

Dependency Annotation of Coordination for Learner Language

Markus Dickinson
Indiana University
md7@indiana.edu

Marwa Ragheb
Indiana University
mragheb@indiana.edu

Abstract

We present a strategy for dependency annotation of corpora of second language learners, dividing the annotation into different layers and separating linguistic constraints from realizations. Specifically, subcategorization information is required to compare to the annotation of realized dependencies. Building from this, we outline dependency annotation for coordinate structures, detailing a number of constructions such as right node raising and the coordination of unlikes. We conclude that branching structures are preferable to treating the conjunct as the head, as this avoids duplicating annotation.

1 Introduction and Motivation

While corpora containing the language of second language learners have often been annotated for errors (e.g., Nicholls, 2003; Rozovskaya and Roth, 2010), they have rarely been annotated for linguistic properties. Those which mark part-of-speech (POS) tend to do so only for illicit forms (e.g., Granger, 2003) and those with syntactic annotation generally first map the learner forms to target forms (e.g., Hirschmann et al., 2010). While these annotations serve many purposes, what has been lacking is linguistic annotation of the learner data itself, in particular syntactic annotation (Dickinson and Ragheb, 2009). As argued in Ragheb and Dickinson (to appear), such annotation has the potential to be beneficial for much second language acquisition (SLA) research, to address questions such as complexity (e.g., Pendar and Chapelle, 2008) and stage of acquisition (e.g., Pienemann, 1998). Such annotation is also suited to evaluate the parsing of learner data (Ott and Ziai, 2010).

We outline an annotation framework for applying syntactic dependency annotation to learner

corpora, focusing on the challenges stemming from coordination for learner structures. The first issue in annotating dependencies for learner language has to do with the fact that learner data diverges from canonical language use. We build from proposals which thus split the annotation into separate levels, one for each piece of evidence. In (1), from (Díaz Negrillo et al., 2010), the word *jobs* is distributionally in a singular noun slot, but has the English plural marker. Díaz Negrillo et al. propose separate layers of part-of-speech (POS) annotation to account for this (see section 2).

- (1) ... for almost every **jobs** nowadays ...

Splitting annotation into different layers for different types of linguistic evidence is applicable to dependency annotation (Dickinson and Ragheb, 2009), but as we will describe in section 3, there is also a need to separate linguistic constraints from the actual realizations, in order to capture non-native properties. Subcategorization requirements, for example, do not always match what is realized.

Coordination is one particularly difficult area for dependency annotation (e.g., Nivre, 2005). When linguistic constraints are separated from realizations, coordination becomes a prominent issue for learner annotation, as the constraints (subcategorization) and the realizations (dependencies) need to be appropriately matched up. Our annotation scheme should: 1) be useful for SLA research (Ragheb and Dickinson, to appear), 2) be as simple as possible to annotate, and 3) cover any learner sentence, regardless of the proficiency level. Balancing these concerns and taking our multi-layered approach to annotation into account (sections 2 and 3), we will advocate a branching approach to coordination in section 4. Such an approach treats every dependency independently, avoiding the duplication of information.

2 Annotating learner language

There has been a recent trend in annotating the grammatical properties of learner language, independent of errors (Díaz Negrillo et al., 2010; Dickinson and Ragheb, 2009; Rastelli, 2009). While error annotation has been the standard annotation in learner corpora (e.g., Granger, 2003; Díaz Negrillo and Fernández Domínguez, 2006), annotation of linguistic properties such as POS and syntax provides SLA researchers direct indices to categories of interest for studying interlanguage (Pienemann, 1992; Ragheb and Dickinson, to appear). One does not posit a correct version of a sentence, but annotates only what is observed.

Consider again example (1): a single POS is not appropriate, as the distributional evidence for *jobs* is of a singular noun, and the morphological evidence is plural. Díaz Negrillo et al. (2010) propose annotating 3 tags, representing the morphological, distributional, and lexical evidence. Each POS layer, then, contains a separate description of a linguistic property. The POS is not claimed to be a single category; rather, the evidence is represented in different layers, thereby providing access for searching. Errors in this framework are epiphenomena, arising from conflicts between layers.

Using SUSANNE tags (Sampson, 1995), we see an example of two layers in (2), where the distributional layer contains a present tense verb (VVZt) and the morphological layer a base form verb (VV0t).¹ In a sense, this parallels the multi-layered annotation in Lüdeling et al. (2005), where each error interpretation is given its own layer.

(2) Tin Toy can **makes** different music ...
NP1x NP1x VMo **VVZt** JJ NN1u ...
NP1x NP1x VMo **VV0t** JJ JJ ...

These annotation efforts are still in the early stages of development, making the conceptual issues clear. Because much SLA research is framed in terms of linguistic categories—e.g., the use of extraction from embedded clauses (e.g., Juffs, 2005; Wolfe-Quintero, 1992)—the annotation has much potential to be useful. We turn next to annotating dependencies in this framework.

3 Dependencies for learner language

We will provide a sketch of the annotation layers we use, emphasizing the split between the anno-

¹Unless otherwise noted, our learner examples come from a corpus of narratives from the 1990s (Bardovi-Harlig, 1999).

tation of realized dependencies (section 3.2) and subcategorization (section 3.3).

3.1 Completeness, Coherence, & Consistency

Leaving aside the separation of linguistic evidence for the moment, we start with the general use of dependencies, which directly capture selection and modification relations. We focus on capturing selectional properties, which means dealing with issues of: 1) completeness, 2) coherence, and 3) consistency (cf. Lexical-Functional Grammar (LFG), Bresnan, 2001). Violations of these are given in the constructed examples in (3). Example (3a) represents an incomplete structure, in that the verb *devour* selects for an object, which is not realized. For *completeness* to hold, all the arguments of a predicate must be realized.

- (3) a. *Max devoured.
b. *Max slept a tree.
c. *Max devoured of a sandwich.

In (3b), there is an incoherent structure, as there is an extra argument: for *coherence*, there must be no additional arguments. Finally, (3c) is inconsistent, as there is a prepositional phrase, but *devoured* selects a noun phrase. To be *consistent*, the realized arguments must match those selected for. Since learners produce structures with a mismatch between the selectional requirements and the realized arguments, we want to represent both.

3.2 Modeling dependencies

3.2.1 Distributional dependencies

We first annotate the relations occurring in the sentence, using the target language (English) as a reference frame to define the relations, e.g., what it means to be a subject. By **distributional dependencies**, we refer to dependencies between words based strictly on syntactic distribution, i.e., primarily word order. Building from Dickinson and Ragheb (2009), we focus on these dependencies; other layers are discussed in section 3.2.3.

In (4), for example, *baby* is in the distributional slot of the subject of *had*, as defined by English declarative structure.

- (4) The **baby** had no more interest ...

To see the need for defining dependencies on a strictly syntactic basis, consider (5). The word *dull* (cf. *doll*) is ambiguous: it could be an object of *escape* (with a missing subject), or it could be

the subject in the wrong location. To fully disambiguate requires knowing learner intention, a difficult proposition for consistent and reliable annotation. Looking only at distribution, however, this position in English is an object position.

(5) After the baby down, escape the **dull**.

The tree for this example is shown in figure 1, where *dull* is the object (OBJ). The non-nativeness of this sentence is captured via the encoding of subcategorization requirements (section 3.3).

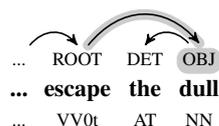


Figure 1: Distributionally-based dependencies, with distributional POS tags

We use the CHILDES annotation scheme (Sagae et al., 2010, 2007) as the basis for our annotation, as it was developed for language being acquired (albeit, first language), with two main differences: 1) They treat main verbs as heads, with auxiliaries and infinitive markers (*to*) as dependents, whereas we mark auxiliaries as heads, following work treating them on a par with raising verbs (e.g., Pollard and Sag, 1994). 2) They treat the conjunct in coordinate structures as the head, whereas we investigate this approach and a binary-branching approach, ultimately arguing for branching. For branching, we introduce a new label, CC (coordinating conjunction), for the relation with the conjunction as a dependent.

3.2.2 Secondary dependencies

Given the widely-held assumption that each word has only one head in a dependency graph (Kübler et al., 2009, ch. 2), basic dependencies cannot capture every relationship. In the learner example (6), for instance, *I* is the subject for the verbs *hope* and *do*. Allowing for additional dependencies to be specified (cf. Kromann, 2003; Sgall et al., 2004), this can be fully represented.

(6) ... the only thing that **I** hope to do ...

We thus annotate **secondary dependencies**, which encode non-local syntactic relationships between words. Such secondary dependencies are represented in figure 2 with arcs below the words. One could argue that secondary dependencies are

semantic; we try to restrict usage to cases where: a) a syntactic process is involved, in this case control, and b) the subcategorization of predicates is at stake (section 3.3). As we will see in section 4, secondary dependencies are crucial to capturing the selected dependents of coordinated functors.

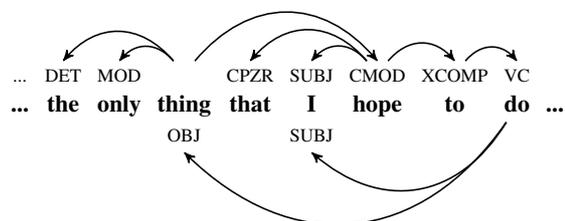


Figure 2: Encoding secondary dependencies

3.2.3 Other types of dependencies

We focus on distributional dependencies in this paper, as this is sufficient to illustrate the issues faced with coordination. Other types of dependencies can and should be annotated for learner language, including **morpho-syntactic** and **semantic dependencies**. Splitting dependencies into different layers of evidence has precedence in a variety of frameworks (e.g., Mel'čuk, 1988; Debusmann et al., 2004; Deulofeu et al., 2010).

For morpho-syntactic dependencies, consider the constructed example (7): *Him* is in the subject distributional position, but morphologically has object marking. The interplay between morphological and distributional layers will vary for different language types (e.g., freer word order).

(7) **Him** slept.

Semantic dependencies would capture the canonical linking of dependencies to meaning (e.g., Ott and Ziai, 2010; Hirschmann et al., 2010). Consider *see* in (8). The distributional position of the subject is filled by *Most (of the movie)*, while the object is *adults*, but on a semantic layer of dependencies, *adults* may be the subject and *Most* the object. Again, this is an orthogonal issue.

(8) **Most of the movie** is seem to see **adults**, but the childern like to movie.

3.3 Modeling subcategorization

Dependencies are based on evidence of what learners are doing, but to capture completeness, coherence, and consistency, we need to model

which dependencies are selected for, namely **subcategorization** information.

We annotate subcategorization frames on the basis of the requirements in the target language (English). For example, in (5), the subordinate clause is missing a verb. One way to capture this is in figure 3, where *baby* is the subject (SUBJ) of *down*, but *down* has an empty subcategorization list (<>). Since subjects are arguments, this mismatch indicates an issue with coherence. By contrast, *baby* subcategorizes for a determiner (<DET>), which is realized.

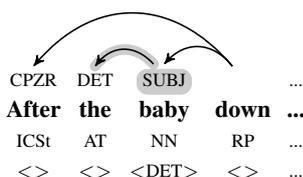


Figure 3: Partial tree with dependencies, distributional POS tags, and subcategorization frames

Words may have many subcategorization frames (Levin, 1993), and we annotate the one which is the best fit for a given sentence. In the constructed cases in (9), for example, *loaded* receives different annotations. In (9a), it is <SUBJ, OBJ>, while in both (9b) and (9c), it is <SUBJ, OBJ, IOBJ-*with*>. For (9c), this is the best fit; while still not matching what is in the sentence, it means that only one element (OBJ) is missing, as opposed to, e.g., <SUBJ, OBJ, IOBJ-*into*>, where two elements would be wrong.

- (9) a. Max **loaded** the wagon.
 b. Max **loaded** the wagon with hay.
 c. *Max **loaded** with hay.

Treatment of raising and control Consider (6) again: in *hope to do*, the subject of *do* is essentially the same as that of *hope*, and in many theories, *to* “raises” the subject, keeping relations local. We can see subcategorization information in figure 4.

It is not immediately clear whether we should explicitly annotate raising and put SUBJ on *to*’s subcategorization frame. We are trying to base the annotation on well-founded grammatical theory, but the primary criteria are: a) to make the data useful for SLA research, and b) to be able to annotate efficiently. Thus, even if a theoretical model supports the annotation, we do not necessarily need to annotate all parts of it.

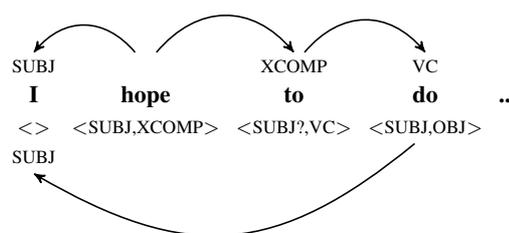


Figure 4: Treating raising and control

We advocate not annotating raising in all cases. This is simpler for annotation, especially as we get into the sharing of elements between conjuncts. We expect more efficient and reliable annotation by annotating the minimal required elements. Additionally, keeping subcategorization simple makes us less committed to any theoretical claims for, for example, right node raising (section 4.2). When coordinated verbs share an object, we do not have to determine whether the object is percolated up to the conjunction; there is simply a long-distance relationship where appropriate.

Technical details We encode our annotation by extending the CoNLL format (Buchholz and Marsi, 2006) to account for secondary dependencies (see details in Dickinson and Ragheb, 2009). We are also extending the format to encode both distributional and morpho-syntactic dependencies.

4 Our treatment of coordination

There are many ways to handle coordination in dependency annotation (see, e.g., Osborne, 2008, sec. 5), of which we will examine two main ones.² With our basic layers as defined above, we will show that a binary-branching analysis is preferable for annotating learner language, in that it minimizes the number of mismatches between subcategorization and realization.

4.1 Basic coordination

In the learner example (10), two arguments (of *about*) are conjoined. One treatment of this is with the conjunction as the head, as in figure 5,³ while an alternate view is to have a branching structure, as in figure 6.⁴ We will use these two treatments of coordination throughout, in order to illustrate what

²If one allows for limited amounts of constituency, there are even more ways to treat coordination (cf. Hudson, 1990).

³We often abbreviate: C=COORD, S=SUBJ, O=OBJ.

⁴Branching could go in either direction; while we choose right-branching, nothing hinges on this.

needs to be captured for learner language; these are also the main analyses considered for parsing (Kübler et al., 2009). The conjunction-as-head analysis treats coordination as involving some degree of a “phrase,” whereas right-branching treats the conjuncts independently.

(10) The story about a tin toy and a baby .

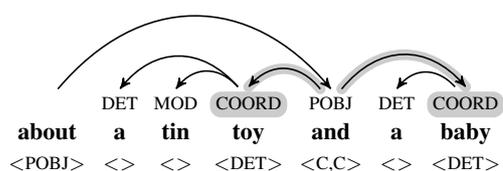


Figure 5: Conjunction-as-head coordination

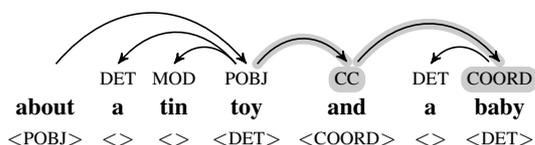


Figure 6: Right-branching coordination

For either analysis, we must consider how subcategorization interacts with the dependencies. In this case, it must be clear that *about*—which selects for a prepositional object (POBJ)—actually realizes it. Both analyses meet this requirement.

Additionally, we need to consider how subcategorization should be handled for the conjunction itself. A learner could potentially use a conjunction like *and* without one of its conjuncts. Thus, it should select for at least one coordinating element. In figure 5, this is done by *and* selecting for two COORD elements, while in figure 6, it selects for one element, as only one conjunct is realized at a time. The CC relation is not selected for, consistent with the fact that the head of *and* is not required to have a conjoined phrase.⁵

For the moment, we are simplifying the dependency graphs; in section 4.3, we will discuss the need to further articulate the COORD labels. In this case, we will have <COORD-POBJ> in the branching analysis, i.e., passing down the POBJ requirement from the head of *and* onto *and* itself.

⁵Another branching analysis has the conjunct be a dependent of the second noun (*baby*) (e.g., Buch-Kromann, 2009). While selection works differently, our general points about branching analyses should apply.

Saturated functors For the coordination of functors—i.e., words selecting for arguments—these can be treated on a par with basic argument coordination if they have realized all their requirements. Looking at the coordination of sentences in (11), for example, both *found* and *hid* are functors, but are saturated when they coordinate. Thus, the treatment of coordination is the same as before (trees not shown for space reasons).

(11) the tin toy **found** the very safety place where he should hide , and he **hid** under a sofar .

4.2 Coordination of unsaturated functors

Consider now the case where two unsaturated elements are coordinated, i.e., both words are still looking for an argument. In (12), for example, *walk* and *run* both have the same subject. The trees in figures 7 and 8 show that *He* is the subject of *begins*, with *walk* and *run* having a secondary connection to it. For this sentence, there is not a great difference between the two different analyses, in terms of connecting dependencies and subcategorizations. If the sentence were *He walks and runs*, however, then *and* would take *He* as a SUBJ for the conjunction-as-head analysis and thus also explicitly include SUBJ on its subcategorization; we take this issue up in the next section.

(12) He begins to walk and at to run .

As a side point, note in this example that *at* has an empty subcategorization list because we cannot determine what it is distributionally. For the morphologically-defined tree (see section 3.2.3), the subcategorization for *at* would be <POBJ> without a POBJ being realized.

Right node raising Moving from a fairly straightforward analysis of shared subjects, let us now consider the more challenging shared object between conjuncts, as in the constructed example (13), a case of right node raising (cf. Ross, 1967).⁶

(13) He begins to walk and to run the race.

Trees for this example are presented in figures 9 and 10. In both cases, the analyses are relatively theory-neutral, in that they do not state anything explicitly about how the object came to be shared between these verbs (see section 3.3).

⁶Most of the remaining examples in the paper are constructed, due to these types of coordination not having been observed in our data thus far.

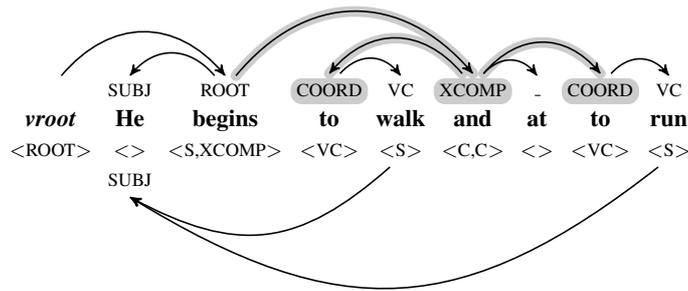


Figure 7: Functor coordination, where functors are unsaturated (conjunction-as-head)

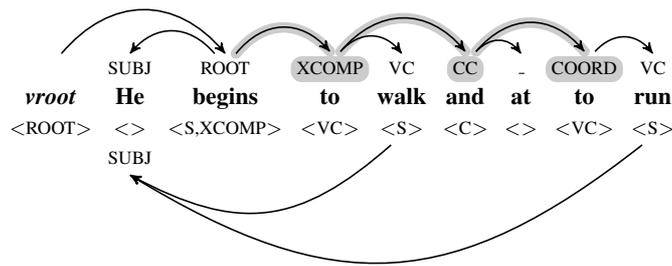


Figure 8: Functor coordination, where functors are unsaturated (right-branching)

What is noticeable in comparing the figures is the extra secondary dependency in the conjunction-as-head analysis. Recall that part of our goal is to accurately encode whether a learner’s sentence obeys completeness, coherence, and consistency. With *and* as the head of the coordinate structure, it must have the object as its dependent and must thus have the object on its subcategorization list. This means that all three words (*walk*, *and*, *run*) have the same object in their subcategorization.

Consider now if there were to be an error in consistency, as in the constructed example (14), where the verbs expect OBJ, but instead find the prepositional IOBJ. There are now 3 mismatches, as *bakes*, *eats*, and *and* all have the same OBJ subcategorization requirement. In general, the conjunction-as-head analysis reduplicates dependency requirements, leading to more mismatches.

(14) He bakes and eats **to** the cookies.

In the branching analysis in figure 10, on the other hand, only the verbs have the object requirement listed in their subcategorization, and the number of secondary dependencies is reduced from 4 to 3. To handle (14), there would be only two mismatches, one for each verb. As we argue below, this is desirable, as each verb can have its

own separate requirements.

Note that we are not claiming that the branching analysis is better theoretically. We are claiming that it is a simpler way to annotate learner language, especially as it posits fewer errors.

Functor coordination with different requirements

Consider an example of right node raising where there are slightly different verbal requirements. In the constructed example (15), for instance, *is fond of* selects for a prepositional object (POBJ), while *buys* selects for an object.

(15) She is fond of and buys toys.

In figures 11 and 12, this is partly handled by the (secondary) dependencies between *of* and *toys*, on the one hand, and between *buys* and *toys*, on the other. The relation is POBJ in the former cases, and OBJ in the latter. Whether primary or secondary, each relation has a unique label.

The issue is in the label between *and* and *toys* in the conjunction-as-head analysis (figure 11): should it be POBJ or OBJ? We can posit a category hierarchy (e.g., POBJ as a subtype of OBJ) or an intersection of categories (e.g., OBJ+POBJ), but this requires additional machinery. The branching analysis (figure 12) requires nothing extra, as no extra relations are used, only those between the

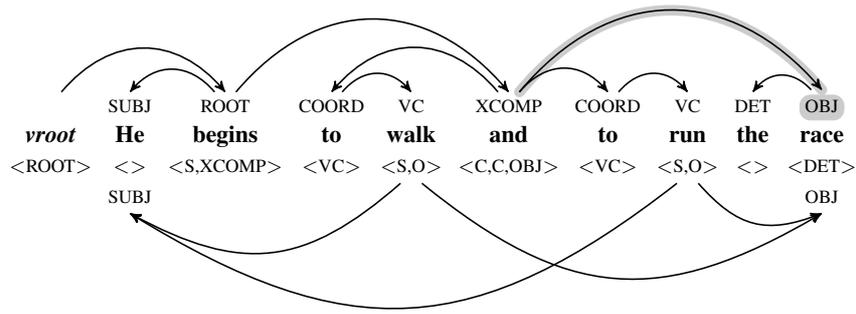


Figure 9: Functor coordination, with right node raising (conjunction-as-head)

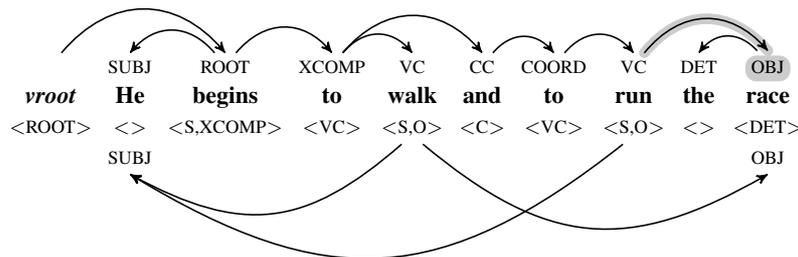


Figure 10: Functor coordination, with right node raising (right-branching)

functors and *to*s. This independent treatment of verbs also means that if verb saturation differs, the conjunction does not have to represent this, as in the learner example (16), where *run* is saturated and *stumbled over* is not (missing POBJ).

(16) ... it **run** after him and **stumbled over** and began to cry.

4.3 Coordination of unlikes

One difficulty that arises in annotating coordination is in how we annotate the coordination of unlike elements. Coordination of unlikes is well-known (Sag, 2003; Sag et al., 1985), though when we refer to the coordination of unlike elements, we are referring to elements which have different dependency relations. For instance, (17) features a coordination of an adjective and a noun phrase. But, in terms of their dependencies, they are both predicatives, so their dependency will be the same (PRED), as our dependency inventory does not distinguish adjectival from nominal predicatives.

(17) Pat is [wealthy and a Republican]. [AP & NP] (Sag et al., 1985)

The kind of case we are concerned about occurs in the constructed example (18), where we have a

non-finite and a finite verb conjoined.⁷ Because learners can head a sentence with a non-finite verb (e.g., *to apparer a baby*) or no verb at all (e.g., *the baby down* in (5)), we distinguish finite ROOT relations from non-finite ROOT-nf. In (18), then, we have one conjunct (*running*) which should be ROOT-nf and one (*eats*) which should be ROOT.

(18) He running and eats.

Walking through figures 13 and 14, we first consider the label on the arc between *and* and its head. For the conjunction-as-head analysis, we need to indicate that the whole *and* phrase is not consistent. This is essentially the same issue we saw with OBJ+POBJ; in this case, we need to annotate the label as ROOT+ROOT-nf or use a hierarchy. This makes the connection to the subcategorization list transparent: *vroot* looks for ROOT, but finds both ROOT and ROOT-nf. The branching structure, on the other hand, only takes the first conjunct is its dependent. Thus, if *running* comes first—as it does in figure 14—its label is ROOT-nf; if *eats* were first, the label would be ROOT.

⁷We have an attested example of unlike coordination in *I want to make happy and love and nice family*, but use the simpler (18) to explain our approach; the points are similar.

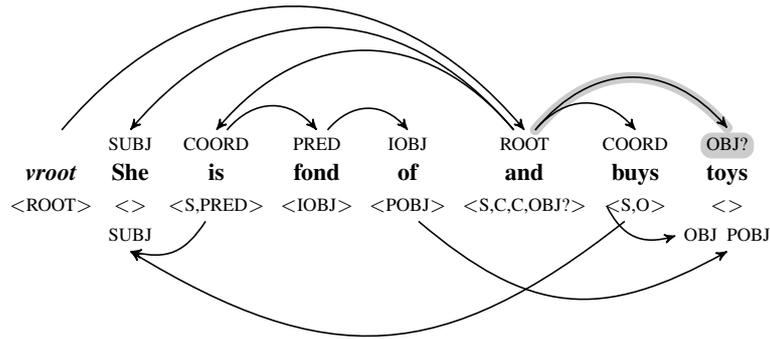


Figure 11: Coordination between two elements with different requirements (conjunction-as-head)

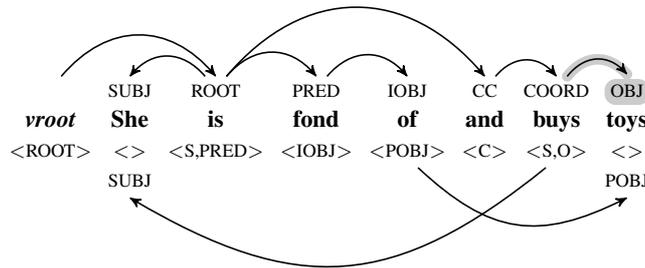


Figure 12: Coordination between two elements with different requirements (right-branching)

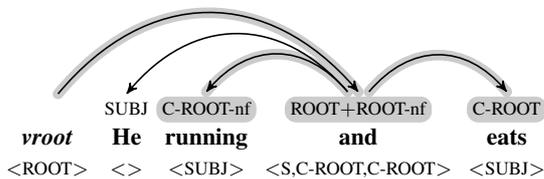


Figure 13: Coordination of unlikes; secondary dependencies not shown (conjunction-as-head)

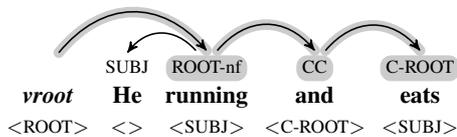


Figure 14: Coordination of unlikes; secondary dependencies not shown (right-branching)

Secondly, there is the relation between *and* and its dependents. To determine which conjunct is finite and which non-finite for the conjunction-as-head analysis and to exactly pinpoint the inconsistency, we augment the COORD labels. COORD only tells us that the element is a coordinating element, but does not tell us if the word is functioning

as a subject, a verbal complex, etc. Incorporating the actual relation, we create COORD-ROOT and COORD-ROOT-nf labels in this case.

For subcategorization, the requirements of the head of *and* (the virtual root *vroot*) are passed down to *and* and added to its conjunct requirements. Thus, in figure 13, *and* selects for two COORD-ROOT elements: COORD because it is a conjunction, and ROOT because its head selects for a ROOT. Thus, in the case of *running*, we identify a mismatch between the selected-for COORD-ROOT and the realized COORD-ROOT-nf.

For the branching analysis in figure 14, we also use COORD-ROOT. If the sentence were *He eats and running*, we would want to know that *and* selects for COORD-ROOT, but realizes COORD-ROOT-nf (*running*). Though not indicated in previous figures, this applies for all the trees in this paper, to ensure that requirements can be checked.

Again, the conjunction-as-head analysis is more complicated to annotate: in figure 13, there are two mismatches—between the subcategorization and realization for *vroot* and also for *and*—for what is only one issue. And unlike the use of ROOT+ROOT-nf, with the branching analysis

there is no confusion about the problem's source.

5 Summary and Outlook

We have outlined a way of annotating dependencies for learner language, relying upon a division of labor between basic dependencies, secondary dependencies to capture long-distance relations, and subcategorization marking for every word. Comparing two different exemplar analyses of coordination, we illustrated why a branching analysis is preferable over one which duplicates information, in terms of keeping annotation simple and allowing one to find mismatches between annotation layers. We are attempting to maintain a relatively simple annotation scheme, but as coordination illustrates, even this can become complex.

This treatment handles the cases of coordination we have observed so far, and in this paper we covered the main constructions we expect to see in learner language. A few other cases need to be fully borne out in the future, however, including cases of missing conjunctions and of non-constituent coordination (Steedman and Baldrige, 2011). For missing conjunctions, one would have to use a non-conjunction head, i.e., one of the conjuncts, in the conjunction-as-head analysis (e.g., Sagae et al., 2010, p. 716), while for the right-branching analysis, there has to be a direct link between conjuncts. This means a CC relation will not have a conjunction as its dependent. Working out the details requires a fuller treatment of modification, but neither case seems to supersede our proposal.

The annotation effort is still relatively new, and we are beginning to move out of the pilot phase. With the different layers in place, we are currently investigating inter-annotator agreement.

Acknowledgments

We thank Detmar Meurers for discussion and four anonymous reviewers for their helpful feedback.

References

Kathleen Bardovi-Harlig. 1999. Examining the role of text type in L2 tense-aspect research: Broadening our horizons. In *Proceedings of the Third Pacific Second Language Research Forum*, volume 1, pages 129–138. Tokyo.

Joan Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell Publishing, Oxford.

Matthias Buch-Kromann. 2009. *Discontinuous Grammar. A dependency-based model of human parsing and language learning*. VDM Verlag.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of CoNLL-X*, pages 149–164. New York City.

Ralph Debusmann, Denys Duchier, and Geert-Jan M. Kruijff. 2004. Extensible dependency grammar: A new methodology. In *Proceedings of the COLING 2004 Workshop on Recent Advances in Dependency Grammar*. Geneva/SUI.

José Deulofeu, Lucie Duffort, Kim Gerdes, Sylvain Kahane, and Paola Pietrandrea. 2010. Depends on what the french say - spoken corpus annotation with and beyond syntactic functions. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 274–281. Uppsala.

Ana Díaz Negrillo and Jesús Fernández Domínguez. 2006. Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada (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).

Markus Dickinson and Marwa Ragheb. 2009. Dependency annotation for learner corpora. In *Proceedings of the TLT-8*. Milan, Italy.

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

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

Richard A. Hudson. 1990. *English Word Grammar*. Blackwell, Oxford, UK.

Alan Juffs. 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.

Matthias Trautner Kromann. 2003. The danish dependency treebank and the underlying linguistic theory. In *Proceedings of TLT-03*. Växjö, Sweden.

- Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. Dependency parsing. In Graeme Hirst, editor, *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.
- Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.
- Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics*. Birmingham.
- Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.
- Diane Nicholls. 2003. The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003)*, pages 572–581. Lancaster University.
- Joakim Nivre. 2005. Dependency grammar and dependency parsing. MSI report 05133, Växjö University: School of Mathematics and Systems Engineering.
- Timothy Osborne. 2008. Major constituents and two dependency grammar constraints on sharing in coordination. *Linguistics*, 46(6):1109–1165.
- Niels Ott and Ramon Ziai. 2010. Evaluating dependency parsing performance on German learner language. In *Proceedings of TLT-9*, volume 9, pages 175–186.
- Nick Pendar and Carol Chapelle. 2008. Investigating the promise of learner corpora: Methodological issues. *CALICO Journal*, 25(2):189–206.
- Manfred Pienemann. 1992. Coala—a computational system for interlanguage analysis. *Second Language Research*, 8(1):58–92.
- Manfred Pienemann. 1998. *Language Processing and Second Language Development: Processability Theory*. John Benjamins, Amsterdam.
- Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press.
- Marwa Ragheb and Markus Dickinson. to appear. Avoiding the comparative fallacy in the annotation of learner corpora. In *Second Language Research Forum Conference Proceedings*. Cascadilla Proceedings Project, Somerville, MA.
- Stefano Rastelli. 2009. Learner corpora without error tagging. *Linguistik online*.
- John Robert Ross. 1967. *Constraints on Variables in Syntax*. Ph.D. thesis, MIT.
- Alla Rozovskaya and Dan Roth. 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, CA.
- Ivan Sag, Gerald Gazdar, Thomas Wasow, and Steven Weisler. 1985. Coordination and how to distinguish categories. *Natural Language and Linguistic Theory*, 3:117–171.
- Ivan A. Sag. 2003. Coordination and underspecification. In *Proceedings of the Ninth International Conference on HPSG*. CSLI Publications, Stanford.
- Kenji Sagae, Eric Davis, Alon Lavie, and Brian MacWhinney and Shuly Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3):705–729.
- Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 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.
- Geoffrey Sampson. 1995. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Clarendon Press, Oxford.
- Petr Sgall, Jarmila Panevová, and Eva Hajičová. 2004. Deep syntactic annotation: Tectogrammatical representation and beyond. In *Proceedings of the Workshop on Frontiers in Corpus Annotation*, pages 32–38. Boston.
- Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. In Robert Borsley and Kersti Borjars, editors, *Non-Transformational Syntax: Formal and Explicit Models of Grammar*. Wiley-Blackwell.
- Kate Wolfe-Quintero. 1992. Learnability and the acquisition of extraction in relative clauses and wh questions. *Studies in Second Language Acquisition*, 14:39–70.