered that this clean split into multiple layers is not really clean at all







# Making the annotation concrete

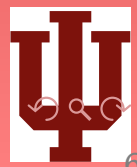
This splitting of layers gets us quite far

- But we discovered that this clean split into multiple layers is not really clean at all

We tried to precisely define three syntactic layers of annotation & discovered many questions & redundancies:

- Subcategorization (e.g., *catch*: <SUBJ,OBJ>)





# Making the annotation concrete

This splitting of layers gets us quite far

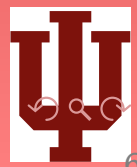
- But we discovered that this clean split into multiple layers is not really clean at all

We tried to precisely define three syntactic layers of annotation & discovered many questions & redundancies:

- Subcategorization (e.g., *catch*: <SUBJ,OBJ>)
- Morphosyntactic dependencies (e.g., 

)





# Making the annotation concrete

This splitting of layers gets us quite far

- But we discovered that this clean split into multiple layers is not really clean at all

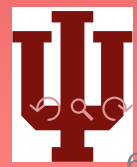
We tried to precisely define three syntactic layers of annotation & discovered many questions & redundancies:

- Subcategorization (e.g., *catch*: <SUBJ,OBJ>)
- Morphosyntactic dependencies (e.g., 

)
- Distributional dependencies (e.g., 

)





# Making the annotation concrete

This splitting of layers gets us quite far

- But we discovered that this clean split into multiple layers is not really clean at all

We tried to precisely define three syntactic layers of annotation & discovered many questions & redundancies:

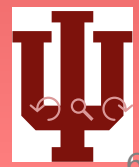
- Subcategorization (e.g., *catch*: <SUBJ,OBJ>)
- Morphosyntactic dependencies (e.g., 

)
- Distributional dependencies (e.g., 

)

**Overall tension:** annotate *every* possible analysis or just contextually-appropriate one(s)?





# Making the annotation concrete

This splitting of layers gets us quite far

- But we discovered that this clean split into multiple layers is not really clean at all

We tried to precisely define three syntactic layers of annotation & discovered many questions & redundancies:

- Subcategorization (e.g., *catch*: <SUBJ,OBJ>)
- Morphosyntactic dependencies (e.g., 

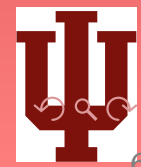
)
- Distributional dependencies (e.g., 

)

**Overall tension:** annotate *every* possible analysis or just contextually-appropriate one(s)?

- We will just sketch the issue: see Ragheb and Dickinson (2012) for more details

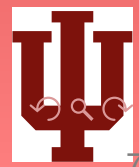




# Sketch: defining morphosyntax

Morphosyntax based on the visible forms of words





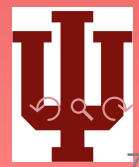
# Sketch: defining morphosyntax

Morphosyntax based on the visible forms of words

- We use distribution to disambiguate:

(2) And also I can heart the step of my **walked** .





# Sketch: defining morphosyntax

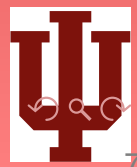
Morphosyntax based on the visible forms of words

- We use distribution to disambiguate:

  (2) And also I can heart the step of my **walked** .

  - 3 POS for *can*, 2 for *step*, 2 for *walked*: 3x2x2=12 trees





# Sketch: defining morphosyntax

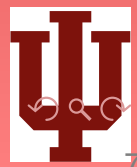
Morphosyntax based on the visible forms of words

- We use distribution to disambiguate:

(2) And also I can heart the step of my **walked** .

  - 3 POS for *can*, 2 for *step*, 2 for *walked*: 3x2x2=12 trees
  - What would be the tree for past tense of *walked*?
    - Adjectival/past participle fits better to the context





# Sketch: defining morphosyntax

Morphosyntax based on the visible forms of words

- We use distribution to disambiguate:

  (2) And also I can heart the step of my **walked** .

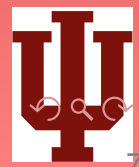
  - 3 POS for *can*, 2 for *step*, 2 for *walked*: 3x2x2=12 trees
  - What would be the tree for past tense of *walked*?
    - Adjectival/past participle fits better to the context

- Distribution likewise uses morphology

  (3) ... having **an**$_{DET}$ experience

  - MOD & QUANT are possible here, but *an* = DET





# Sketch: defining morphosyntax

Morphosyntax based on the visible forms of words

- We use distribution to disambiguate:

  (2) And also I can heart the step of my **walked** .

  - 3 POS for *can*, 2 for *step*, 2 for *walked*: 3x2x2=12 trees
  - What would be the tree for past tense of *walked*?
    - Adjectival/past participle fits better to the context

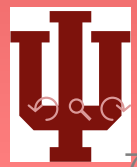
- Distribution likewise uses morphology

  (3) ... having **an**$_{DET}$ experience

  - MOD & QUANT are possible here, but *an* = DET

**Overall point:** The layers are inter-defined





# Sketch: defining morphosyntax

Morphosyntax based on the visible forms of words

- We use distribution to disambiguate:

  (2) And also I can heart the step of my **walked** .

  - 3 POS for *can*, 2 for *step*, 2 for *walked*: 3x2x2=12 trees
  - What would be the tree for past tense of *walked*?
    - Adjectival/past participle fits better to the context

- Distribution likewise uses morphology

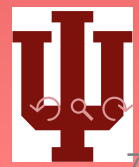
  (3) ... having **an**$_{DET}$ experience

  - MOD & QUANT are possible here, but *an* = DET

**Overall point:** The layers are inter-defined

- We do not annotate distributional dependencies: mostly covered by other layers



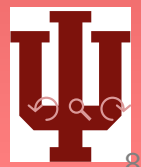


# Broadly-applicable annotation

The annotation is based on linguistic evidence from the L2 grammatical system, making it applicable:

- to learners with different L1s
- to learners at different stages of interlanguage development
- across contexts (e.g., text type)
- regardless of one's (SLA) theory



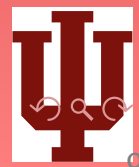


# Broadly-applicable annotation

Three aspects of this broad applicability to examine:





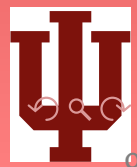


# Broadly-applicable annotation

Three aspects of this broad applicability to examine:

- **Before:** Heuristics which allow us to take a set of possible interpretations & land on a single annotation
  - The goal is to be robust to learner innovations & cases lacking evidence or with ambiguous evidence



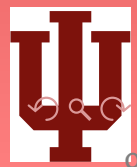


# Broadly-applicable annotation

Three aspects of this broad applicability to examine:

- **Before:** Heuristics which allow us to take a set of possible interpretations & land on a single annotation
  - The goal is to be robust to learner innovations & cases lacking evidence or with ambiguous evidence
- **During:** Applying the annotation to difficult usage shows the applicability to innovations



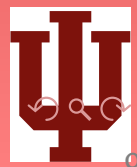


# Broadly-applicable annotation

Three aspects of this broad applicability to examine:

- **Before:** Heuristics which allow us to take a set of possible interpretations & land on a single annotation
  - The goal is to be robust to learner innovations & cases lacking evidence or with ambiguous evidence
- **During:** Applying the annotation to difficult usage shows the applicability to innovations
- **After:** Comparable occurrences of usage, found across different learner variables





# Broadly-applicable annotation

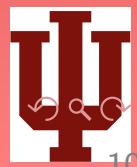
## Our heuristics to define the annotation

**Primary focus:** use surface evidence always at hand

- Make categories systematic & broadly applicable







# Broadly-applicable annotation

## Our heuristics to define the annotation

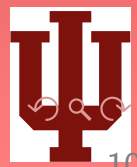
**Primary focus:** use surface evidence always at hand

- Make categories systematic & broadly applicable

(4) Although it not big and famous , but **it still has something <u>own</u> itself** .







# Broadly-applicable annotation

## Our heuristics to define the annotation

**Primary focus:** use surface evidence always at hand

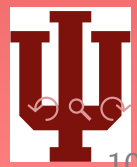
- Make categories systematic & broadly applicable

(4) Although it not big and famous , but **it still has something own itself** .

1. We give the learner the benefit of the doubt (if there is no way to disambiguate based on context).







# Broadly-applicable annotation

## Our heuristics to define the annotation

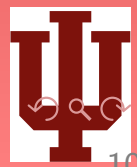
**Primary focus:** use surface evidence always at hand

- Make categories systematic & broadly applicable

(4) Although it not big and famous , but **it still has something own itself** .

1. We give the learner the benefit of the doubt (if there is no way to disambiguate based on context).
2. We try not to assume too much about the intended meaning of the learner, focusing on morpho-syntax.





# Broadly-applicable annotation

## Our heuristics to define the annotation

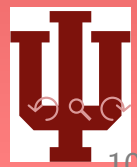
**Primary focus:** use surface evidence always at hand

- Make categories systematic & broadly applicable

(4) Although it not big and famous , but **it still has something own itself** .

1. We give the learner the benefit of the doubt (if there is no way to disambiguate based on context).
2. We try not to assume too much about the intended meaning of the learner, focusing on morpho-syntax.
3. As much as possible, we annotate the language "as is", i.e., in terms of linguistic evidence.





# Broadly-applicable annotation

## Our heuristics to define the annotation

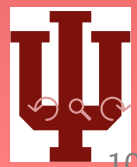
**Primary focus:** use surface evidence always at hand

- Make categories systematic & broadly applicable

(4) Although it not big and famous , but **it still has something own itself** .

1. We give the learner the benefit of the doubt (if there is no way to disambiguate based on context).
2. We try not to assume too much about the intended meaning of the learner, focusing on morpho-syntax.
3. As much as possible, we annotate the language "as is", i.e., in terms of linguistic evidence.
4. When the linguistic properties cannot be fully determined, we may underspecify the annotation.





# Broadly-applicable annotation

## Our heuristics to define the annotation

**Primary focus:** use surface evidence always at hand

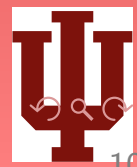
- Make categories systematic & broadly applicable

(4) Although it not big and famous , but **it still has something own itself** .

1. We give the learner the benefit of the doubt (if there is no way to disambiguate based on context).
2. We try not to assume too much about the intended meaning of the learner, focusing on morpho-syntax.
3. As much as possible, we annotate the language "as is", i.e., in terms of linguistic evidence.
4. When the linguistic properties cannot be fully determined, we may underspecify the annotation.
5. If nothing else works, we choose the more "primary" grammatical form (e.g., based on dictionary forms).







# Broadly-applicable annotation

## Our heuristics to define the annotation

**Primary focus:** use surface evidence always at hand

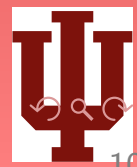
- Make categories systematic & broadly applicable

(4) Although it not big and famous , but **it still has something own itself** .

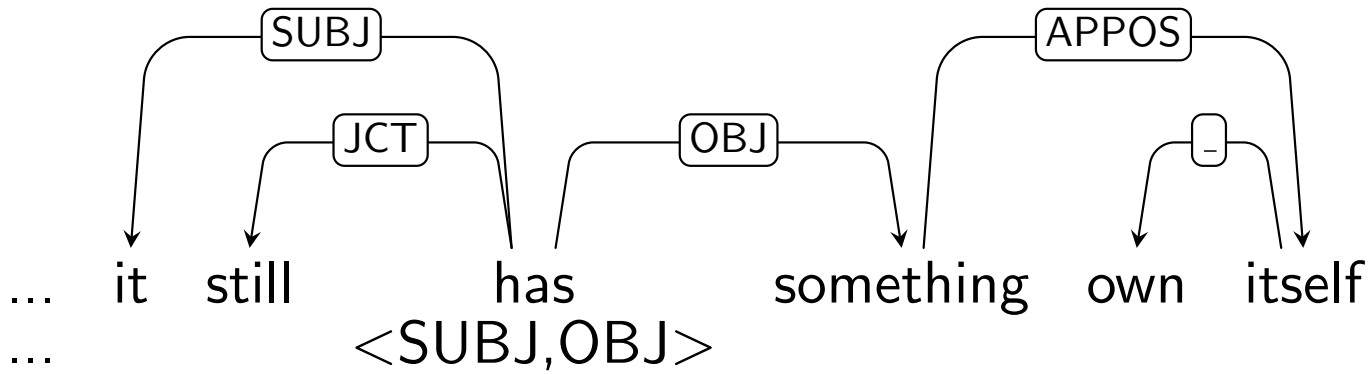
1. We give the learner the benefit of the doubt (if there is no way to disambiguate based on context).
2. We try not to assume too much about the intended meaning of the learner, focusing on morpho-syntax.
3. As much as possible, we annotate the language "as is", i.e., in terms of linguistic evidence.
4. When the linguistic properties cannot be fully determined, we may underspecify the annotation.
5. If nothing else works, we choose the more "primary" grammatical form (e.g., based on dictionary forms).

$\Rightarrow$ Principles ensure researchers can interpret categories



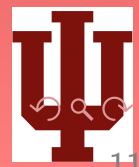


# Example annotation







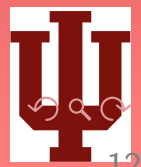
# Broadly-applicable annotation

## Applying the annotation

One of the most challenging aspects for annotators is our strict focus on syntax







# Broadly-applicable annotation

## Applying the annotation

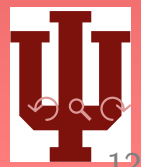
One of the most challenging aspects for annotators is our strict focus on syntax

(5) All these me felt better .

- In context, this probably means that *all these [things]* were things that made the writer feel better
  - But there is no *made* here!







# Broadly-applicable annotation

## Applying the annotation

One of the most challenging aspects for annotators is our strict focus on syntax

(5) All these me felt better .

- In context, this probably means that *all these [things]* were things that made the writer feel better
  - But there is no *made* here!

Regardless, there are some easy properties to annotate:

- *felt* is ROOT, with *better* as its predicate (PRED)
- *All* is QUANT dependent, of *these* or *me*



# Broadly-applicable annotation

## Applying the annotation

One of the most challenging aspects for annotators is our strict focus on syntax

(5) All these me felt better .

- In context, this probably means that *all these [things]* were things that made the writer feel better
  - But there is no *made* here!

Regardless, there are some easy properties to annotate:

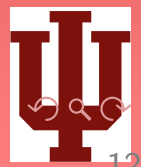
- *felt* is ROOT, with *better* as its predicate (PRED)
- *All* is QUANT dependent, of *these* or *me*

Our (morphologically-based) solution rests on:

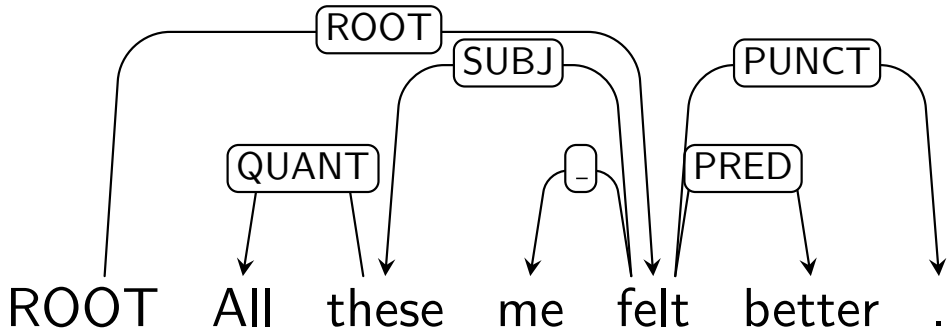
- *All these* making for a proper subject
- An allowance for underspecification







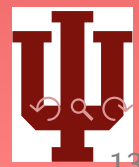
# Applying the annotation (2)



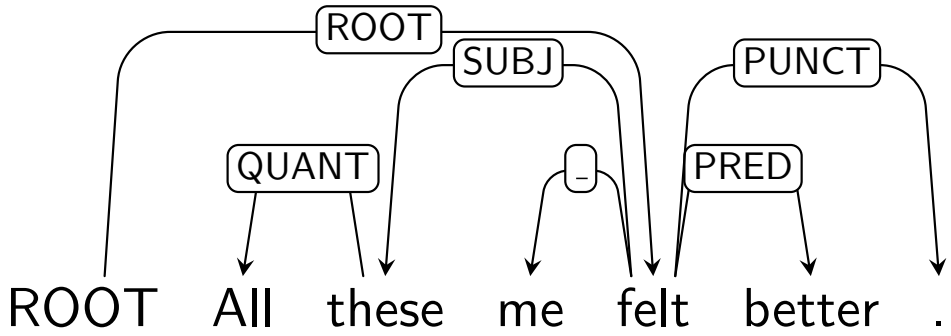

- *me* is some unspecified dependent of *felt*







# Applying the annotation (2)



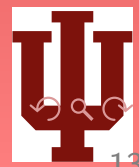

- *me* is some unspecified dependent of *felt*

What meaning does this correspond to?

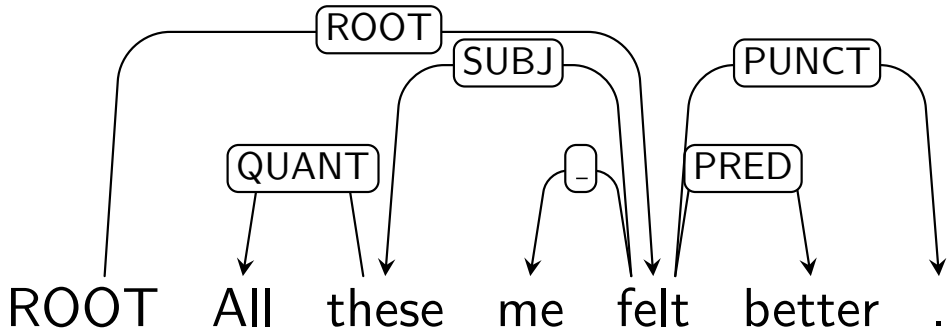
- We don't know precisely (future work?)







# Applying the annotation (2)




- *me* is some unspecified dependent of *felt*

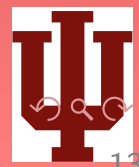
What meaning does this correspond to?

- We don't know precisely (future work?)

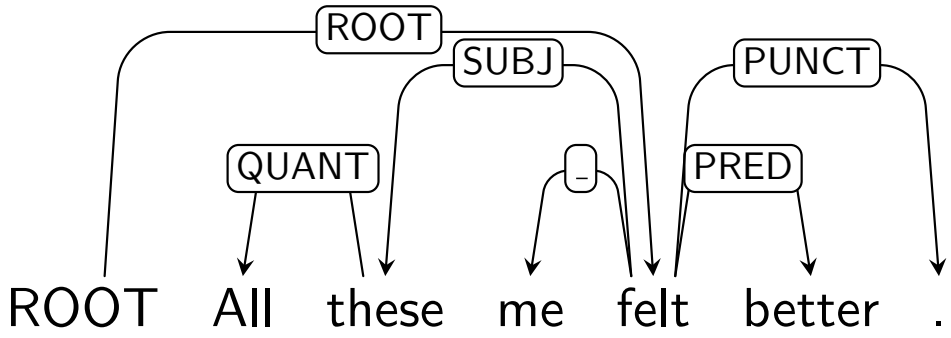
Does this correspond to the intended meaning?

- Probably not (need more meta-information to know?)





# Applying the annotation (2)




- *me* is some unspecified dependent of *felt*

What meaning does this correspond to?

- We don't know precisely (future work?)

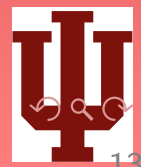
Does this correspond to the intended meaning?

- Probably not (need more meta-information to know?)

These analyses would be useful, but this analysis:

- Fits the evidence at hand & still provides access to useful categories





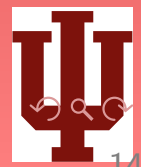
# Broadly-applicable annotation

## Comparable occurrences

e.g., learners using multiple arguments when one is required







# Broadly-applicable annotation

## Comparable occurrences

e.g., learners using multiple arguments when one is required

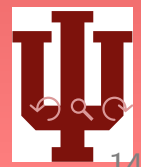
- Some cases are found by a search for TOP (topic)

(6) **Korea (level 3):** So , My **goals**$_{TOP}$ I catch **them**$_{OBJ}$

(7) **Thailand (4):** First , **atiquet**$_{TOP}$ of eating , **It**$_{SUBJ}$ 's very different from atiquet of eating in my country ...

(8) **Afghanistan (6):** So the **person**$_{TOP}$ who has gotten education by the government money and help , **he**$_{SUBJ}$ is owned to serve as hard as he can .





# Broadly-applicable annotation

## Comparable occurrences

e.g., learners using multiple arguments when one is required

- Some cases are found by a search for TOP (topic)

(6) **Korea (level 3):** So , My **goals**$_{TOP}$ I <u>catch</u> **them**$_{OBJ}$

(7) **Thailand (4):** First , **atiquet**$_{TOP}$ of eating , **It**$_{SUBJ}$ <u>'s</u> very different from atiquet of eating in my country ...

(8) **Afghanistan (6):** So the **person**$_{TOP}$ who has gotten education by the government money and help , **he**$_{SUBJ}$ <u>is</u> owned to serve as hard as he can .

$\Rightarrow$ What each TOP label means may differ, but we are able to group cases together across L1 & level



# Comparable occurrences (cont.)

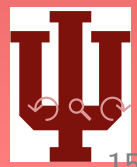
- Other example are found by finding multiple dependents of the same type when only one is required

(9) **Korea (3, same as previous):** I 'm not decide yet **which**$_{DET}$ **my**$_{DET}$ job .

(10) **Thailand (4, same):** It 's not suiable for **dress**$_{POBJ}$ **something**$_{POBJ}$ complecated .

(11) **Thailand (4, same):** They can **leave**$_{VC}$ **supapreate**$_{VC}$ with their family .





# Comparable occurrences (cont.)

- Other example are found by finding multiple dependents of the same type when only one is required

  (9) **Korea (3, same as previous):** I 'm not decide yet **which**$_{DET}$ **my**$_{DET}$ job .

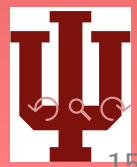
  (10) **Thailand (4, same):** It 's not suiable for **dress**$_{POBJ}$ **something**$_{POBJ}$ complecated .

  (11) **Thailand (4, same):** They can **leave**$_{VC}$ **supapreate**$_{VC}$ with their family .

$\Rightarrow$ Whether this corresponds to the same type of thing as with TOP is up to the researcher

- The annotation brings the sentences to light



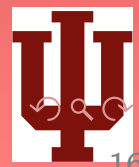


# Ways to utilize the annotation

We will consider different ways to utilize the annotation





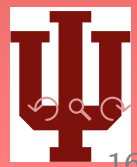


# Ways to utilize the annotation

We will consider different ways to utilize the annotation

- Simple gathering of statistics, e.g.:
  - 10 uses of the ELL (elliptical) label
  - 24 past tense copulas (VBD) taking a predicate (PRED)



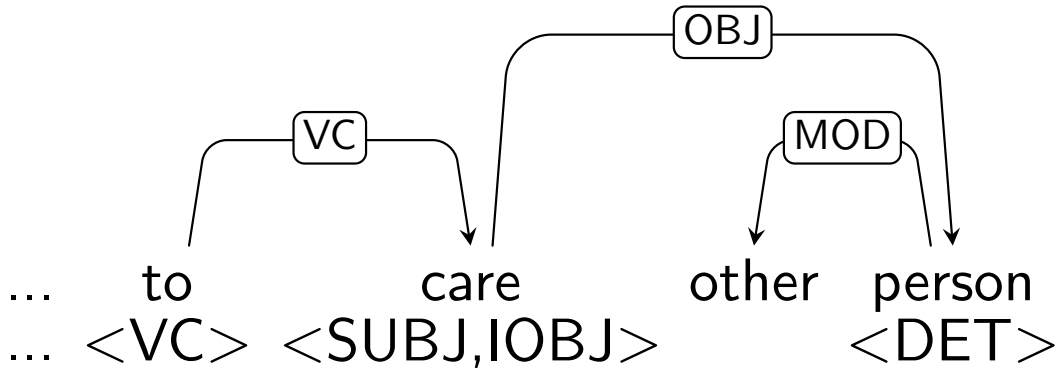


# Ways to utilize the annotation

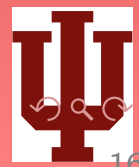
We will consider different ways to utilize the annotation

- Simple gathering of statistics, e.g.:
  - 10 uses of the ELL (elliptical) label
  - 24 past tense copulas (VBD) taking a predicate (PRED)
- Finding of mismatches:
  - POS mismatches (e.g., *have* **see**)
  - Subcategorization vs. realization mismatches

(12) I do n't have to care other person .





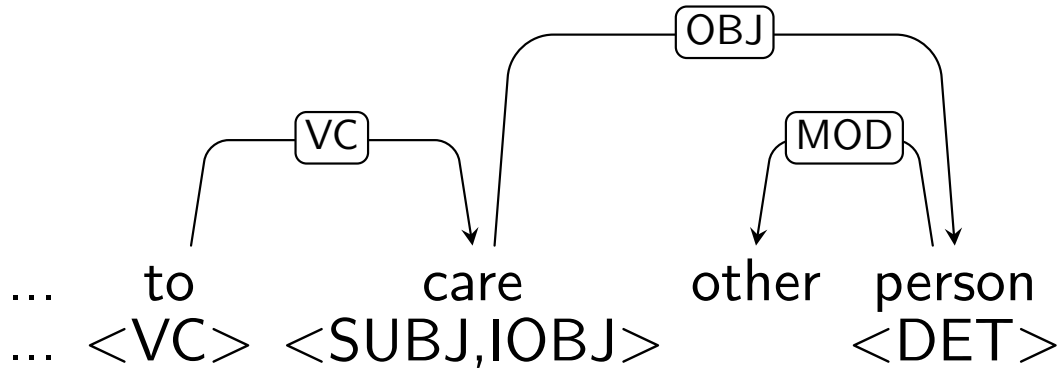



# Ways to utilize the annotation

We will consider different ways to utilize the annotation

- Simple gathering of statistics, e.g.:
  - 10 uses of the ELL (elliptical) label
  - 24 past tense copulas (VBD) taking a predicate (PRED)
- Finding of mismatches:
  - POS mismatches (e.g., *have* **see**)
  - Subcategorization vs. realization mismatches

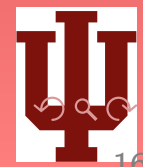
(12) I do n't have to care other person .




- Examining underspecified labels (_)

(13) ... I can n't **losing**_ these all things .



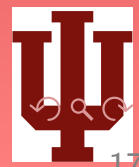


# Ways to utilize the annotation (2)

- Investigating rare/interesting labels: ELL, APPOS, ...

(14) I think if I was in my hometown , or **in**$_{ELL}$ better word , in my home country , I would n't be as hard working person , as now I am .





# Ways to utilize the annotation (2)

- Investigating rare/interesting labels: ELL, APPOS, ...

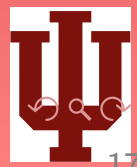
  (14) I think if I was in my hometown , or **in**$_{ELL}$ better word , in my home country , I would n't be as hard working person , as now I am .

- Quantitatively & qualitatively investigating correct and innovative uses of specific labels/constructions
  - e.g., We are currently investigating all the instances of determiner (non)use in the annotated part of the corpus

  (15) a. **Correct:** In addition, **the government** provides for **students food** and facilitative **dormitories**

  b. **Mismatch:** So the person who has gotten education by **the government money** and help ..



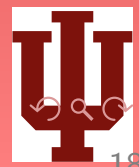


# Ways to utilize the annotation (3)

- Examining words with more than one syntactic head
  - e.g., right node raising:

(16) All countries in the world , especially United States of America helped and are helping our **coutry**$_{OBJ}$ to be in a better situation .



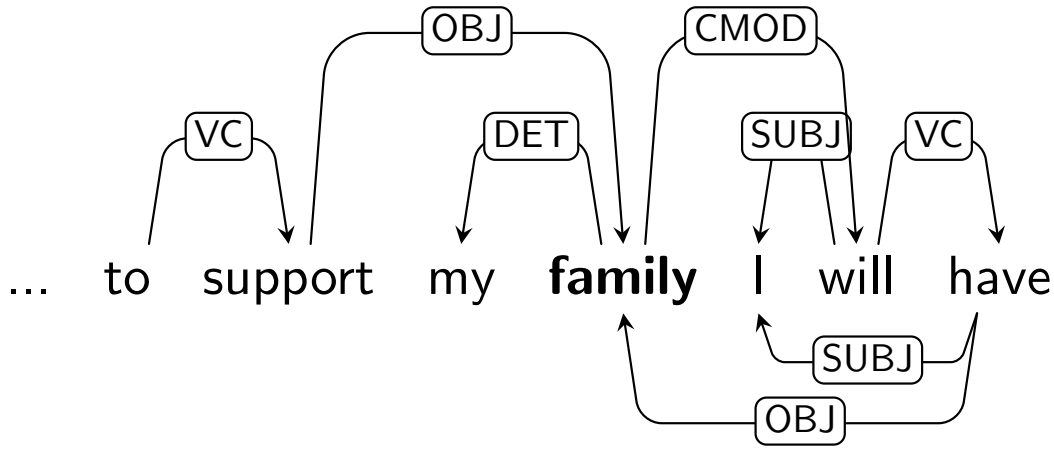


# Ways to utilize the annotation (3)

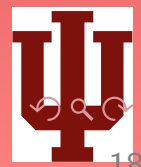
- Examining words with more than one syntactic head
  - e.g., right node raising:

    (16) All countries in the world , especially United States of America helped and are helping our **coutry**$_{OBJ}$ to be in a better situation .

  - e.g., relative clauses





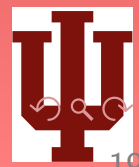



# Summary and Outlook

**Take-home point:** Corpus categories as well-documented indices provide data for researchers to interpret







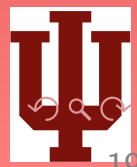
# Summary and Outlook

**Take-home point:** Corpus categories as well-documented indices provide data for researchers to interpret

We have discussed three points to our annotation scheme:

- Challenges in defining (morpho)syntactic categories for second language learner annotation
- Notes on how the annotation applies in different settings
- Opportunities in utilizing the annotation for research





# Summary and Outlook

**Take-home point:** Corpus categories as well-documented indices provide data for researchers to interpret

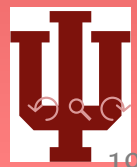
We have discussed three points to our annotation scheme:

- Challenges in defining (morpho)syntactic categories for second language learner annotation
- Notes on how the annotation applies in different settings
- Opportunities in utilizing the annotation for research

Some immediate next steps:

- Annotate more data, including parsing as a first step
- More thoroughly utilize the annotation
  - e.g., ongoing determiner study (Ragheb forthcoming)
- More deeply explore the connection between this annotation & error annotation
  - More thoroughly investigate how the annotation connects to semantic forms



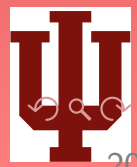


# Current status of the corpus/annotation

The annotation:

- Has been tested for inter-annotator agreement (Ragheb and Dickinson 2013)
- Has extensive guidelines already available online (Dickinson and Ragheb 2013)
  - These guidelines let you know what the categories mean





# Current status of the corpus/annotation

The annotation:

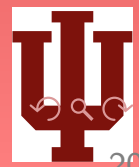
- Has been tested for inter-annotator agreement (Ragheb and Dickinson 2013)
- Has extensive guidelines already available online (Dickinson and Ragheb 2013)
  - These guidelines let you know what the categories mean

Target corpus release: within the next year (fingers crossed)

- Currently: 24 essays, 474 sentences, 7208 words ... with 5 main layers of annotation for each sentence
- This still needs more hand-checking

See: http://cl.indiana.edu/~salle/





# References


Ana Díaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2):139–154. Special Issue on New Trends in Language Teaching.

Markus Dickinson and Marwa Ragheb. 2013. Annotation for learner English guidelines, v. 0.1. Technical report, Indiana University, Bloomington, IN. June 9, 2013.

Sylviane Granger. 2003. Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, 20(3):465–480.

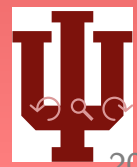
John A. Hawkins and Paula Buttery. 2010. Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1(1):1–23.

Hagen Hirschmann, Anke Lüdeling, Ines Rehbein, Marc Reznicek, and Amir Zeldes. 2010. Syntactic overuse and underuse: A study of a parsed learner corpus and its target hypothesis. Talk given at the Ninth Workshop on Treebanks and Linguistic Theory.

Marwa Ragheb. forthcoming. *Building a Syntactically-Annotated Corpus of Learner English*. Ph.D. thesis, Indiana University, Bloomington, IN.

Marwa Ragheb and Markus Dickinson. 2011. Avoiding the comparative fallacy in the annotation of learner corpora. In *Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions*, pages 114–124. Cascadilla Proceedings Project, Somerville, MA.



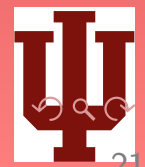



Marwa Ragheb and Markus Dickinson. 2012. Defining syntax for learner language annotation. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012), Poster Session*. Mumbai, India.

Marwa Ragheb and Markus Dickinson. 2013. Inter-annotator agreement for dependency annotation of learner language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, GA.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. 2010. Using parse features for preposition selection and error detection. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 353–358. Uppsala, Sweden.





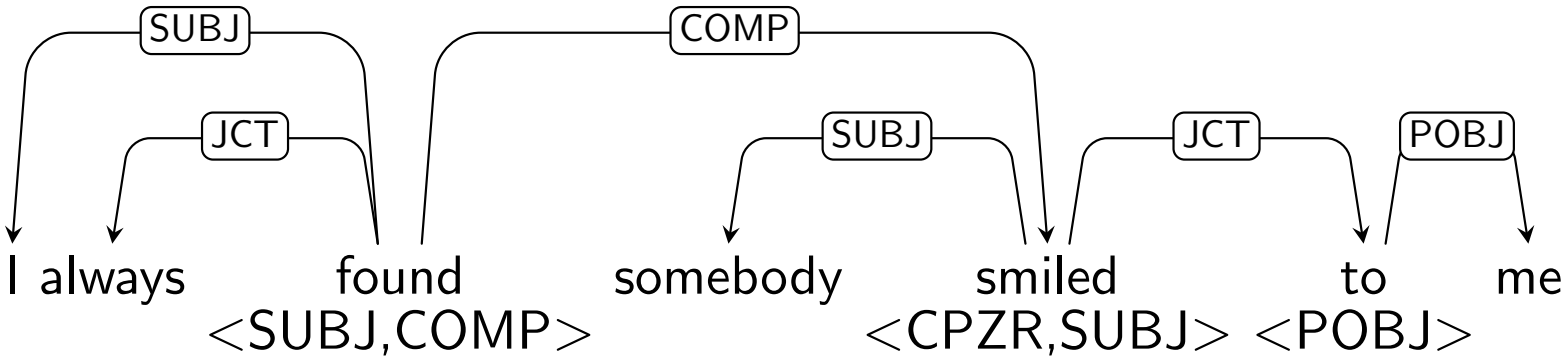
# Broadly-applicable annotation

## Applying the annotation

Consider this complement clause:

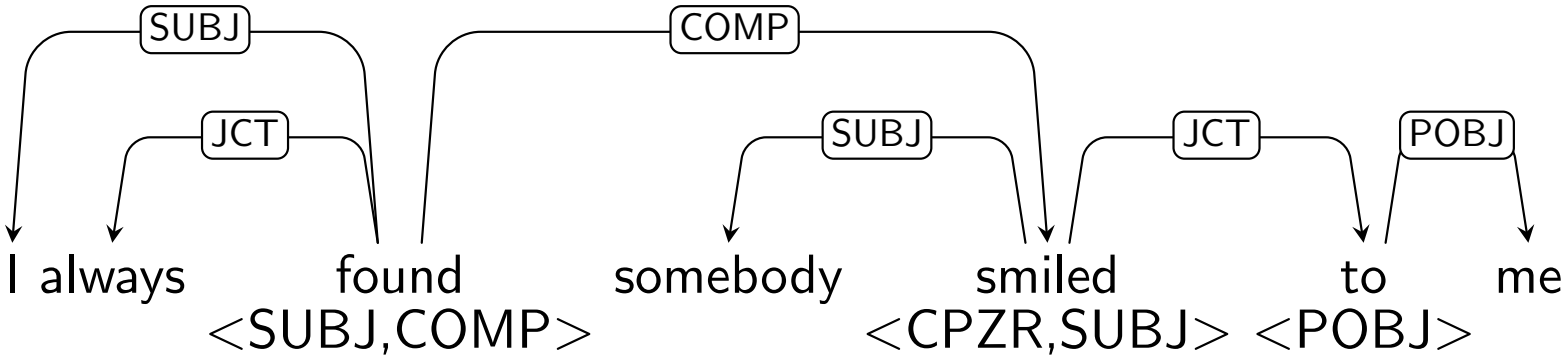
(17) When I walked on the road , **I always found somebody smiled to me** .

- Even though it is odd, most of the annotation decisions are relatively easy
  - e.g., *to* is treated like a normal preposition



# Applying the annotation

## How is the sentence ungrammatical?

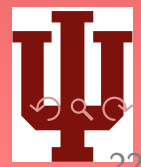



- *smiled* is a finite verb heading a clause, as a complement (COMP) of *found*, cf.

  (18) I found that somebody was smiling at me.

- The only mismatch is the complementizer (CPZR) on the subcategorization list of *smiled*
  - We do not need to know why the absence of *that* makes a difference here, only annotate as such



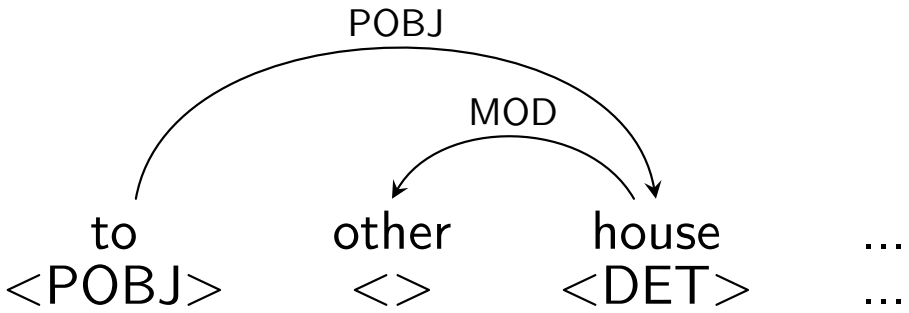


# Defining subcategorization

Subcategorization encodes general constraints

- It can help capture discrepancies in argument structure

(19) . . . we moved again to other **house** . . .

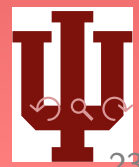



- Annotate every possible subcategorization or only one?
  - One annotation makes it context-specific, overlapping with distributional evidence

  (20) One [goal] is to contribute to both global and local **community** ...

  - <DET> fits discourse; <> is a general possibility





# Defining morphosyntax

Morphosyntax based on the visible forms of words

- e.g., analysis should be appropriate for 3 sg. verb:

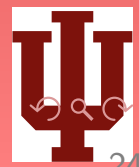
  (21) I had a problem a bout **chooses** my car.

- We use distribution to disambiguate:

  (22) And also I can heart the step of my **walked** .

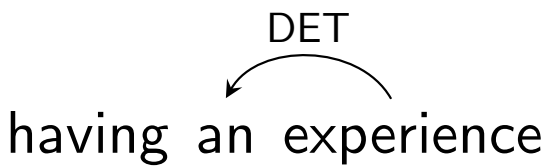
  - 3 POS for *can*, 2 for *step*, 2 for *walked*: 3x2x2=12 trees
  - What would be the tree for past tense of *walked*?
    - Adjectival/past participle fits better to the context





# Defining distribution

- Syntactic distributional slot: position where token with particular properties (e.g., singular noun) predicted to occur, on the (syntactic) basis of surrounding tokens
- Does morphology play a role in disambiguation?

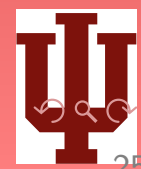



  - MOD & QUANT are possible here, but *an* = DET

**Overall point:** The layers are inter-defined

- Morphosyntax + subcategorization covers nearly all distributional dependency distinctions
- We thus do not to annotate distributional dependencies





# Interannotator agreement

## Initial layers

Overview of agreement rates before & after discussion (phases 2 & 4):

| Annotators | lemma | | $POS_m$ | | $POS_d$ | |
|---|---|---|---|---|---|---|
| | P2 | P4 | P2 | P4 | P2 | P4 |
| A, B | 93.4 | 96.9 | 99.0 | 98.7 | 99.2 | 98.7 |
| B, C | 94.4 | 97.7 | 99.0 | 99.5 | 98.7 | 99.3 |
| C, A | 92.4 | 96.9 | 99.7 | 99.7 | 98.5 | 99.3 |

- $POS_m$ & $POS_d$: High agreement rates reflect annotators making very few changes to automatic pre-annotation
- Lemmas: Improvement in agreement—although it could be higher, given the simplicity of lemma information





# Interannotator agreement

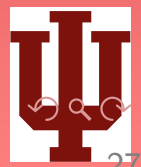
## Dependencies

Overview of agreement rates before & after discussion (phases 2 & 4):

| Annotators | Subcat. | | UAA | | LAA | |
|---|---|---|---|---|---|---|
| | P2 | P4 | P2 | P4 | P2 | P4 |
| A, B | 85.5 | 94.0 | 86.6 | 97.0 | 80.0 | 95.2 |
| B, C | 86.1 | 95.7 | 86.7 | 97.1 | 80.3 | 96.0 |
| C, A | 86.1 | 96.6 | 86.9 | 97.7 | 82.4 | 96.7 |

- Initial agreement rates around 80–85%: moderately high
- Agreement rates improved after discussion, achieving approximately 95% agreement





# Interannotator agreement

## Dependencies (cont.)

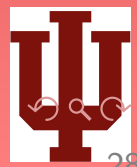
| Ann. | UAA | | LAA | | LOA | |
|---|---|---|---|---|---|---|
| | P2 | P4 | P2 | P4 | P2 | P4 |
| A, B | 81.8 | 96.1 | 73.6 | 93.4 | 80.3 | 95.5 |
| B, C | 80.9 | 96.2 | 73.4 | 94.4 | 79.3 | 97.1 |
| A,C | 83.6 | 97.6 | 79.7 | 96.7 | 81.8 | 97.9 |

Table: MASI percentages for dependencies, Text 1

| Ann. | UAA | | LAA | | LOA | |
|---|---|---|---|---|---|---|
| | P2 | P4 | P2 | P4 | P2 | P4 |
| A, B | 92.6 | 98.1 | 87.8 | 97.4 | 89.3 | 97.8 |
| B, C | 93.8 | 98.3 | 88.7 | 97.9 | 90.2 | 98.6 |
| A, C | 90.9 | 97.9 | 85.7 | 96.8 | 87.6 | 97.9 |

Table: MASI percentages for dependencies, Text 2





# Interannotator agreement

## Subcategorization

| Ann. | MASI | | $GCM_1$ | | $GCM_2$ | |
|---|---|---|---|---|---|---|
| | P2 | P4 | P2 | P4 | P2 | P4 |
| A,B | 84.3 | 92.4 | 81.9 | 90.8 | 72.8 | 88.1 |
| B,C | 83.6 | 93.8 | 74.4 | 91.6 | 73.6 | 90.2 |
| A,C | 84.9 | 96.1 | 83.0 | 96.4 | 73.1 | 92.2 |

Table: Agreement rates for subcategorization, Text 1

| Ann. | MASI | | $GCM_1$ | | $GCM_2$ | |
|---|---|---|---|---|---|---|
| | P2 | P4 | P2 | P4 | P2 | P4 |
| A,B | 87.1 | 95.9 | 88.9 | 96.0 | 77.2 | 94.1 |
| B,C | 89.3 | 98.0 | 88.3 | 98.0 | 82.0 | 96.8 |
| A,C | 87.6 | 97.2 | 91.2 | 97.3 | 73.7 | 94.2 |

Table: Agreement rates for subcategorization, Text 2