Avoiding the Comparative Fallacy in the Annotation of Learner Corpora

Marwa Ragheb and Markus Dickinson
Indiana University

1. Introduction

It is becoming more common to use corpora of second language learner data in order to support research on various second language acquisition (SLA) topics (e.g., Römer, 2009; Wulff et al., 2009), but there has been little use of corpus annotation. For many questions in SLA research, using a corpus is simple and in no need of annotation: one can search a corpus for specific words to find relevant examples. For example, if one wants to examine how modal verbs are used by L2 learners (cf., e.g., Aijmer, 2002), one can search for those specific lexical items (*can*, *should*, etc.) and analyze the output by hand.

Consider a search for syntactic patterns, however, such as examining *wh* movement (e.g., Juffs, 2005; Wolfe-Quintero, 1992; Schachter, 1989). These types of questions require more linguistic abstraction (cf., e.g., Lüdeling, 2010). If we take the learner sentence (1), for example, what kind of search involving specific words addresses questions about the function of *whom*?[1] If we search for all instances of *whom* in a corpus, we still have to determine whether this is a relative clause marker, whether this is subject or object extraction, or what the depth of embedding is; and then we need to do the same for *that*, *which*, or even other prepositional objects. We need the data marked with syntactic annotation.

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

Examining abstract linguistic properties is a common issue in SLA research, and if corpora are to increase in usage, they must able to address searches for different realizations of negation (e.g., Tomaselli & Schwartz, 1990), the marking of definiteness or indefiniteness (or lack thereof) (cf., e.g., Ionin et al., 2004), the usage of (headless) relative clauses (e.g., Izumi, 2003; O'Grady et al., 2003), and so forth. Currently, these properties must be searched for by hand.

To investigate such issues, we need the data marked up, or *annotated*, with grammatical properties, to serve as indices for searching. Otherwise, relevant instances will not be found, and many non relevant instances will have to be sorted through (see, e.g., Meurers & Müller, 2009). One can use natural language processing (NLP) technology to automatically add part of speech and syntactic information to the data (e.g., Biber, 1995), but this is unsatisfactory, as these tools were generally developed for the language of native speakers. To add annotation to learner data, we first need to know what the annotation should look like, in order to account for the complexities inherent in learner language. We thus focus on the question of *how* we can define annotation which supports the investigation of learner language.

Specifically, in defining an annotation scheme which is appropriate for learner language, we must account for the comparative fallacy (Bley-Vroman, 1983): how can we define linguistic categories which do not fall prey to an overcomparison to either the L1 or the L2? How can we add linguistic annotation that facilitates looking at interlanguage as a system in its own right? To explore this question, we review existing annotation schemes, both for learner language and for general linguistic annotation, in section 2. In light of these schemes, we discuss the comparative fallacy in section 3 before outlining our proposal

*We would like to thank Detmar Meurers, Kathleen Bardovi-Harlig, David Stringer, Rex Sprouse, Stuart Davis and the IU Computational Linguistics discussion group and Second Language Studies colloquium audience for discussions and advice about our work. We also thank the two anonymous reviewers for their useful comments.

†mragheb@indiana.edu, md7@indiana.edu

[1]Unless otherwise noted, all examples are from our corpus, as outlined in section 2.3.

in section 4. We argue that by: a) annotating all words, b) relying on linguistic evidence in the data, and, most importantly, c) breaking the linguistic annotation into multiple layers for each piece of evidence, we can provide useful information for SLA research which avoids the comparative fallacy, as we discuss in section 5. To date, there has been some research on this topic (Díaz-Negrillo et al., 2010; Dickinson & Ragheb, 2009; Rastelli, 2009), but none fully working out the ramifications for the comparative fallacy.

2. Existing types of annotation

2.1. Error annotation

Annotation adds information beyond the raw text present in a corpus. Learner corpora have received some attention in terms of annotation, and the majority of work on annotation of learner corpora has focused on annotating learner errors (e.g., Suri & McCoy, 1993; Granger, 2003; Nicholls, 2003; Lüdeling et al., 2005; Boyd, 2010; Hana et al., 2010; Rozovskaya & Roth, 2010). We will briefly survey a few annotation schemes, highlighting the types of information encoded and how learner data is viewed (see also Díaz-Negrillo & Fernández-Domínguez, 2006). In section 5.2, we will return to some of the issues involved in annotating errors and target forms.

A good starting point to examine error annotation is the FRIDA corpus (Granger, 2003). The annotation scheme makes use of three levels: error domain (e.g., form, morphology, grammar, lexis, syntax), error category (e.g., inflection, gender, noun complementation, word order) and word category (i.e., part of speech). Every error is tagged with this information, as well as given a correction, i.e., target form. An example of this type of annotation can be seen in (2), taken from Granger (2003), where the error is in the grammar domain (G), of the type number (NBR), and with a word category of finite simple (VSC). The correction (*pensent*) is indicated before the actual word used by the learner (*pense*).

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

This annotation makes it possible to retrieve different types of errors, for example, all errors related to finite simple verbs by searching for the tag <VSC>. Thus, a corpus annotated in this way can be used to find statistics about errors and their types, and rank them according to frequency of occurrence, as is done in Granger (2003). Such linguistic annotation can be seen as an advantage over annotation not encoding part of speech. For example, in the MELD corpus, Fitzpatrick & Seegmiller (2004) mark errors but do not label them with tags. Rather, they mark the correction after an error, such as in (3). An example of this type of error annotation can be seen in (3).

(3) I would prefer to study at home rather {**then/than**} {**going/go**} to a traditional {**schools/school**}.

While linguistic annotation can be seen as a benefit, in both of these examples it is important to note that not every word is annotated, only the erroneous ones.

Relatedly, there have been schemes developed for specific types of errors, such as collocations (Ramos et al., 2010), Korean particles (Lee et al., 2009), etc. Rozovskaya & Roth (2010) focus on ESL errors of prepositions, articles, grammar, word choice, and word order. The goal of their annotation is to "correct all mistakes in the sentence," but they do not have an expansive annotation taxonomy. By focusing on not just errors, but specific types of errors, much information is left unannotated, or only very coarsely annotated (e.g., "grammar" error).

As has been noted by several researchers (e.g., Granger, 2008; Díaz-Negrillo & Fernández-Domínguez, 2006), there is a lot of subjectivity in the process of error annotation and interpretation. Because error annotation can fall prey to an annotator selecting one interpretation over another, some annotation schemes have subsequently made use of multi layered annotation. For the FALKO corpus of L2 German (Lüdeling et al., 2005), for example, they use multi layered standoff annotation — i.e., annotation that is stored separately from the raw text. This allows them to make sure that they can encode multiple target hypotheses, if annotators see multiple interpretations, as well as code the respective errors, using independent levels of annotation.[2] Multiple layers of annotation also allow one to treat error annotation as an incremental process, e.g., building from smaller constituents to larger ones (Boyd, 2010; Hana et al., 2010). We will discuss multi layer (error) annotation more in sections 4.3 and 5.

[2]This format also has some technical advantages, such as making it easy to annotate errors that span more than one word.

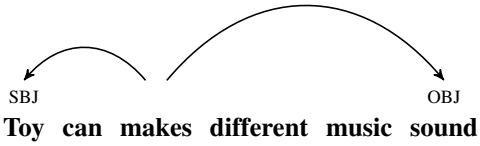

Figure 1: Basic grammatical relations (i.e., **dependencies**): the verb *makes* is the head of the sentence, with *Toy* as its subject (SBJ) and *sound* as its object (OBJ).

Also germane to our approach is how errors are categorized. Lüdeling et al. (2005) encode errors in L2 German that have to do with orthography, word formation, agreement, government, tense, mood, word order and expression. Boyd (2010) uses an error annotation scheme that encodes grammatical errors (word form, selection, agreement), as well as word order and punctuation errors. If we were to look at the details of their taxonomies, these two examples would illustrate that, even with error annotation, there needs to be a well defined set of linguistic properties. After all, errors are defined with respect to these properties. An alternative which we pursue is to take these same types of properties (e.g., selection/government) and encode them for all words, whether or not errors are encoded.

2.2. Linguistic annotation

But how easy is it to encode such linguistic properties? If we consider word category, or part of speech (POS) annotation from the FRIDA corpus, as in the tag VSC in example (2), we can see that POS tags are only defined for erroneous words. More than that, whenever they are defined, it is not clear what the tag should be for novel POS uses, as we see in example (4) from Díaz-Negrillo et al. (2010).

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

In this example, the word *subjectives* distributionally appears to be an adjective, and morphologically a noun, since it has the English plural marker. In this case, both pieces of information seem to be relevant to describing the features of the learner's developing interlanguage, a topic we take up again in section 4.

Outside of learner language, the issue of annotating linguistic properties is not a new one. Linguistic annotation can contain information about lemmata, morphology, and part of speech (POS) (e.g., Leech, 1997; Sampson, 1995; Schiller et al., 1995); syntactic constituencies and dependencies (e.g., Marcus et al., 1993; Hajič, 1998; Skut et al., 1997); semantic roles and word senses (e.g., Palmer et al., 2005; Hajičová, 1998; Erk et al., 2003); and discourse properties (e.g., Allen & Core, 1996). In other words, all different levels of linguistic analysis are annotated. The question we want to investigate is how we can apply these types of annotation to learner language.

Specifically, when we talk about linguistic properties, our focus is going to be on morphosyntactic and syntactic annotation, such as in figure 1, where we see a subset of dependency relations for a sentence. By focusing on linguistic properties, we can adapt linguistic annotation, i.e., the specific tagging schemes and best practices, from previous annotation efforts.

2.3. Our context

We have begun annotating data from learners of varying levels. To pilot the annotation, we used narratives collected from the 1990s (Bardovi-Harlig, 1999). Students were asked to watch a short, silent cartoon and write about the events in that story. We also sampled from another corpus of essays collected through the Intensive English Program (IEP) at Indiana University; these were timed exams used for course placement, where students responded to a prompt such as "What are your plans for life?" We are in the early stages of applying the annotation, and annotated data will be released at a future point. In determining the process of annotation, we have observed the importance of not falling into the trap of the comparative fallacy, which we briefly outline in the following section.

3. The comparative fallacy

The *comparative fallacy* refers to the "mistake of studying the systematic character of one language by comparing it to another" (Bley-Vroman, 1983). In his criticism about how some SLA research methodology proceeds, Bley-Vroman (1983) stresses that the language system constructed by a second language (L2) learner is not a "degenerate form" of the target language (TL), and interlanguage (IL) is a system in itself that should be studied. According to him, studies which commit the comparative fallacy include studies tabulating errors or attempting to classify the interlanguage according to certain criteria such as *omission*, *substitution*, etc., as these types of studies depend on comparing the interlanguage to "hypothesized corresponding TL sentences." On the other hand, Bley-Vroman (1983) mentions that investigations of interlanguage could be furthered if researchers focus on constructing a *linguistic description* of a learner's language (p. 16).

While Bley-Vroman (1983) talks about the comparative fallacy with respect to comparisons with the target language, Lakshmanan & Selinker (2001) also stress the point made by Adjemian (1976) to investigate IL independent of the native language (L1) as well. Lakshmanan & Selinker (2001) give examples of how researchers can be influenced by the L1 of the learners when analyzing certain phenomena (e.g., the use of null subjects by L1 Spanish speakers who are learners of English), concluding that comparing with the L1 could obscure systematicity in interlanguage.

As for annotation, to avoid the comparative fallacy is to provide annotation which does not overcompare a learner's language system to either the L2 or the L1, thereby avoiding the comparative fallacy by not viewing learner production as a "degenerate form" of the target language. One such type of annotation would be a linguistic description of learner production applied to all the text, marking linguistic phenomena in all their occurrences, whether target like or non target like. Such a linguistic description will make it possible to capture the first emergence of syntactic phenomena (cf., Lakshmanan & Selinker, 2001), such as relative clauses or preposition stranding, or phenomena which are relatively rare. We will discuss the implications of different types of annotation with respect to the comparative fallacy in section 5, after providing more specifics of our proposal for linguistic annotation next.

4. Our proposal

Taking into account our desire to annotate linguistic properties in a way which avoids the comparative fallacy, we advocate certain principles for annotating learner language. Although we will present examples from a specific annotation scheme, the important part is not the details of the scheme (e.g., NP1x to mean 'proper noun'), but the methodology we use to annotate.

The first principle we propose is to *encode linguistic properties for every word, not just "errors."* This way we can access all the learner's forms and categories, and we avoid focusing only on obligatory contexts. The second is to *use linguistic evidence when assigning linguistic properties*. We emphasize annotating observable forms, based on the evidence present in the surrounding context (i.e., the sentence). The third principle is to *describe the data as it appears, by separating linguistic properties into multiple layers*. In this way, each different piece of linguistic evidence can be encoded on a separate layer. With such principles, annotation efforts will be less likely to fall into the comparative fallacy. We will briefly illustrate these three principles in what follows.

4.1. All words

The first principle is simple: annotate all words. Errors are inherently exclusive, applying only to erroneous words, whereas linguistic annotation is by its nature applicable to all words. Figure 2 illustrates a particular way that we realize linguistic annotation for learner language , wherein we encode part of speech (POS) and syntactic dependency information.

As we will describe the motivation for in section 4.3, the POS and dependency annotations come in multiple layers. For POS tags,[3] we use two layers, one for morphological information and the other for distributional information. In figure 2, these are represented in the two rows under each word. We

[3]We use POS tags from SUSANNE tagset (Sampson, 1995).

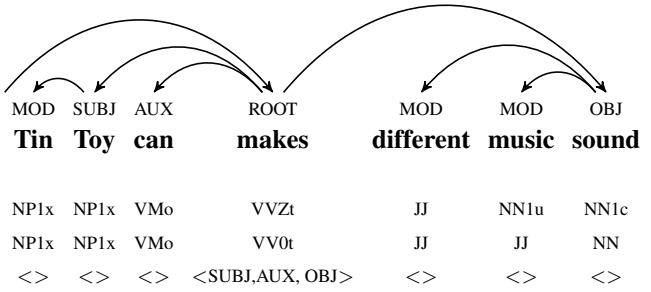

Figure 2: Linguistic annotation of a sample sentence. Annotation includes dependency relations (top row & arcs), part of speech information (rows 3 & 4), and subcategorization (bottom row).

encode surface dependencies,[4] as shown in the arcs going from each head to its dependent(s), and the labels on these arcs. For example, we can see that *makes* is the head of *Toy*, and the grammatical relation is SUBJ (i.e., subject). We also include subcategorization information in the last row, with arguments between angled brackets. The linguistic properties are determined on the basis of linguistic evidence, explained next.

4.2. Evidence based

Again, the second principle may be obvious: use linguistic evidence. By this, however, we mean that, instead of relying on learner intention, knowledge of L1 and how that may affect L2 production, SLA theory, or interpreting what this structure may mean in the learner's interlanguage, we rely on *linguistic* evidence to annotate the data.

This is not a trivial matter, though. Consider POS tags, where POS is defined by both morphological and distributional criteria (e.g., Sampson, 1995). A learner may have evidence pointing different ways, as we can see in (5):

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

For the verb *makes*, the morphological evidence shows it is a 3rd person present tense verb, whereas the distributional evidence points to its being in the slot of a base form verb. A single POS tag cannot capture these two pieces of conflicting evidence. This leads to the use of multiple layers, described next.

4.3. Multi layered description

As as can see from the above example, relying on observable evidence (in this case, about *makes*) leads to multi layered annotation, where we can encode separate pieces of evidence on separate layers of linguistic annotation (cf. Díaz-Negrillo et al., 2010). For *can makes*, *makes* can be annotated as a 3rd person present tense verb (VVZ) on the *morphological layer* and as a base form verb (VV0) on the *distributional layer*, as highlighted in figure 3. This means that each layer can contain a description of a linguistic property.[5] In a sense, this is parallel to the multi layered annotation in Lüdeling et al. (2005), but instead of making each error hypothesis explicit, each grammatical component is made explicit.

With these three general principles, we have outlined a way of encoding linguistic annotation for learner data, capturing relevant properties for much SLA research. The remaining question is: is this a valid way to do annotation? Or, will we be falling into the same traps which plague other forms of annotation?

[4] We use the dependency annotation scheme developed for CHILDES data (Sagae et al., 2007, 2004).

[5] We are in the process of adding an additional layer of dependencies, to represent linguistic evidence also diverging in terms of morphology and distribution (see Dickinson & Ragheb, 2011). One could consider annotating semantically driven dependencies, but this would require more information than is generally available in an essay.

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

Figure 3: Morphological and distributional POS tags. The highlighted tags reflect conflicting evidence between the layers for the token *makes*.

5. Avoiding the comparative fallacy

5.1. Our proposal

If we stop and consider the proposal, we are still defining each layer of annotation in terms of L2 properties. Taking our running example of *makes* in (5), the morphology of "3rd person singular present tense" is defined by the presence of *-s* on a verb, and 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, in this case English. Are we avoiding the comparative fallacy, and, if so, how?

At this point, it is important to remember what the annotation does and does not say. It says that the morphology is indicative of 3rd singular present tense, while the distributional properties are those of a base form verb. Crucially, the annotation does not say that the learner is using (or intending to use) this as either a 3rd singular or a base form verb. Multi layered annotation does not make a definitive claim about a single annotation, and thus the annotation in no way indicates that the sentence is a *degenerate* L2 form or *should be* any other form; it provides a description that is closely tied to the data. Also, basing the labels on the L2 (instead of, e.g., the L1) should keep the annotation comparable across learners.

Another way to view this is that the annotation labels are intended to be indices to help access the grammatical categories used by learners; they are not in themselves theoretical entities of the interlanguage. Researchers may use this annotation to access categories of interest, ascribing to them their own interpretation. In general, then, such multi layered annotation seems well suited to avoiding the comparative fallacy, insofar as it: a) treats interlanguage as a system in its own right, and b) enables SLA research that avoids the comparative fallacy. One lingering question, though, is: which other types of annotation do or do not readily lead to the comparative fallacy?

5.2. Error annotation

To address this, we must first make clear what it means to annotate errors and target forms. Annotating errors implies that there is a distance from what the learner produced to some notion of correctness (i.e., grammaticality), but one is not necessarily forced to state what is correct. To state that there is subject verb disagreement, for example, is different than correcting the subject or the verb. Even more neutral is to simply say that there is a problem without specifying its nature. In our annotation scheme, for example, noting discrepancies between layers points to something "problematic" but does not state what the issue is, nor definitively claim that it is an error so much as an area with divergent linguistic evidence.

Annotating target forms, on the other hand, commits one to a particular analysis, or set of analyses (Lüdeling et al., 2005). However, the distance from the learner production to some target form may or may not indicate that this target was intended. [6] In Meurers et al. (2010), for instance, the target forms are defined with respect to answers for an exercise, thereby not implying that the learner form is corrupted so much as it does not answer the question. Therefore, a distinction should be made with respect to different types of target forms/hypotheses: those that arise in the context of answers to exercises — and the semantic content of which could be predicted — and those that attempt to get at learner intention, or reconstruct the form to what the learner "should have said." It can be easy to define the answer to an exercise, whereas the difficulty of getting at learner intention has been noted by many (e.g., Lakshmanan

[6]Thanks to Detmar Meurers for pointing this out to us.

& Selinker, 2001), and it has also been noted that attempts at reconstructing learner text have issues of low inter annotator agreement (Fitzpatrick & Seegmiller, 2004; Meurers, 2011).

Which of these annotations are prone to the comparative fallacy is perhaps a matter of debate, but there are a few things we can say with some degree of certainty. First, annotating target forms which express a form a learner *should have said* or learner intention in the L2 falls into the comparative fallacy, as these are cases of the learner language being treated as a degenerate form of the L2. Secondly, explicit error annotation which provides an *interpretation* of what the learner did in terms of distance from the L2 commits the comparative fallacy. For example, specifying a noun gender error says that the learner production is distant from an assumedly correct form where the gender is different, as opposed to, for example, a correct form with a different noun number — which could be an equally plausible interpretation of the same "error".[7] What is less clear is how to classify error annotation which simply says that a particular position within a sentence is an "error," as there is no specification of how it differs from the L2. Likewise, target forms defined with respect to exercise answers specify a semantic distance based on a desired answer and seem to be orthogonal to discussions of the comparative fallacy.

Tenfjord et al. (2006) argue that annotating errors does not necessarily succumb to the comparative fallacy because "error analysis is a *method*, not a theory of SLA" (p. 102). Citing Corder (1973), they argue that error annotation is a superficial description of the data and does not make theoretical claims about SLA processes. More specifically, Tenfjord et al. (2006) take a learner text, reconstruct it, and annotate grammatical properties of the reconstructed text. A crucial property of this work is the stage of reconstruction, i.e., altering the learner text into well formed L2 text. Although it rests on a similar descriptive approach to annotating the data as we take, it is done in terms of *deviations* from the target language grammar (p. 103).

As Rastelli (2009) points out, this type of methodology "is at risk of being misleading for SLA research purposes." Discussing a "missing auxiliary," he says:

> In fact the error tag "missing auxiliary" seems to indicate that a learner violates a real rule of the TL. In which sense do we say that a 'missing auxiliary' violates a rule? Does such a rule exist in a learner's mind? Is it a rule of the TL or a rule of the IL? The procedure of error tagging is risky because it ontologizes TL grammar rules as if they were psychological *realia* in a learner's mind.

Part of the issue here seems to be what constitutes a "theory": in Tenfjord et al. (2006), a theory seems to specifically refer to SLA theories about processes the learner is using. In that sense, annotation is a method and not an SLA theory. Another viewpoint stems from the idea that any annotation is inherently interpretive (Leech, 2004); in some sense, then, all annotation is theoretical. While it is too far afield to get into such a debate, we can note the following: whether a method or theory, error annotation which explicitly encodes deviations from an L2 grammar best supports SLA theories based on deviations from an L2 grammar, and such theories are prone to the comparative fallacy.

5.3. Our proposal revisited

While our annotation in principle avoids the comparative fallacy (see section 5.1), there are some caveats regarding the degree of interpretation needed to annotate. Firstly, there will be examples which are so uninterpretable as to have parts with no possible reasonable annotation. In general, however, there will still be parts which can be annotated. In (6), for instance, we may not know what the learner intended, but we can still annotate linguistic properties relevant to agreement, adverb placement, etc.

(6) That make me really kindly strong.

Additionally, while interpreting the linguistic properties of learner data may differ in quantity from native speaker data, it may not actually differ in quality; for example, the Penn Treebank tagset — which has largely been applied to newspaper text — defines X for "unknown, uncertain, or unbracketable" structures (Bies et al., 1995).

[7] Allowing for multiple interpretations and being explicit about the nature of the interpretations alleviates this to some extent by making the assumptions clear for users of the annotation (see Lüdeling et al., 2005).

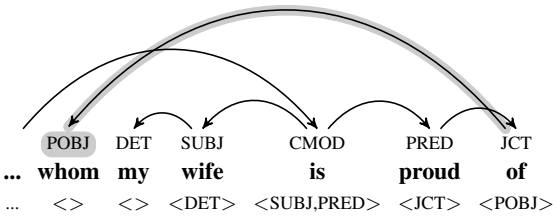

Figure 4: Dependency arcs encode the relationship between a *wh* word *whom* and its head *of*.

Secondly, our annotation requires some degree of interpretation — as all annotation does, since annotation is inherently interpretive (Leech, 2004). Crucially, though, the interpretation is on the level of determining a linguistic category, not in terms of determining the deviation from a target or in terms of any SLA process. Since it is on the "same" level as general linguistic annotation, such annotation schemes and practices can be re used, such as how to handle ambiguous structures (e.g., PP attachment).

In addition to how linguistic annotation for learner language can avoid the comparative fallacy, there are also additional benefits to our approach. Providing a description of the linguistic properties of the entire text allows one to study issues that may not be as clear with error annotation, such as avoidance (Ellis, 2008). Learners of a certain L1 may be making more "errors" with respect to a certain syntactic phenomenon, but that might be because they are making more attempts at it (cf. the study of the acquisition of relative clauses by Schachter (1974), as cited in Ellis (2008)). A linguistic description of the whole text can show what the learner is doing "right," as well as what the learner is doing "wrong." Indeed, this type of annotation can augment studies of contrastive interlanguage analysis (CIA) which focus on comparing the use of particular grammatical properties across different groups (Granger, 2004). Likewise, this approach can also help avoid issues that arise due to only looking at obligatory contexts (Ellis & Barkhuizen, 2005), as again, we are not focused on target forms.

Practically speaking, the type of annotation we propose is suitable for investigating a range of theoretical questions. The encoded information will make it easier to search for specific linguistic properties in the learner corpus, as we can now talk about things beyond the words, i.e., linguistic classes of surface forms. Consider *wh* words again, as in the abbreviated example in figure 4. Here, we see that *whom* is a prepositional object of *of*, despite preceding it. Furthermore, when contrasting this example with the one in figure 5, we can tell that the prepositional object occurs at a different level of embedding. Other research questions focusing on IL syntax can also benefit from this type of annotation. As but one example, our syntactic annotation differentiates between matrix and embedded clauses. This distinction would help researchers with different questions related to the acquisition of IL syntax that involve main and subordinate clauses, such as question formation (cf., e.g., Haznedar, 2003; Lightbown & Spada, 1993; Pienemann et al., 1988; White, 1986), extraction and *wh* movement (e.g., Juffs, 2005; Wolfe-Quintero, 1992; Schachter, 1989), word order (e.g., Schwartz & Sprouse, 1994; DuPlessis et al., 1987; Clahsen & Muysken, 1986), and relative clauses (e.g., Lee-Ellis, 2011).

6. Conclusions and Outlook

We have discussed designing an annotation scheme for learner language, in a way which emphasizes encoding linguistic properties, as opposed to learner intention, and thereby avoids the comparative fallacy. The key elements of this annotation scheme lie in annotating all words, using linguistic evidence, and breaking the annotation into different layers for different pieces of evidence. Using such annotation will allow for better descriptions of interlanguage properties and thus eventually better searching. Error annotation has proven to be useful for contrastive interlanguage analysis (CIA) (Granger, 2004) and in providing data for intelligent computer assisted language learning (ICALL) systems (see discussion in Meurers, 2011), and it will continue to serve those purposes well. However, such annotation cannot provide the information needed for many strands of SLA research, where linguistic information for all the data is needed. Additionally, to engage in error annotation, one often needs to be working with data

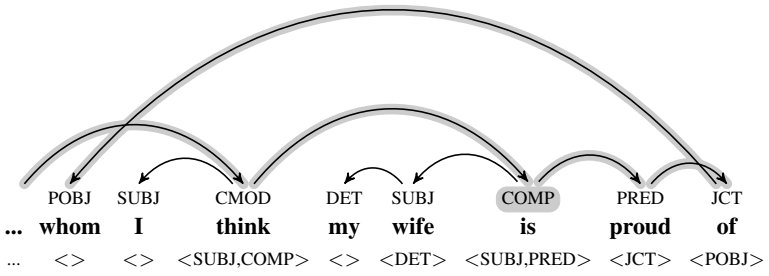

Figure 5: Dependency arcs encode the depth of embedding

where the intention is clear (see discussion in section 5), thus generally favoring the data of advanced learners. By contrast, what we have presented here is applicable for learners of all levels.

There are several next steps to pursue, the first being to refine the annotation scheme and test inter annotator agreement. In tandem with this, we are collecting and annotating learner data that will eventually be made publicly available. This annotation does not answer SLA questions directly, but it provides a platform for others to answer such questions. Finally, we will develop computational technology for tagging and parsing the type of multi layered annotation we propose. Splitting the annotation out into different layers presents challenges for adapting technology for one layer, but if researchers are going to search over a lot of learner data, or use a new set of learner data, this new data will not be annotated, and so we will have to rely on (semi)automatic methods to provide the desired annotation.

References

Adjemian, Christian (1976). On the nature of interlanguage systems. *Language Learning* 26:2, 297–320.

Aijmer, Karin (2002). Modality in advanced Swedish learners' written interlanguage. Granger, Sylviane, Joseph Hung & Stephanie Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Language Learning and Language teaching, John Benjamins, 55–76.

Allen, James & Mark Core (1996). Draft of DAMSL: Dialog act markup in several layers. Tech. rep., University of Rochester, Department of Computer Science.

Bardovi-Harlig, Kathleen (1999). Examining the role of text type in L2 tense-aspect research: Broadening our horizons. Robinson, Peter & Nicholas O. Jungheim (eds.), *Proceedings of the Third Pacific Second Language Research Forum*, PacSLRF, Tokyo, vol. 1, 129–138.

Biber, Douglas (1995). *Dimension of Register Variation*. Cambridge University Press, Cambridge.

Bies, Ann, Mark Ferguson, Karen Katz & Robert MacIntyre (1995). *Bracketing Guidelines for Treebank II Style Penn Treebank Project*. University of Pennsylvania.

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

Boyd, Adriane (2010). EAGLE: an error-annotated corpus of beginning learner German. *Proceedings of LREC-10*, Malta.

Clahsen, Harald & Pieter Muysken (1986). The availability of universal grammar to adult and child learners - a study of the acquisition of German word order. *Second Language Research* 2:2, 93–119.

Corder, Stephen Pit (1973). *Introducing Applied Linguistics*. Penguin Education, Harmondsworth.

Díaz-Negrillo, Ana & Jesús Fernández-Domínguez (2006). Error tagging systems for learner corpora. *RESLA* 19, 83–102.

Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera & Holger Wunsch (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, Markus & Marwa Ragheb (2009). Dependency annotation for learner corpora. *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*, Milan, Italy, 59–70.

Dickinson, Markus & Marwa Ragheb (2011). Dependency annotation of coordination for learner language. *Proceedings of the International Conference on Dependency Linguistics (Depling 2011)*, Barcelona, Spain.

DuPlessis, Jean, Doreen Solin, Lisa Travis & Lydia White (1987). UG or not UG, that is the question: a reply to Clahsen and Muysken. *Second Language Research* 3:1, 56–75.

Ellis, Rod (2008). *The Study of Second Language Acquisition*. Oxford University Press, Oxford, second edn.

Ellis, Rod & Gary Barkhuizen (2005). *Analyzing Learner Language*. Oxford University Press, Oxford.

Erk, Katrin, Andrea Kowalski, Sebastian Pado & Manfred Pinkal (2003). Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, Sapporo, Japan, 537–544.

Fitzpatrick, Eileen & M.S. Seegmiller (2004). The Montclair electronic language database project. *Language and Computers* 52, 223–237(15).

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

Granger, Sylviane (2004). Computer learner corpus research: current status and future prospects. Connor, Ulla & Thomas A. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, Rodopi, Amsterdam & Atlanta, 123–145.

Granger, Sylviane (2008). Learner corpora. Lüdeling, Anke & Merja Kytö (eds.), *Corpus Linguistics: An international handout*, Mouton de Gruyter, vol. 1, 259–275.

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

Hajičová, Eva (1998). Prague Dependency Treebank: From analytic to tectogrammatical annotation. *Proceedings of the First Workshop on Text, Speech, Dialogue*, Brno, Czech Republic, 45–50.

Hana, Jirka, Alexandr Rosen, Svatava Škodová & Barbora Štindlová (2010). Error-tagged learner corpus of Czech. *Proceedings of the Fourth Linguistic Annotation Workshop*, Association for Computational Linguistics, Uppsala, Sweden, 11–19.

Haznedar, Belma (2003). The status of functional categories in child second language acquisition: evidence from the acquisition of cp. *Second Language Research* 19:1, 1 – 41.

Ionin, Tania, Heejeong Ko & Kenneth Wexler (2004). Article semantics in L2 acquisition: The role of specificity. *Language Acquisition* 12:1, 3–69.

Izumi, Shinichi (2003). Processing difficulty in comprehension and production of relative clauses by learners of English as a second language. *Language Learning* 53:2, 285–323.

Juffs, Alan (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.

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

Lee, Sun-Hee, Seok Bae Jang & Sang kyu Seo (2009). Annotation of Korean learner corpora for particle error detection. *CALICO Journal* 26:3.

Lee-Ellis, Sunyoung (2011). The elicited production of Korean relative clauses by heritage speakers. *Studies in Second Language Acquisition* 33:01, 57–89.

Leech, Geoffrey (1997). *A Brief Users' Guide to the Grammatical Tagging of the British National Corpus*. UCREL, Lancaster University.

Leech, Geoffrey (2004). Adding linguistic annotation. Wynne, Martin (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxbow Books, Oxford, 17–29.

Lightbown, Patsy & Nina Spada (1993). *How languages are learned*. Oxford University Press, Oxford.

Lüdeling, Anke (2010). Syntactic overuse and underuse. a study of a parsed learner corpus and its target hypothesis. Invited talk given at the Ninth International Workshop on Treebanks and Linguistic Theories. University of Tartu.

Lüdeling, Anke, Maik Walter, Emil Kroymann & Peter Adolphs (2005). Multi-level error annotation in learner corpora. *Proceedings of Corpus Linguistics 2005*, Birmingham.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19:2, 313–330.

Meurers, Detmar (2011). Natural language processing and language learning. Chapelle, Carol A. (ed.), *Encyclopedia of Applied Linguistics*, Blackwell. To appear.

Meurers, Detmar, Niels Ott & Ramon Ziai (2010). Compiling a task-based corpus for the analysis of learner language in context. *Proceedings of Linguistic Evidence 2010*, Tübingen, 214–217.

Meurers, Walt Detmar & Stefan Müller (2009). Corpora and syntax (article 42). Lüdeling, Anke & Merja Kytö (eds.), *Corpus linguistics*, Mouton de Gruyter, Berlin, vol. 2, 920–933.

Nicholls, Diane (2003). The cambridge learner corpus - error coding and analysis for lexicography and ELT. Archer, Dawn, Paul Rayson, Andrew Wilson & Tony McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster University, University Centre for Computer Corpus Research on Language, Technical Papers, vol. 16, 572–581.

O'Grady, William, Miseon Lee & Miho Choo (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.

Palmer, Martha, Daniel Gildea & Paul Kingsbury (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31:1, 71–105.

Pienemann, M., M. Johnston & G. Brindley (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition* 10, 217–243.

Ramos, Margarita Alonso, Leo Wanner, Orsolya Vincze, Gerard Casamayor del Bosque, Nancy Vzquez Veiga, Estela Mosqueira Surez & Sabela Prieto Gonzlez (2010). Towards a motivated annotation schema of collocation errors in learner corpora. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Malta.

Rastelli, Stefano (2009). Learner corpora without error tagging. *Linguistik online* 38.

Römer, Ute (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics* 7, 140–162.

Rozovskaya, Alla & Dan Roth (2010). Annotating ESL errors: Challenges and rewards. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, Los Angeles, California, 28–36.

Sagae, Kenji, Brian MacWhinney & Alon Lavie (2004). Adding syntactic annotations to transcripts of parent-child dialogs. *Proceedings of LREC-04*, Lisbon.

Sagae, Kenji, Eric Davis, Alon Lavie, Brian MacWhinney & Shuly Wintner (2007). High-accuracy annotation and parsing of childes transcripts. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, Prague, 25–32.

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

Schachter, Jacquelyn (1974). An error in error analysis. *Language Learning* 27, 205–214.

Schachter, Jacquelyn (1989). Testing a proposed universal. Gass, Susan M. & Jacquelyn Schachter (eds.), *Linguistic Perspectives on Second Language Acquisition*, Cambridge University Press, 73–88.

Schiller, Anne, Simone Teufel, Christiane Stöckert & Christine Thielen (1995). Vorläufige guidlines fr das taggen deutscher textcorpora mit STTS. Tech. rep., IMS, Univ. Stuttgart and SfS, Univ. Tübingen.

Schwartz, Bonnie D. & Rex A. Sprouse (1994). Word order and nominative case in non-native language acquisition: A longitudinal study of (L1 Turkish) German interlanguage. Hoekstra, Teun, Kenneth Wexler & Bonnie D. Schwartz (eds.), *Language Acquisition Studies in Generative Grammar: Papers in Honor of Kenneth Wexler from the 1991 GLOW Workshops*, Language Acquisition & Language Disorders, John Benjamins, Philadelphia, 317–368.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants & Hans Uszkoreit (1997). An annotation scheme for free word order languages. *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*, Washington, D.C., 88–95.

Suri, Linda Z. & Kathleen F. McCoy (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.

Tenfjord, Kari, Jon Erik Hagen & Hilde Johansen (2006). The hows and whys of coding categories in a learner corpus (or "how and why an error-tagged learner corpus is not *ipso facto* one big comparative fallacy. *Rivista di psicolinguistica applicata* 6:3, 93–108.

Tomaselli, Alessandra & Bonnie D. Schwartz (1990). Analysing the acquisition stages of negation in L2 German: support for UG in adult second language acquisition. *Second Language Research* 6, 1–38.

White, Lydia (1986). Implications of parametric variation for adult second language acquisition: an investigation of the pro-drop parameter. Cook, Vivian (ed.), *Experimental Approaches to Second Language Learning*, Pergamon Press, Oxford.

Wolfe-Quintero, Kate (1992). Learnability and the acquisition of extraction in relative clauses and wh questions. *Studies in Second Language Acquisition* 14, 39–70.

Wulff, Stefanie, Nick C. Ellis, Ute Roemer, Kathleen Bardovi-Harlig & Chelsea LeBlanc (2009). The acquisition of tense-aspect: Converging evidence from corpora and telicity ratings. *The Modern Language Journal* 93:3, 354–369.

Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions

edited by Gisela Granena, Joel Koeth, Sunyoung Lee-Ellis, Anna Lukyanchenko, Goretti Prieto Botana, and Elizabeth Rhoades

Cascadilla Proceedings Project Somerville, MA 2011

ISBN 978-1-57473-448-5 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Ragheb, Marwa and Markus Dickinson. 2011. Avoiding the Comparative Fallacy in the Annotation of Learner Corpora. In *Selected Proceedings of the 2010 Second Language Research Forum*, ed. Gisela Granena et al., 114-124. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2620.