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)

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?
  - 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

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

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), . . .
  - 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):

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

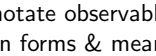
(3) . . . television, radio are very **subjectives** . . .

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:

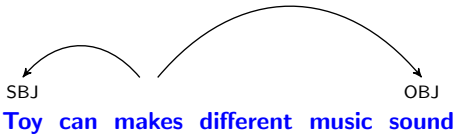
- lemmata, morphology, & part-of-speech (POS) (e.g., Leech 1997; Sampson 1995; Schiller et al. 1995)
- syntactic constituencies & dependencies (e.g., Marcus et al. 1993; Hajič 1998; Skut et al. 1997)
- semantic roles & word senses (e.g., Kingsbury et al. 2002; Hajičová 1998; Erk et al. 2003)
- discourse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

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)



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

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

Learner Language Annotation

We advocate the following for annotating learner language:

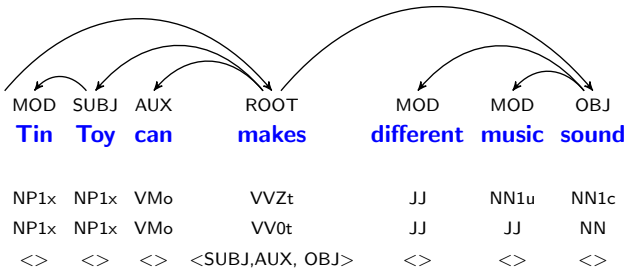
1. Encode linguistic properties for every word, not just 'errors'
2. Use linguistic evidence when assigning linguistic properties
3. Describe the data as it appears, by separating linguistic properties into multiple layers

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):



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)

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

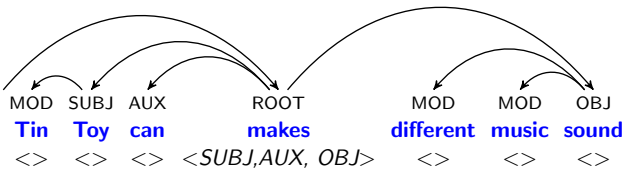
Tin	Toy	can	makes	different	music	sound
NP1x	NP1x	VMo	VVZt	JJ	NN1u	NN1c
NP1x	NP1x	VMo	VV0t	JJ	JJ	NN

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)

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
- 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

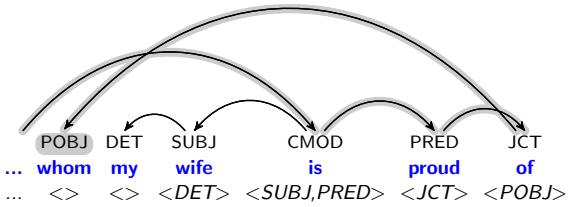
- i.e., we avoid combining evidence from different descriptive classes (e.g., a single POS tag)

Usefulness of the annotation

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

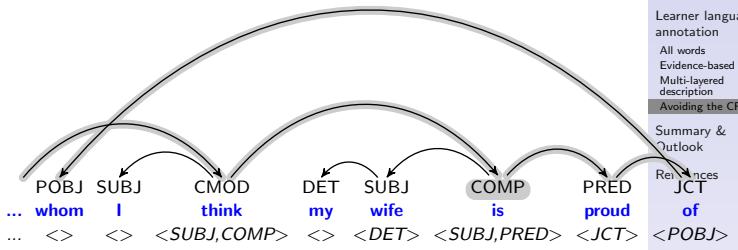
- We can now talk about things beyond the words, i.e., linguistic classes of surface forms

Consider *wh*-words again:



Usefulness of the annotation (2)

The annotation allows us to determine the depth of embedding:



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

References

Aijmer, K. (2002). Modality in advanced swedish learners' written interlanguage. In Granger, S., Hung, J., and Petch-Tyson, S., editors, *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Language Learning and Language teaching, pages 55–76. John Benjamins.

Allen, J. and Core, M. (1996). Draft of DAMSL: Dialog act markup in several layers. Technical report, University of Rochester, Department of Computer Science.

Bardovi-Harlig, K. (1999). Examining the role of text type in l2 tense-aspect research: Broadening our horizons. In Robinson, P. and Jungheim, N., editors, *Proceedings of the Third Pacific Second Language Research Forum*, volume 1, pages 129–138, Tokyo. PacSLRF.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, 33(1):1–17.

Boyd, A. (2010). Eagle: an error-annotated corpus of beginning learner german. In *Proceedings of LREC-10*, Malta.

Díaz-Negrillo, A., Meurers, D., Valera, S., and Wunsch, H. (2010). Towards interlanguage pos annotation for effective learner corpora in sla and flt. *Language Forum*, 36(1–2). Special Issue on New Trends in Language Teaching.

Dickinson, M. and Ragheb, M. (2009). Dependency annotation for learner corpora. In *Proceedings of TLT-8*, Milan, Italy.

Erk, K., Kowalski, A., Pado, S., and Pinkal, M. (2003). Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In *Proceedings of the 41st Annual Meeting of the Association*

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

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

Hajič, J. (1998). Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In Hajičová, E., editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pages 12–19. Prague Karolinum, Charles University Press.

Hajičová, E. (1998). Prague Dependency Treebank: From analytic to tectogrammatical annotation. In *Proceedings of the First Workshop on Text, Speech, Dialogue*, pages 45–50, Brno, Czech Republic. http://ufal.ms.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/index.html.

Ionin, T., Ko, H., and Wexler, K. (2004). Article semantics in l2 acquisition: The role of specificity. *Language Acquisition*, 12(1):3–69.

Juffs, A. (2005). The influence of first language on the processing of wh-movement in english as a second language. *Second Language Research*, 21(2):121–151.

Kingsbury, P., Palmer, M., and Marcus, M. (2002). Adding semantic annotation to the Penn Treebank. In *Proceedings of the Human Language Technology Conference (HLT 2002)*, San Diego, California. http://www.cis.upenn.edu/~ace/HLT2002-propbank.pdf.

Lakshmanan, U. and Selinker, L. (2001). Analysing interlanguage: how do we know what learners know?. *Second Language Research*, 17(4):393 – 420.

Leech, G. (1997). *A Brief Users' Guide to the Grammatical Tagging of the British National Corpus*. UCREL, Lancaster University. http://www.hcu.ox.ac.uk/BNC/what/gramtag.html.

Lüdeling, A., Walter, M., Kroymann, E., and Adolphs, P. (2005). Multi-level

error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*, Birmingham.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330. http://acl.ldc.upenn.edu/J/J93/J93-2004.pdf.

Meurers, W. D. and Müller, S. (2009). Corpora and syntax (article 42). In Lüdeling, A. and Kytö, M., editors, *Corpus linguistics*, volume 2, pages 920–933. Mouton de Gruyter, Berlin.

O'Grady, W., Lee, M., and Choo, M. (2003). A subject-object asymmetry in the acquisition of relative clauses in korean as a second language. *Studies in Second Language Acquisition*, 25(03):433–448.

Rozovskaya, A. and Roth, D. (2010). Annotating esl errors: Challenges and rewards. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36, Los Angeles, California. Association for Computational Linguistics.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., and Wintner, S. (2007). High-accuracy annotation and parsing of childes transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32, Prague.

Sagae, K., MacWhinney, B., and Lavie, A. (2004). Adding syntactic annotations to transcripts of parent-child dialogs. In *Proceedings of LREC-04*, Lisbon.

Sampson, G. (1995). *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Clarendon Press, Oxford.

Schachter, J. (1989). Testing a proposed universal. In Gass, S. and Schachter, J., editors, *Linguistic Perspectives on Second Language Acquisition*, pages

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

73–88. Cambridge University Press.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1995). Vorläufige guidlines fr das taggen deutscher textcorpora mit STTS. Technical report, IMS, Univ. Stuttgart and SfS, Univ. Tübingen. http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts-guide.ps.gz.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*, pages 88–95, Washington, D.C. http://www.coli.uni-sb.de/~thorsten/publications/Skut-ea-ANLP97.ps.gz.

Suri, L. Z. and McCoy, K. F. (1993). A methodology for developing an error taxonomy for a computer assisted language learning tool for second language learners. Technical Report 93–16, Department of Computer and Information Sciences, University of Delaware, Newark, DE.

Tomaselli, A. and Schwartz, B. (1990). Analysing the acquisition stages of negation in l2 german: support for ug in adult second language acquisition. *Second Language Research*, 6:1–38.

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