On Grammaticality in the Syntactic Annotation of Learner Language

Markus Dickinson
Indiana University
Bloomington, IN USA
md7@indiana.edu

Marwa Ragheb
Indiana University
Bloomington, IN USA
mragheb@indiana.edu

Abstract

We examine some non-canonical annotation categories that license missing material (ellipses and enumerations). In extending these categories to learner data, the distinctions seem to require an annotator to determine whether a sentence is grammatical or not when deciding between particular analyses. We unpack the assumptions surrounding the annotation of learner language and how these particular phenomena compare to competing analyses, pointing out the implications for annotation practice and second language analysis.

1 Introduction and Motivation

The grammatical principles underlying linguistic annotation are often only implicit. The implicitness and undercommittal to any particular theory can be beneficial, as it: 1) allows multiple users of the annotation to utilize it in different ways; 2) frees annotators to extend existing categories to unforeseen constructions; and 3) treats annotation as indices for others to derive theories from. Without necessarily having to be a theoretically-driven corpus (Oepen et al., 2004), there are cases, however, where a grammatical model for annotation may need to be made more explicit and the annotation categories more precise. For non-canonical data (e.g., historical, second language, and internet data), a thorough definition of language categories should lead to a consistent application throughout a corpus. As one example, knowing whether a hashtag denotes a syntactic unit (e.g., *Got #college admissions questions ?*) is important for obtaining a syntactic tree for Twitter data (Kong et al., 2014). Even for canonical data, annotation categories are not truly meaningful without some specification or guidelines (Rambow, 2010). We here explore *non-canonical categories* for *non-canonical data*, specifically categories that license "missing" material (ellipsis, enumeration) in the context of second language learner data, and we demonstrate that one needs to make clear to what extent the categories in the grammar underlying the annotation extend to novel constructions.

To gauge the impact on second language data of categories designed to cover more "peripheral" phenomena involving missing material requires investigating, first, how these categories apply in general, and, secondly, how they extend to learner data and how they compare to competing, learner-specific analyses. We refer to categories which license missing (or additional) semantic material as *non-canonical categories*. Applying such categories to learner data makes us question to what degree we need to know whether a sentence is grammatical—where *grammatical* refers to being licensed by the grammar underlying the annotation.

Focusing on the data of second language learners and the annotation of syntactic dependencies, the question of grammaticality is compounded, not just by novel constructions, but by various research practices. First, there is a long literature in second language acquisition (SLA) as to the nature of a second language grammar (*interlanguage*) (Selinker, 1972; Adjemian, 1976; Ellis, 1985; Lakshmanan and Selinker, 2001). Secondly, and sometimes competingly, there are many schemes for annotating learner errors in corpora (Díaz-Negrillo and Fernández-

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 158–167,
Denver, Colorado, June 5, 2015.

Domínguez, 2006; Granger, 2003; Nicholls, 2003; Lüdeling et al., 2005), where direct or indirect reference is made to target (i.e., native) grammars in the annotation of corrections. Part of the tension between these approaches is to what extent the grammatical categories used for native language are applicable to learner data.

Thus, non-canonical categories are worth investigating not just to improve corpus annotation, but also to provide insight into these traditions. In particular, there has been much discussion in SLA regarding the comparative fallacy (Bley-Vroman, 1983; Lakshmanan and Selinker, 2001; Tenfjord et al., 2006), wherein learner language is (over)compared to the target language, and the degree to which such comparison affects the conclusions drawn. The grammatical annotation of learner language is in some sense ideal for providing insight, as it provides a systematic characterization of everything in the data and thus allows one to assess the degree of over-comparison (Ragheb, 2014).

In section 2 we discuss the aims of linguistic annotation for learner data, which leads directly to a unpacking of the grammaticality assumed in such annotation in section 3—examining both the source of the grammar and the way innovative learner examples do or do not fit within the categories given by that grammar. After setting this stage, we turn to our two main areas of phenomena: 1) ellipsis and missing heads (section 4); and 2) coordination, enumeration, and missing conjunctions (section 5). After seeing the issues involved in these categories and in the decision procedure for annotation (section 6)—at least for one annotation scheme—we conclude in section 7 that the main options for annotation are: 1) apply the native categories even to learner innovations; 2) develop tighter restrictions on the native categories; and/or 3) reference sentence-level grammaticality in the definitions of categories.

This paper will likely raise more questions than it provides answers, as "answers" are ultimately going to be specific to one's particular goals and project. However, we believe the questions are crucial to annotating learner language: indeed, our own motivation for raising these questions stems from syntactically annotating our own learner corpus (Ragheb, 2014; Ragheb and Dickinson, 2014; Dickinson and Ragheb, 2013) and realizing we needed clarification of certain categories, in particular those dealing with missing elements.

We examine phenomena surrounding ellipsis and enumeration because they are the main ones in our annotation scheme that license missing material, and missing material is important to investigate in the context of learner language, as learners often omit structures, e.g., determiners (see (Ragheb, 2014), ch. 7, and references therein). One other category could potentially be confused with categories licensing missing material, namely serial verb (SRL), which licenses a sequence of two verbs without a connector (similar to enumeration). In *come hang with us*, for example, *hang* is a SRL dependent of *come*. We ignore this category because: a) it is restricted to *come* and *go*; b) what we say about distinguishing coordination from enumeration (section 5) can more or less be applied to SRL; and c) we have not noticed it specifically causing confusion.

2 Linguistic Annotation for Learner Data

As argued in (Ragheb and Dickinson, 2011), one way to approach the annotation of learner corpora is by annotating linguistic properties. A starting assumption is that the categories used for learner language are similar enough to those for native language to use native categories. However, one quickly finds that linguistic categories for native speaker data are inadequate to represent the full range of learner productions (Díaz-Negrillo et al., 2010). For example, in (1),[1] the word *he* cannot simply be marked as a nominative or accusative pronoun because in some sense it is both. Thus, one may want to annotate multiple layers, in this case one POS layer for morphological evidence and one for syntactic distributional evidence (i.e., position).

(1) I must play with **he**.

While errors (i.e., ungrammaticalities) can be derived from mismatches between annotation layers, they are not primary entities. The multi-layer linguistic annotation is primarily based on linguistic evidence, not a sentence's correctness.

There are two main wrinkles to separating linguistic annotation from error annotation, however:

[1] Example sentences in this paper come from the SALLE corpus, comprised of essays from an Intensive English Program.

1) annotation categories could employ a notion of grammatical correctness to define; and 2) the decision process for ambiguous cases could reference a sentence's correctness. In the former case, the issue often has to do with using categories that are not always clearly defined for native data, while in the latter case, the issue is in having categories which—even if well-defined on different annotation layers—are insufficient to handle the usage the learner presents. In the next few sections we discuss issues surrounding non-canonical annotation categories and discuss the effect of the decision procedure in section 6.

To make the issues concrete, we rely on the syntactic annotation of the SALLE (Syntactically Annotating Learner Language of English) project (Ragheb, 2014; Ragheb and Dickinson, 2014), which employs multi-layer annotation. The issues are not specific to this annotation, but it illustrates the difficulties in applying native categories to learner data. That is, the SALLE annotation scheme (Dickinson and Ragheb, 2013) helps define questions of what constitutes appropriate linguistic annotation for interlanguage.[2]

3 Grammatical Annotation

When annotating learner data, it is important to know what is meant by *grammatical*. For error annotation, for example, this defines what an error is; e.g., in Korean, a missing postpositional particle may be an error or not depending on the level of formality underpinning grammaticality (Lee et al., 2012). The SALLE framework assumes a grammar based on the target language as an underpinning to the annotation (section 3.1), but, in the face of innovative learner usage, has focused on annotating the language as it appears and not on whether each sentence deviates from that grammar, i.e., is ungrammatical or not (section 3.2).

3.1 Target language grammar

To see the need to make clear the source of grammaticality, consider morphological POS annotation (section 2). In a verbal sequence like *can promotes*, for example, *promotes* intuitively has the morphological evidence of a third person singular verb. But to reference these morphological properties requires some notion of how these properties are defined, e.g., how *-s* stands for third person singular.

One obvious source of information is that "third person singular" comes from the definition of the *-s* morpheme in English. To annotate this way means referencing grammatical concepts from the target language (L2). If a different grammar is chosen to define categories, such as the learner's first language (L1), one might posit, e.g., *-et* as an indicator of "third person singular" (cf. Russian). In (Ragheb and Dickinson, 2012), we argue for using the L2 as the source of the grammar, as learners share many aspects of development in the L2 (Ellis, 2008) and as this can ensure annotation reliability.

3.2 Emerging categories

Annotation deals with the way facts from the grammar interact with phenomena occuring within a sentence. Consider objects, for example: a constellation of properties allows one to specify that two different sentences both contain them. Objects can be defined as: a) occurring, roughly speaking, after a verb (**syntactic distribution**); b) fitting into the argument structure of a verb, typically as a patient/theme (**semantic distribution**); and c) taking accusative case, as appropriate, e.g., *him* (**morphological distribution**). The class of objects emerges from these same patterns occurring across sentences within, in this case, English, and the task of annotation is to see whether a new instance fits into this class.

A distinction between categories—e.g., subjects and objects—arises from them having different sets of (typical) properties. With learner phenomena, there appear to be *new* kinds of emergent categories, ones which may overlap with previously-defined categories. When this happens, one has to specify which of the two categories a particular language instance falls into, and one way may be to say, "Category X is grammatical/native-like; category Y is not." It such cases we cover in the next two sections.

Before examining non-canonical categories, though, consider objects as they relate to the the usage of, for instance, *one* in (2). Does *one* fit the (target) category of object (OBJ), some other target category, or something else entirely?

(2) When I was in my country , I dreamed **one** I

[2] Guidelines at: `http://cl.indiana.edu/~salle/`

can go to a typical American city .

One possible approach (Reznicek et al., 2013; Rehbein et al., 2012) is to say: 1) the usage of *one* is non-native; 2) a native-like target is *I dreamed that one day I could go ...*; and 3) the grammatical annotation of *one* can thus be based upon this target form (e.g., as a type of temporal adjunct of *can go*).

The approach used in SALLE, by contrast, assumes that, after splitting out the lingusitic evidence into different layers (section 2), many learner innovations should be able to fit into an existing target category. In this case, the **morphosyntactic dependency** annotation layer ignores the semantic definition of OBJ and focuses on the fact that *one* occurs as a post-verbal nominal and is consistent with being accusative case. Thus, it can be annotated as conveying the evidence of the target category OBJ.

The point here is that this style of annotation employs definitions from a target grammar, in lieu of creating learner-specific categories or creating target forms that make clear a discrepancy between non-native and native categories, i.e., which deem a sentence ungrammatical. For canonical categories, individual learner instances can be difficult to categorize, but the categories themselves are, generally speaking, relatively well-defined.

4 Ellipsis and Missing Heads

4.1 Ellipsis

Ellipsis concerns omitted material in a sentence. In SALLE, an ELL label marks the relation between two categories that normally would not have a relation, but nonetheless do because of missing material. This ELL label collapses several elliptical relations in the CHILDES project (MacWhinney, 2000) (sec. 12.2), where pairs of labels denote the chain of dependencies that, in a sense, should be present between the two words (e.g., DET-OBJ). ELL is used when no other relation is possible and the dependent relation is not possible to specify locally, i.e., without crossing branches. An example is given in (3).

(3) I am a graduated **Biologist** actually an **Ecologist** .

Here, *Ecologist* restates *Biologist* as an appositive; the adverb *actually*, however, is a verbal modifier. To indicate an elliptical structure (cf. *actually* [*I am*] *an Ecologist*), *actually* is annotated as an ELL dependent of *Ecologist*, as in figure 1. The word missing its head takes the ELL label (*actually*) and attaches to the head of the construction (*Ecologist*).[3]

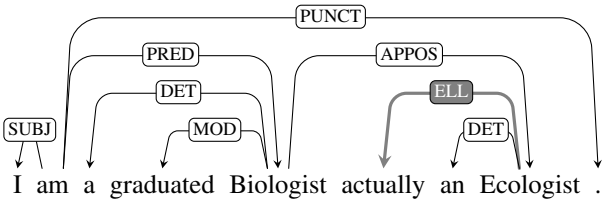


Figure 1: Appositive with an elliptical modifier

4.2 Missing heads

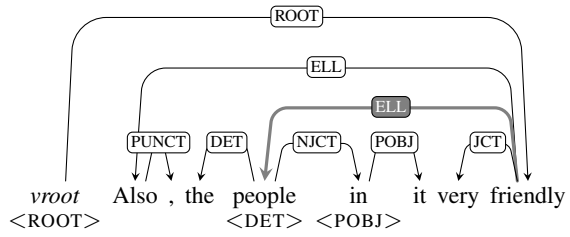
There are other cases of missing heads which are more clearly ungrammatical. One common case for learners concerns the omission of a finite verb in a sentence, as in (4). An analysis which continues the usage of the ELL label would annotate it as in figure 2(a), where the label mitigates the relation between *people* (what would be the subject if *are* were present) and *friendly* (what would be the predicate). Also shown here is a **subcategorization** layer, indicating which arguments each word is selecting for.

(4) Also , the people in it very friendly .

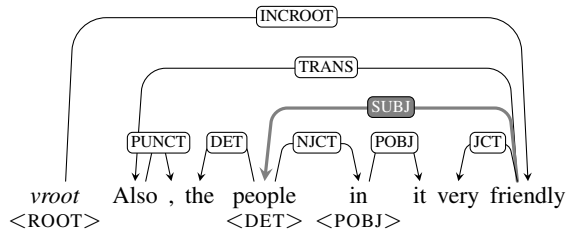
There is something satisfying and dissatisfying about the analysis. On the one hand, it stays in line with the annotation scheme by not marking anything peculiar. On the other hand, it poses two problems: 1) given the general side effect of mismatches between annotation layers when something is ungrammatical, one expects there to be a mismatch here, yet there is not; and 2) given the goal to annotate based on the evidence at hand, one would hope to provide a more informative label than ELL when possible. For example, *people* is a SUBJ of *friendly*, at least in some semantic sense.

Unlike the cases of ellipsis in section 4.1, there is no head recoverable from the context; i.e., unlike in (3) where *am* is present but just non-local, we do not have *are* anywhere in the context. The evidence

[3]Note that the ELL label only concerns missing heads, whereas the term *ellipsis* is generally used more broadly (e.g., (Sag, 1976)); missing dependents are handled differently, as discussed in (Dickinson and Ragheb, 2013) (sec. 5.1.2).



(a) ELL analysis of missing copula



(b) Missing head analysis of missing copula

Figure 2: Example of a missing copula

for this particular case is thus qualitatively different than in the more traditional elliptical cases—and so one may want to treat such cases differently.

There is additional reason for a separate missing head analysis: for some sentences, it is almost unavoidable to posit a missing head. Consider (5), where a purpose clause lacks the infinitive marker *to*. The construction *in order to* is more of a fixed form, and it is clear that a particular function word is missing. While ellipsis is governed by some principles (syntactic or otherwise) (e.g., (Sag, 1976; Goldberg, 2005; Culicover and Jackendoff, 2005)), learners can freely omit heads (and dependents) of various kinds—content or function words, fixed forms or open-ended constructions, etc.—and learner language annotation thus seems to need a separate treatment of missing heads.

(5) . . . I need more natural and friendly place to live with my wife **in order understand** each values and natures . . .

The treatment of (4) in SALLE is shown in figure 2(b). Here, *people* is the SUBJ of *friendly*; unlike ELL, SUBJ is an argument label, meaning it should be subcategorized for, but here it is not (indicated by having no <SUBJ>). Thus, there is a mismatch in annotation, and an informative, evidence-based label (SUBJ) being used. However, the sentence is treated differently than some other cases with missing heads, namely ones deemed elliptical.

4.3 Ellipsis vs. missing head

The details of each particular analysis are less important than noting the decision to make: should ellipsis annotation extend to non-native missing head constructions? There is evidence suggesting that at least some types of these cases are different (e.g., non-local presence/absence of the locally missing head) and thus the ellipsis category may no longer apply.[4] Additionally, there is an open question as to whether one wishes to refer to elliptical constructions as grammatical and missing heads as ungrammatical in determining the distinction.

5 Coordination and Enumeration

Coordination and enumeration feature a similar dichotomy, potentially dependent upon a sentence's grammaticality when no conjunction is present.

5.1 Coordination

Coordination in SALLE is right-branching. In figure 3, for example, *knowledge* serves as the prepositional object (POBJ); *and* is the CCC dependent of *knowledge*; and *personality* is the final coordination (COORD) element. An MCOORD (modificatory coordination) label is used between non-final elements in coordinations of three or more elements. COORD is an argument label and is thus subcategorized for (<COORD>), whereas MCOORD is not.[5]

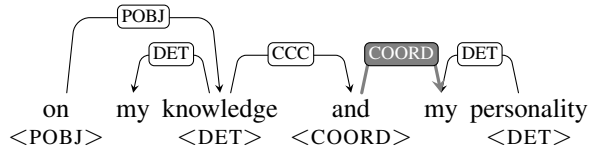


Figure 3: Treatment of basic coordination

[4] There are various other distinctions between figure 2(a) and figure 2(b), owing to other annotation scheme criteria, which we do not delve into here, i.e., ROOT vs. INCROOT, ELL vs. TRANS. See (Dickinson and Ragheb, 2013) for details.

[5] The right-branching analysis handles interactions with subcategorization for learner innovations; nothing hinges on this choice for the current paper, but for more details and argumentation, see (Dickinson and Ragheb, 2011).

5.2 Enumeration

SALLE also includes an enumeration label for lists of things. In line with coordination, they are treated as right-branching, with an ENUM label, as illustrated in figure 4. ENUM is not an argument label and thus does not need to be subcategorized for.

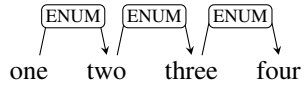


Figure 4: Treatment of an enumerated list (constructed)

This distinction is borrowed from the CHILDES annotation scheme (MacWhinney, 2000; Sagae et al., 2010), but the exact defintion of enumeration is difficult to pin down. Its prototypical properties include not needing a conjunction and often implying a continuation. Otherwise, the semantics are similar to coordination: multiple items are functioning in a parallel fashion. Further, some coordinations in some languages allow for no conjunction (Mithun, 1988), and enumeration might be considered a form of degenerate coordination (Wälchli, 2005).

5.3 Missing conjunction

The question of determining what enumeration refers to has a strong bearing on learner language, where there are constructions which could be either characterized as enumerations or as coordinations without a conjunction. Consider (6), with two separate sequences to consider. Focusing on the sequence of *is*es, there may be something amiss in being able to link them without a conjunction (in addition to the anomalous connection between the noun *Santiago* and the following three adjectives).

(6) I am Chilean , my hometown is Santiago , is beautiful , is big , is nice .

A partial dependency tree for the missing conjunction analysis in SALLE is given in figure 5. The analysis here is to use a COORD relation that is not subcategorized for as the final dependency, thus creating a mismatch indicating ungrammaticality.

It is hard to pinpoint exactly when a missing conjunction analysis should be utilized, and in this case part of the motivation has to do with capturing a formal written register of English. Additionally, garden-variety run-on sentences could be analyzed as missing conjunctions—as the connection between the main clauses in (6) could be. Furthermore, there are sentences where the units being combined are non-parallel, as in the link between *readings* and *swim and running* in (7), again opening the door for a possible missing conjunction analysis.

(7) Besides , I like swim and running , readings

It should be noted that there is also an option of treating the construction as involving two distinct elements with the same function; for example, in *my these tasks*, *tasks* could have two separate determiners. This option can complicate annotation, but does not change the question of how to separate coordination from enumeration, and so we set it aside here.

5.4 Enumeration vs. missing conjunction

Again, the pertinent question is: should enumeration annotation extend to non-native missing conjunctions? As pointed out, there is some evidence suggesting that they are different constructions, and as with missing heads and ellipses (section 4.3), missing conjunction coordinations can thus be defined as not being enumerations. For example, to be an enumeration might mean that no conjunction is required by the context and can be indicated with evidence such as an *etc*, as in (8).

(8) and i sing in church , street , station etc .

Again, an open question is whether one wishes to explicitly reference grammaticality (see, e.g., (Dickinson and Ragheb, 2013), p. 71). Note that such questions could arise for native language annotation, but the greater variability in learner forms exacerbates the problem: a string of items in sequence does not now necessarily mean it is an enumerated list.

6 Annotation Decision Procedure

Learner language can be multi-ways ambiguous—especially when categories license missing material—so annotation needs to provide multiple analyses (Reznicek et al., 2012; Lüdeling et al., 2005), provide enough contextual (meta-data) information to sort through analyses (Ott et al., 2012), and/or have a clear decision procedure for annotation. Due to having minimal meta-data

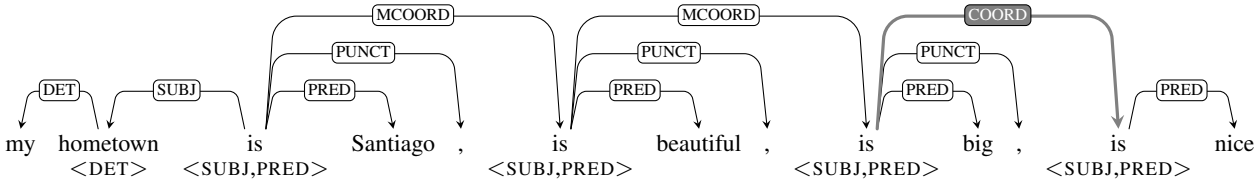


Figure 5: Missing conjunction (secondary SUBJs not shown)

and a small number of annotators, the SALLE project focuses on this last point. The annotation scheme is in some sense independent of the decision procedure involved in assigning the annotation—but the procedure itself could employ a notion of grammaticality in choosing a best analysis.

As mentioned in section 2, the issue here is in having L2 categories that are too specific to handle the usage the learner presents, i.e., no categories fit the usage. For example, in (9), the usage of *what* is not really a (question) determiner (DDQ) and the form is not that of a (subordinating) conjunction (CST).

(9) So when I admit to korea university , I decide **what** i find my own way .

There are a number of possible analyses for handling *what i find my own way*, including:

1. *what* as an extraneous word with no clear function and with a missing auxiliary (e.g., *would*);
2. *what* as a type of infinitival marker, with *i* as an extraneous word; or
3. *what* as a complementizer, albeit lexically anomalous, with the clause as valid (if odd).

A main SALLE heuristic is to "give the learner the benefit of the doubt." This heuristic favors analyses with fewer *mismatches*, i.e., discrepancies between different annotation layers, when no other evidence can distinguish the analyses. In this case, the third analysis is chosen because the lexical anomaly is the only indication of a learner-specific innovation.

Giving the learner the benefit of the doubt stems from treating the learner's language as a system in its own right (section 7.2) and does reduce the ambiguity for annotation. However, to give the benefit of the doubt—in lieu of other evidence—means annotators are arguably aware of how good or bad a sentence is, as they use a lack of errors as a guide. This is still qualitatively different than using explicit target hypotheses—as it is in terms of categories—but the degree to which this procedure references sentence correctness is a question that deserves closer investigation in the future. As mentioned, alternatives are to include more trees or more meta-information to disambiguate, each of which has its own costs.

7 Implications

We have seen non-canonical categories that license missing material (ellipses and enumerations), distinctions which could involve an annotator determining whether a sentence is grammatical when deciding between analyses. The decision procedure to obtain a single annotation may also reference grammaticality. The investigation in this paper and one's particular choices in practice have implications for both annotation practice (section 7.1) and second language analysis (section 7.2).

7.1 Impact on annotation

There are several takeaway points here for annotation of native or non-native data. First, these non-canonical categories seem to require one to consider to what extent annotation labels are merely indices and to what extent they reflect some grammatical properties worth capturing; that is, is there truly a grammar underlying the annotation? One must also consider the effect of annotation heuristics on the definitions in the grammar.

Secondly, when faced with non-canonical data and potentially a new set of competing analyses, one must choose how to apply the non-canonical categories. The main options seem to be the following:

1. Apply the native categories even to learner innovations, thereby extending the original definitions of the categories and making sentences potentially more ambiguous. For example, an

ellipsis category may license nearly any connection between two words.

2. Develop tighter restrictions on the native categories, so that differences in native and non-native instances emerge naturally. For example, ellipsis might be licensed only when the elided words can be literally recovered from the previous context. It should be noted that, in the general case, this option may only be available for data with enough meta-data to consistently distinguish the categories.
3. Reference sentence-level grammaticality in the definitions of categories. In essence, solution #3 is a subtype of solution #2, where the tighter restriction references grammaticality.

We have shied away from #1 because: a) it allows for too many possible analyses, and b) it treats the learner innovations exactly on a par with constructions that seem different. But note that this option seems to be consistent with the annotation practice of extending grammatical categories to new constructions (cf. (Pustejovsky and Stubbs, 2013), ch. 4)), while options #2 and #3 seem to be more in line with treating the underlying grammar as generative, i.e., as defining the set of allowable sentences in a language (cf. work back to (Chomsky, 1965)).

In this light, option #3 could have an unusual interpretation: as we understand it, to say that a missing head is not ellipsis *because it is ungrammatical* is to say that it is not in the target grammar (as ellipsis) because it is not in the grammar. Defining a category in terms of grammaticality may thus be a useful diagnostic for annotation practice, but further work should tease apart how principled this is. In general, being able to properly define a target category so that cases clearly do or do not fit (cf. sections 4.3 and 5.4), i.e., continuing to be evidence-based, seems to be worth pursuing. Option #3 also impacts acquisition research, a point we turn to next.

7.2 Impact on the comparative fallacy

The comparative fallacy in SLA is the notion that a researcher may be over-comparing a learner's interlanguage to the L2, and in that way treating the interlanguage as a corrupt form of the L2 (Bley-Vroman, 1983). (Ragheb and Dickinson, 2011) argue that linguistic annotation avoids the comparative fallacy in a way that error annotation doesn't, but relying on sentence-level grammaticality judgments would make that picture more muddled.

Without delving too deeply into the issue here (including how much one should want to avoid the comparative fallacy), our discussion of non-canonical categories implies that, at least for annotation, the comparative fallacy is not a simple binary distinction. Stemming from section 3, there is a distinction between analyzing target forms and target categories to consider in discussions of comparison, as well as a question of analyzing emerging constructions by making some reference to the correctness of a sentence, irrespective of a specific target. Non-canonical categories such as ellipsis seem to force an investigation into these issues; perhaps not coincidentally, these structures have often been relegated to peripheral phenomena in the theoretical literature (Culicover and Jackendoff, 2005).

8 Outlook

By applying categories appropriate for native language to learner language, we have discovered non-canonical categories that are difficult to apply. Further annotation for English and other languages will likely reveal other nuances, perhaps for distinctions generally difficult for dependency grammar, e.g., relative clauses. An immediate next step is to study categories which license extra arguments, such as topics and appositives.

Learner-specific annotation, such as underspecified categories, may also prove to impact how one sees non-canonical data. In that light, we have only scratched the surface of the implications for second language research, and we have not begun to examine other kinds of non-canonical data (e.g., dialectal). Additionally, one would like to know which categories are indeed useful for acquisition research, and studies utilizing this and other annotation schemes should shed light on this question (Ragheb, 2014; Alexopoulou et al., to appear).

Acknowledgments

We would like to thank Detmar Meurers, Heike Zinsmeister, and James Pustejovsky for discussion surrounding these topics, as well as the three anonymous reviewers for useful comments.

References

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

Theodora Alexopoulou, Jeroen Geertzen, Anna Korhonen, and Detmar Meurers. to appear. Exploring large educational learner corpora for sla research: perspectives on relative clauses. *International Journal of Learner Corpus Research*, 1(1):96–129.

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

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambrdige, MA.

Peter W. Culicover and Ray Jackendoff. 2005. *Simpler Syntax*. Oxford University Press, Oxford.

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

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. 2011. Dependency annotation of coordination for learner language. In *Proceedings of the International Conference on Dependency Linguistics (Depling 2011)*, pages 135–144, Barcelona, Spain.

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

Rod Ellis. 1985. Sources of variability in interlanguage. *Applied Linguistics*, 6(2):118–131.

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

Lotus Goldberg. 2005. *Verb-Stranding VP Ellipsis: A Cross-Linguistic Study*. Ph.D. thesis, McGill University.

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

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, October. Association for Computational Linguistics.

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

Sun-Hee Lee, Markus Dickinson, and Ross Israel. 2012. Developing learner corpus annotation for Korean particle errors. In *Proceedings of the Sixth Linguistic Annotation Workshop*, LAW VI '12, pages 129–133, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition. Electronic Edition, updated April 25, 2012, Part 2: the CLAN Programs: `http://childes.psy.cmu.edu/manuals/CLAN.pdf`.

Marianne Mithun. 1988. The grammaticization of coordination. In John Haiman and Sandra A. Thompson, editors, *Clause Combining in Grammar and Discourse*, volume 18 of *Typological Studies in Language*, pages 331–359. John Benjamins.

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

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. 2004. Lingo redwoods: A rich and dynamic treebank for hpsg. *Research on Language and Computation*, 2(4):575–596.

Niels Ott, Ramon Ziai, and Detmar Meurers. 2012. Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Thomas Schmidt and Kai Wrner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam.

James Pustejovsky and Amber Stubbs. 2013. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc., Sebastopol, CA.

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, Somerville, MA. Cascadilla Proceedings Project.

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*, pages 965–974, Mumbai, India.

Marwa Ragheb and Markus Dickinson. 2014. Developing a corpus of syntactically-annotated learner language for English. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13), Poster Session*, Tübingen, Germany.

Marwa Ragheb. 2014. *Building a Syntactically-Annotated Corpus of Learner English*. Ph.D. thesis, Indiana University, Bloomington, IN, August.

Owen Rambow. 2010. The simple truth about dependency and phrase structure representations: An opinion piece. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 337–340, Los Angeles, CA, June.

Ines Rehbein, Hagen Hirschmann, Anke Lüdeling, and Marc Reznicek. 2012. Better tags give better trees - or do they? *Linguistic Issues in Language Technology (LiLT)*, 7(10).

Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas, 2012. *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01*. Humboldt-Universität zu Berlin, Berlin.

Mark Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, editors, *Automatic Treatment and Analysis of Learner Corpus Data*, pages 101–123. John Benjamins, Amsterdam.

Ivan A. Sag. 1976. *Deletion and logical form*. Ph.D. thesis, Massachusetts Institute of Technology.

Kenji Sagae, Eric Davis, Alon Lavie, and Brian MacWhinney an Shuly Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3):705–729.

Larry Selinker. 1972. Interlanguage. *International Review of Applied Linguistics*, 10(3):209–231.

Kari Tenfjord, Jon Erik Hagen, and 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.

Bernhard Wälchli. 2005. *Co-Compounds and Natural Coordination*. Oxford Studies in Typology and Linguistic Theory. Oxford University Press, Oxford.