Avoiding the Comparative Fallacy in the Annotation of Learner Corpora

Marwa Ragheb & Markus Dickinson

Dept. of Linguistics, Indiana University
SLRF 2010; College Park, MD; October 16, 2010
Introduction & Motivation
Searching for relevant linguistic properties

For many questions in second language research, one can search a corpus for specific words to find relevant examples

▶ e.g., How are modal verbs used by L2 learners? (cf., e.g., Aijmer 2002)
Introduction & Motivation
Searching for relevant linguistic properties

For many questions in second language research, one can search a corpus for specific words to find relevant examples

▶ e.g., How are modal verbs used by L2 learners? (cf., e.g., Aijmer 2002)

But consider a search for syntactic patterns, such as examining *wh*-movement (e.g., Juffs 2005; Schachter 1989)

▶ What kind of search involving specific words addresses questions about the function of *whom* in this sentence?

(1) I want to be a person *whom* my wife and children would be proud of
Introduction & Motivation
Searching for relevant linguistic properties

For many questions in second language research, one can search a corpus for specific words to find relevant examples

▶ e.g., How are modal verbs used by L2 learners? (cf., e.g., Aijmer 2002)

But consider a search for syntactic patterns, such as examining *wh*-movement (e.g., Juffs 2005; Schachter 1989)

▶ What kind of search involving specific words addresses questions about the function of *whom* in this sentence?

(1) I want to be a person *whom* my wife and children would be proud of

▶ How do we know this is subject or object extraction? How do we know the depth of embedding? . . .
Corpora for SLA

Corpora containing data of second language learners provide data for investigating SLA questions

- But how does one search for abstract properties?
Corpora for SLA

Corpora containing data of second language learners provide data for investigating SLA questions

- But how does one search for abstract properties?
  - different realizations of negation (e.g., Tomaselli and Schwartz 1990)
  - definiteness or indefiniteness (or lack thereof) (cf., e.g., Ionin et al. 2004)
  - (headless) relative clauses (e.g., O’Grady et al. 2003)
  - . . .
Corpora for SLA

Corpora containing data of second language learners provide data for investigating SLA questions

- But how does one search for abstract properties?
  - different realizations of negation (e.g., Tomaselli and Schwartz 1990)
  - definiteness or indefiniteness (or lack thereof) (cf., e.g., Ionin et al. 2004)
  - (headless) relative clauses (e.g., O’Grady et al. 2003)
  - ... 

- Currently, these must be searched for by hand
Corpora for SLA

Corpora containing data of second language learners provide data for investigating SLA questions

- But how does one search for abstract properties?
  - different realizations of negation (e.g., Tomaselli and Schwartz 1990)
  - definiteness or indefiniteness (or lack thereof) (cf., e.g., Ionin et al. 2004)
  - (headless) relative clauses (e.g., O’Grady et al. 2003)
  - ...

- Currently, these must be searched for by hand

To investigate such issues, we need the data marked up with grammatical properties (see, e.g., Meurers and Müller 2009)

- Otherwise: relevant instances won’t be found & many non-relevant instances will have to be sorted through
Annotation of Learner Corpora

By providing *annotation* of relevant learner properties, we can provide for better investigation of SLA issues.
Annotation of Learner Corpora

By providing *annotation* of relevant learner properties, we can provide for better investigation of SLA issues.

But what should linguistic, or grammatical, annotation of learner language look like?
Annotation of Learner Corpora

By providing *annotation* of relevant learner properties, we can provide for better investigation of SLA issues.

But what should linguistic, or grammatical, annotation of learner language look like?

▶ *How* do we define annotation which supports the investigation of learner language?

▶ This is a useful question in its own right:
  ▶ We are forced to precise about the linguistic properties of learner language.
Annotation of Learner Corpora

By providing annotation of relevant learner properties, we can provide for better investigation of SLA issues

But what should linguistic, or grammatical, annotation of learner language look like?

- How do we define annotation which supports the investigation of learner language?
- This is a useful question in its own right:
  - We are forced to precise about the linguistic properties of learner language

Goal: Work on defining an annotation scheme appropriate for learner language

- There is only little research on this topic (Díaz-Negrillo et al. 2010; Dickinson and Ragheb 2009)
Outline of talk:

▶ Existing annotation schemes
  ▶ Annotation for learner corpora
  ▶ Linguistic annotation for other corpora
▶ The comparative fallacy
▶ Annotating learner language
  ▶ Annotate all words
  ▶ Rely on linguistic evidence
  ▶ Describe separate properties of the language (multi-layered annotation)
  ⇒ Avoid the comparative fallacy
▶ Summary & Outlook
Existing Annotation of Learner Corpora

For corpora of language of second language learners

- The most common form of annotation focuses on errors
  - Suri and McCoy (1993); Lüdeling et al. (2005); Boyd (2010); Rozovskaya and Roth (2010), . . .
**Existing Annotation of Learner Corpora**

For corpora of language of second language learners

- The most common form of annotation focuses on errors
  - Suri and McCoy (1993); Lüdeling et al. (2005); Boyd (2010); Rozovskaya and Roth (2010), ...  
  - Example of annotation from Granger (2003):
    
    (2) Ces gens <G><NBR><VSC> #pensent$  
    
    pense </VSC></NBR></G> aussi que ...

- Error tags:
  - <G> = Grammar error (Error domain)
  - <NBR> = Number (Error category)
Existing Annotation of Learner Corpora

For corpora of language of second language learners

▶ The most common form of annotation focuses on errors
  ▶ Suri and McCoy (1993); Lüdeling et al. (2005); Boyd (2010); Rozovskaya and Roth (2010), . . .
  ▶ Example of annotation from Granger (2003):

\[
(2) \text{Ces gens } \langle G \rangle \langle NBR \rangle \langle VSC \rangle \# \text{pensent$}
\text{pense } \langle /VSC \rangle \langle /NBR \rangle \langle /G \rangle \text{ aussi que . . .}
\]

▶ Error tags:
  ▶ \(\langle G \rangle\) = Grammar error (Error domain)
  ▶ \(\langle NBR \rangle\) = Number (Error category)

▶ Annotation of part of speech (POS) for errors:
  ▶ \(\langle VSC \rangle\) = Finite simple verb (Word category)
Existing Annotation of Learner Corpora

For corpora of language of second language learners

- The most common form of annotation focuses on errors
  - Suri and McCoy (1993); Lüdeling et al. (2005); Boyd (2010); Rozovskaya and Roth (2010), ...
  - Example of annotation from Granger (2003):

    (2) Ces gens <G><NBR><VSC> #pensent$ pense </VSC></NBR></G> aussi que ...

- Error tags:
  - <G> = Grammar error (Error domain)
  - <NBR> = Number (Error category)
- Annotation of part of speech (POS) for errors:
  - <VSC> = Finite simple verb (Word category)
- Target form: #pensent$
Error annotation

Consider word category, or part-of-speech (POS), annotation in the previous example (VSC):
Error annotation

Consider word category, or part-of-speech (POS), annotation in the previous example (VSC):

- Only defined for erroneous words
Error annotation

Consider word category, or part-of-speech (POS), annotation in the previous example (VSC):

▶ Only defined for erroneous words
▶ Not clear what it should be for novel POS uses, e.g., from Díaz-Negrillo et al. (2010):

(3) ... television, radio are very **subjectives** ...
Error annotation

Consider word category, or part-of-speech (POS), annotation in the previous example (VSC):

- Only defined for erroneous words
- Not clear what it should be for novel POS uses, e.g., from Díaz-Negrillo et al. (2010):

  (3) ... television, radio are very **subjectives** ...

Error annotation does not allow for searching of linguistic properties, e.g., finding different types of question formation

- Annotating target forms often encodes some notion of distance from the L2
Error annotation

Consider word category, or part-of-speech (POS), annotation in the previous example (VSC):

- Only defined for erroneous words
- Not clear what it should be for novel POS uses, e.g., from Díaz-Negrillo et al. (2010):
  
  (3) ... television, radio are very **subjectives** ...

Error annotation does not allow for searching of linguistic properties, e.g., finding different types of question formation

- Annotating target forms often encodes some notion of distance from the L2

Need to annotate linguistic properties of learner language

- i.e., annotate observable surface properties, not ones based on forms & meanings learner intended to use
A range of corpus annotation

Outside of learner language, this issue of annotating linguistic properties is not a new one . . .

Linguistic annotation can contain information about:

- lemmata, morphology, & part-of-speech (POS) (e.g., Leech 1997; Sampson 1995; Schiller et al. 1995)
- syntactic constituencies & dependencies (e.g., Marcus et al. 1993; Hajič 1998; Skut et al. 1997)
- semantic roles & word senses (e.g., Kingsbury et al. 2002; Hajičová 1998; Erk et al. 2003)
- discourse properties (e.g., Allen and Core 1996)
A range of corpus annotation

Outside of learner language, this issue of annotating linguistic properties is not a new one . . .

Linguistic annotation can contain information about:

- lemmata, morphology, & part-of-speech (POS) (e.g., Leech 1997; Sampson 1995; Schiller et al. 1995)
- syntactic constituencies & dependencies (e.g., Marcus et al. 1993; Hajič 1998; Skut et al. 1997)
- semantic roles & word senses (e.g., Kingsbury et al. 2002; Hajičová 1998; Erk et al. 2003)
- discourse properties (e.g., Allen and Core 1996)

Question: How can we apply these types of annotation to learner language?
Linguistic properties

When we talk about linguistic properties, we are talking about morphosyntactic & syntactic annotation

- e.g., Here are two grammatical relations (dependencies) between words (among others)

![Diagram showing grammatical relations between words]

Tin Toy can makes different music sound
Where we’re coming from

We have begun annotating data from learners of varying levels (Dickinson and Ragheb 2009)
Where we’re coming from

We have begun annotating data from learners of varying levels (Dickinson and Ragheb 2009)

- Narratives collected from the 90s (Bardovi-Harlig 1999)
  - Learners watched a short cartoon (*Tin Toy*) and were asked to discuss what happened
- Essays from the Intensive English Program (IEP) at Indiana University, used for course placement
  - Students respond to a prompt such as “What are your plans for life?”
Where we’re coming from

We have begun annotating data from learners of varying levels (Dickinson and Ragheb 2009)

- Narratives collected from the 90s (Bardovi-Harlig 1999)
  - Learners watched a short cartoon (Tin Toy) and were asked to discuss what happened

- Essays from the Intensive English Program (IEP) at Indiana University, used for course placement
  - Students respond to a prompt such as “What are your plans for life?”

We have tried to apply linguistic annotation & noticed issues arising related to the comparative fallacy
Avoiding the Comparative Fallacy in the Annotation of Learner Corpora

**Comparative fallacy**: ‘mistake of studying the systematic character of one language by comparing it to another’ (Bley-Vroman 1983)
Avoiding the Comparative Fallacy in the Annotation of Learner Corpora

**Comparative fallacy**: ‘mistake of studying the systematic character of one language by comparing it to another’ (Bley-Vroman 1983)

- Language system constructed by a second language (L2) learner is not a ‘degenerate form’ of target language
  - Interlanguage is a system in itself that should be studied
Annotation & the Comparative Fallacy

- **Comparative fallacy**: ‘mistake of studying the systematic character of one language by comparing it to another’ (Bley-Vroman 1983)
- Language system constructed by a second language (L2) learner is not a ‘degenerate form’ of target language
  - Interlanguage is a system in itself that should be studied
- Lakshmanan and Selinker (2001) extend this notion:
  - Comparing with native language (L1) could obscure systematicity in interlanguage
Annotation & the Comparative Fallacy

- **Comparative fallacy**: ‘mistake of studying the systematic character of one language by comparing it to another’ (Bley-Vroman 1983)
- Language system constructed by a second language (L2) learner is not a ‘degenerate form’ of target language
  - Interlanguage is a system in itself that should be studied
- Lakshmanan and Selinker (2001) extend this notion:
  - Comparing with native language (L1) could obscure systematicity in interlanguage

Error annotation is inherently prone to comparative fallacy:
- Error interpretation makes learner language seem like a degenerate L2
We advocate the following for annotating learner language:

1. Encode linguistic properties for every word, not just ‘errors’.
2. Use linguistic evidence when assigning linguistic properties.
3. Describe the data as it appears, by separating linguistic properties into multiple layers.

With such principles, annotation efforts will be less likely to fall into the comparative fallacy.
Learner Language Annotation

We advocate the following for annotating learner language:

1. Encode linguistic properties for every word, not just ‘errors’
Learner Language Annotation

We advocate the following for annotating learner language:

1. Encode linguistic properties for every word, not just ‘errors’
2. Use linguistic evidence when assigning linguistic properties
Learner Language Annotation

We advocate the following for annotating learner language:

1. Encode linguistic properties for every word, not just ‘errors’
2. Use linguistic evidence when assigning linguistic properties
3. Describe the data as it appears, by separating linguistic properties into multiple layers
Learner Language Annotation

We advocate the following for annotating learner language:

1. Encode linguistic properties for every word, not just ‘errors’
2. Use linguistic evidence when assigning linguistic properties
3. Describe the data as it appears, by separating linguistic properties into multiple layers

With such principles, annotation efforts will be less likely to fall into the comparative fallacy

- We emphasize annotating observable forms
Encode linguistic properties for every word

Example of what we’re aiming for (Dickinson and Ragheb 2009):

MOD SUBJ AUX ROOT MOD MOD OBJ

Tin Toy can makes different music sound

NP1x NP1x VMo VVZt JJ NN1u NN1c
NP1x NP1x VMo VV0t JJ JJ NN
<><><><SUBJ,AUX, OBJ><><><>
Use linguistic evidence

Rely on *linguistic* evidence to annotate the data

▶ Instead of relying on intention, knowledge of L1, SLA theory, etc.
Use linguistic evidence

Rely on *linguistic* evidence to annotate the data

- Instead of relying on intention, knowledge of L1, SLA theory, etc.

But this is non-trivial
Use linguistic evidence

Rely on linguistic evidence to annotate the data
  ▶ Instead of relying on intention, knowledge of L1, SLA theory, etc.

But this is non-trivial
  ▶ Consider part-of-speech (POS) tags, where POS is defined by both morphological & distributional criteria (e.g., Sampson 1995)
    ▶ A learner may have evidence pointing different ways:

   (4) Tin Toy can makes different music sound.
Use linguistic evidence

Rely on *linguistic* evidence to annotate the data

- Instead of relying on intention, knowledge of L1, SLA theory, etc.

But this is non-trivial

- Consider part-of-speech (POS) tags, where POS is defined by both morphological & distributional criteria (e.g., Sampson 1995)
  - A learner may have evidence pointing different ways:

  (4) Tin Toy can *makes* different music sound.

  - Morphological evidence: 3rd person present tense verb
  - Distributional evidence: base form verb
Multi-layered description

Relying on what’s observable, i.e., on evidence, leads to multi-layered annotation

- Encode separate layers for separate pieces of evidence (cf. also Díaz-Negrillo et al. 2010)
Multi-layered description

Relying on what’s observable, i.e., on evidence, leads to multi-layered annotation

- Encode separate layers for separate pieces of evidence (cf. also Díaz-Negrillo et al. 2010)

For *can makes, makes* can be annotated as:

- Morphological layer: 3rd person present tense verb
- Distributional layer: base form verb
Multi-layered description

Relying on what’s observable, i.e., on evidence, leads to multi-layered annotation

- Encode separate layers for separate pieces of evidence (cf. also Díaz-Negrillo et al. 2010)

For *can makes*, *makes* can be annotated as:

- Morphological layer: 3rd person present tense verb
- Distributional layer: base form verb

This means that each layer can contain a description of a linguistic property
Putting it all together
Different layers for POS annotation

We are using POS tags from SUSANNE tagset (Sampson 1995)
Putting it all together

Different layers for dependency annotation

Grammatical Relations encoded via surface dependencies & subcategorization frames:

We use the dependency annotation scheme developed for CHILDES data (Sagae et al. 2007, 2004)
But wait: We are still defining each layer of annotation in terms of L2 properties

► The morphology of “3rd person singular present tense” is defined by the presence of -s on a verb
► The distribution of “base form verb” is defined by appearing directly after an auxiliary verb
But wait: We are still defining each layer of annotation in terms of L2 properties

- The morphology of “3rd person singular present tense” is defined by the presence of -s on a verb
- The distribution of “base form verb” is defined by appearing directly after an auxiliary verb

These properties are defined by virtue of how they work in the L2 (English)
L2 reference frame

But wait: We are still defining each layer of annotation in terms of L2 properties

- The morphology of “3rd person singular present tense” is defined by the presence of -s on a verb
- The distribution of “base form verb” is defined by appearing directly after an auxiliary verb

These properties are defined by virtue of how they work in the L2 (English)

How, then, are we avoiding the comparative fallacy?
Avoiding the comparative fallacy

For the example of *makes*

- What we say:
  - Morphologically: 3rd singular present tense
  - Distributionally: base form verb slot
Avoiding the comparative fallacy

For the example of *makes*

- **What we say:**
  - Morphologically: 3rd singular present tense
  - Distributionally: base form verb slot
- **What we don’t say:** learner is using (or intending to use) this as 3rd singular or base form verb
Avoiding the comparative fallacy

For the example of *makes*

- What we say:
  - Morphologically: 3rd singular present tense
  - Distributionally: base form verb slot
- What we don’t say: learner is using (or intending to use) this as 3rd singular or base form verb
  - Multi-layered annotation allows us not to make a definitive claim about what the single annotation is
  - Annotation in no way indicates that the sentence is a degenerate L2 form or should be any other form
Avoiding the comparative fallacy

For the example of *makes*

- What we say:
  - Morphologically: 3rd singular present tense
  - Distributionally: base form verb slot
- What we don’t say: learner is using (or intending to use) this as 3rd singular or base form verb
  - Multi-layered annotation allows us not to make a definitive claim about what the single annotation is
  - Annotation in no way indicates that the sentence is a degenerate L2 form or should be any other form

We only use categories which are closely tied to the data
- i.e., we avoid combining evidence from different descriptive classes (e.g., a single POS tag)
Usefulness of the annotation

The encoded information will make it easier to search for specific linguistic properties in the learner corpus

- We can now talk about things beyond the words, i.e., linguistic classes of surface forms
Usefulness of the annotation

The encoded information will make it easier to search for specific linguistic properties in the learner corpus

- We can now talk about things beyond the words, i.e., linguistic classes of surface forms

Consider *wh*-words again:
Usefulness of the annotation

The encoded information will make it easier to search for specific linguistic properties in the learner corpus

- We can now talk about things beyond the words, i.e., linguistic classes of surface forms

Consider *wh*-words again:

... whom my wife is proud of
...

... <> <> <DET> <SUBJ,PRED> <JCT> <POBJ>
Usefulness of the annotation (2)

The annotation allows us to determine the depth of embedding:

... whom I think my wife is proud of ...
Summary and Outlook

We have:

▶ discussed designing an annotation scheme for learner language, in a way which avoids the comparative fallacy

Next steps:

▶ Development and refinement of the annotation scheme
▶ Collect and annotate learner data that will eventually be made publicly available
▶ This annotation does not answer SLA questions, but it provides a platform for others to answer such questions
Summary and Outlook

We have:

- discussed designing an annotation scheme for learner language, in a way which avoids the comparative fallacy
  - Annotate all words
  - Use linguistic evidence
  - Describe different layers of annotation

Summary & Outlook

Next steps:
- Development and refinement of the annotation scheme
- Collect and annotate learner data that will eventually be made publicly available
- This annotation does not answer SLA questions, but it provides a platform for others to answer such questions
Summary and Outlook

We have:

▶ discussed designing an annotation scheme for learner language, in a way which avoids the comparative fallacy
  ▶ Annotate all words
  ▶ Use linguistic evidence
  ▶ Describe different layers of annotation
▶ Using such annotation will allow for better searching for interlanguage properties
Summary and Outlook

We have:

- discussed designing an annotation scheme for learner language, in a way which avoids the comparative fallacy
  - Annotate all words
  - Use linguistic evidence
  - Describe different layers of annotation
- Using such annotation will allow for better searching for interlanguage properties

Next steps:

- Development and refinement of the annotation scheme
- Collect and annotate learner data that will eventually be made publicly available
  - This annotation does not answer SLA questions, but it provides a platform for others to answer such questions
Acknowledgements

Thanks to the following for feedback:

- Kathleen Bardovi-Harlig
- Detmar Meurers
- Rex Sprouse
- David Stringer
- Holger Wunsch
- The CL discussion group & SLS colloquium series at IU


Avoiding the Comparative Fallacy in the Annotation of Learner Corpora

Introduction

Existing annotation

Comparative fallacy

Learner language annotation

All words

Evidence-based multi-layered description

Avoiding the CF

Summary & Outlook

References

for Computational Linguistics (ACL-03), pages 537–544, Sapporo, Japan. http://acl.ldc.upenn.edu/P/P03/P03-1068.pdf.


Avoiding the Comparative Fallacy in the Annotation of Learner Corpora

Introduction

Existing annotation

Comparative fallacy

Learner language annotation

All words

Evidence-based Multi-layered description

Avoiding the CF

Summary & Outlook

References


