Defining syntax for learner language annotation

Marwa RAGHEB[1] *Markus DICKINSON*[1]
(1) Indiana University, Bloomington, IN USA
mragheb@indiana.edu, md7@indiana.edu

ABSTRACT
We discuss making syntactic annotation for learner language more precise, by clarifying the properties which the layers of annotation refer to. Building from previous proposals which split linguistic annotation into multiple layers to capture non-canonical properties of learner language, we lay out the questions which must be asked for grammatical annotation and provide some answers. Our investigation points to the layer of distributional syntax being based on properties of the target language (L2) and largely redundant with the other layers. We show, for example, that subcategorization seems to better be able to underspecify annotation for situations where no single correct solution can be found. While this paves the way for applying the annotation to larger corpus efforts, it also represents a significant step in elucidating syntax for non-canonical language.

KEYWORDS: syntactic annotation, dependency syntax, learner language.

Proceedings of COLING 2012: Posters, pages 965–974,
COLING 2012, Mumbai, December 2012.

1 Introduction

Learner corpora are increasingly gaining attention due to the potential wealth of data they present for a variety of purposes, including the investigation of different aspects of *interlanguage*, the developing language of second language learners. Interlanguage (IL) often differs from the target language (L2), and the annotation of such corpora is an important means of accessing its unique characteristics (Granger, 2003). Such annotation has practical benefits for developing error detection systems and intelligent tutoring systems (e.g., Nagata et al., 2011; Rozovskaya and Roth, 2010), by providing data and insights for the parsing of learner language. While many benefits have been derived from error annotation, a recent approach to annotating learner language is to annotate the linguistic properties of a text to provide direct access to grammatical properties of interest (e.g., Díaz-Negrillo et al., 2010; Dickinson and Ragheb, 2009; Rastelli, 2009; Pienemann, 1992). To account for IL, multiple layers of analysis—e.g., three separate part-of-speech (POS) layers—have been proposed to capture learner innovations. These layers have yet to be properly defined for syntax, however. Our aim is to probe syntactic annotation for learner language, making precise what decisions annotation efforts must make.

Recent work on learner corpora has underscored the importance of providing a linguistic description of learner text for second language acquisition (SLA) (Ragheb and Dickinson, 2011). To see the need more broadly, consider that there has been very little work investigating POS tagging (Thouësny, 2009; van Rooy and Schäfer, 2002; de Haan, 2000) or parsing (Rehbein et al., 2012; Krivanek and Meurers, 2011; Ott and Ziai, 2010) of learner language, due to a lack of annotated data or clear standards. Furthermore, the studies on parsing often first map to a target form, while many situations—such as extracting parse features for error detection (Tetreault et al., 2010) or identifying criterial features indicating learner proficiency level (Hawkins and Buttery, 2010)—require direct parsing of learner language. Defining and applying syntactic annotation provides a clearer picture of the goal for parsing learner language, and evaluation data to do so. Only by developing such annotation can research into POS tagging and syntactic parsing for learner language make serious advancements.

As mentioned, proposals for linguistic annotation split categories into multiple layers. In proliferating categories, however, we must ask what these categories denote and whether they should all be marked by annotators. We specifically look at syntactic dependency annotation, each layer based on different evidence: morphological dependencies, distributional dependencies, and subcategorization (Dickinson and Ragheb, 2011) . In order to define these layers, we must revisit core syntactic principles, to clearly delineate the different layers and their realizations for the in-progress language of learners. Our most important contribution is to outline the questions which need to be addressed for grammatical annotation of learner language.

2 Annotation for Learner Language

Since learner language includes non-native-like constructions, an annotation scheme with single categories for each native-like property does not seem to be adequate (Dickinson and Ragheb, 2011, 2009; Díaz-Negrillo et al., 2010). We thus adopt a multi-layer annotation approach (cf., e.g., Lüdeling et al., 2005), which allows us to capture different pieces of evidence, some of which might conflict for the same token. We thus need to be clear on what the evidence is.

Although we focus on syntactic dependencies, we start by examining part-of-speech (POS) information. Consider two POS layers, one for *morphological* evidence and one for *distributional*. For most native constructions, the layers include the same information, but mismatches arise

with non-canonical structures. For example, in *I have see a movie*,[1] the word *see* has conflicting evidence. The morphological form is the base form or the non-3rd person singular present tense; distributionally, the position is of a dependent of *have*, i.e., a past participle. The use of two POS layers captures the mismatch between morphology and distribution without referencing a unified POS. In this framework, errors are often derivable from mismatches between layers.

Focusing on the evidence relevant for dependency annotation, we build from the POS layers, with a morphosyntactic and a distributional layer of dependencies. We also include subcategorization information, to capture issues relating to the presence or absence of arguments (Dickinson and Ragheb, 2011, 2009). Most of the paper will be spent on defining these three layers of annotation, but we can see the impetus for them by continuing this example. Figure 1 shows the morphosyntactic dependency tree,[2] where the relations are based on the surface form of the tokens and the morphological POS tags.[3] We also see subcategorization frames.

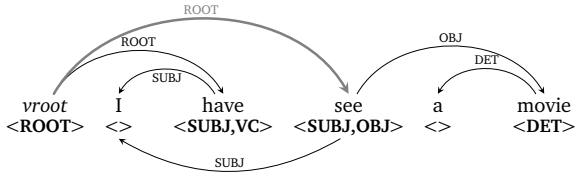

Figure 1: Morphosyntactic dependency tree

By contrast, figure 2 shows the distributional dependency tree. The two trees feature discrepancies in how *see* is treated, as the morphological and distributional evidence diverge.

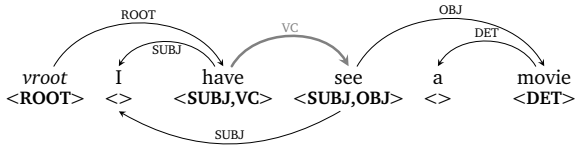

Figure 2: Distributional dependency tree

We can note: a) distribution and morphology often coincide, with four out of five dependency relations repeated; and b) with subcategorization, there is already a mismatch within the morphosyntactic tree (figure 1), as *have* does not realize its verbal complement (VC). We return to issues of redundancy in section 6 after clarifying the three different layers.

3 Defining Subcategorization

While dependencies model what is realized, learner language often contains violations of argument structure (e.g., a missing subject); to capture the nature of these cases, we need to model what is *selected*, in order to identify mismatches (see Dickinson and Ragheb, 2011). Encoding subcategorization does this. As one example, in (1), *house* requires a determiner.

[1] Unless stated, examples are from two corpora collected from English L2 learners, used for developing annotation.
[2] We refer to this layer as *morphosyntactic*, as it incorporates some syntactic information, as described in section 4.
[3] We illustrate with adaptations of the CHILDES dependency scheme (Sagae et al., 2010, 2007).

One way to capture this is as in figure 3, where *house* selects for a determiner on the level of subcategorization (<DET>), though no determiner is present.

(1) . . . we moved again to other **house** . . .

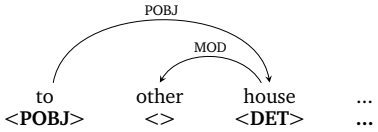

Figure 3: Partial tree with morphosyntactic dependencies and subcategorization frames

Whether derived from the L1, L2, or IL, subcategorization encodes general constraints in the language. This is a different perspective than with annotating dependencies, as dependencies are based on the (local) *evidence* in the sentence (see sections 4 and 5)—e.g., in figure 2, the subcategorization of *see* is for the word in general, while the dependency reflects its immediate context. We model subcategorization on the basis of the requirements in the target language (L2), as constraints most naturally coincide with the language being learned (see also section 5). To say that *house* selects for a determiner in (1) is a fact derived from the L2.

An important question—as with lexical stem information on the POS level (Díaz-Negrillo et al., 2010)—is: what do we do about ambiguity? Words may have many subcategorization frames (Levin, 1993), and we could annotate all of them, since they represent lexical constraints equally well. For a given sentence, however, not all are equally relevant. In (2), for instance, the use of *community* in the whole essay is as a count noun, even though *community* can have uses where a determiner is not required. If we annotate all subcategorizations, then both <DET> and <> are marked, thereby making it unclear whether the learner is doing anything novel.

(2) One [goal] is to contribute to both global and local **community** through my job .

We thus choose to annotate the subcategorization frame which best fits the context of a given sentence. Note what this means: subcategorization annotation is now not totally lexical. It is a lexical property combined with some *contextual* information. Space precludes discussing exactly how context is used to whittle the subcategorization possibilities down to one.

To sum, annotating subcategorization requires answering: 1) What is the source of information? (L1? L2? IL?) 2) How does one handle lexical ambiguity, in particular how much context should be incorporated? Future work can also investigate: 3) how does one disambiguate for an ambiguous context? Our answers of using the L2 as a reference frame and incoporating sentence context means that we will overlap with distribution, as discussed in section 6.

4 Defining Morphosyntax

Morphosyntactic dependencies are based on the visible forms of words. In (3), for instance, regardless of the distribution of *chooses*, its morphology is a third singular present tense verb.

(3) I had a problem a bout **chooses** my car . . .

While this is relatively clear for POS, we must work out what it means to annotate a dependency graph based on morphology, given that dependencies are normally either syntactic or semantic entities (though, see Mel'čuk, 1988). By *morphosyntactic* dependencies, we refer to the syntactic functions derived from the morphological forms. Back in figure 1, for instance, *see* is morphologically a candidate for a ROOT dependency, as it occurs in a (possibly) tensed form.

It is important to distinguish the task: do we base trees on context-sensitive morphological analysis (cf. POS tagging) or context-independent analysis, with multiple analyses? Keeping every possible analysis makes fewer assumptions, but becomes infeasible. If every word in a sentence of n words had 2 possible morphological POS tags, there would potentially be a different tree for every unique sequence of tags—2^n trees. Thus, we choose to annotate only the most contextually-relevant morphological dependency. This is also in line with the idea that some morphological tags are too distant to annotate. Considering (1), for example, VERB is one possibility for the word *house*. Yet in this sentence, the usage is clearly nominal—an orthogonal fact to any non-native properties of the phrase.

How then do we determine which of multiple tags to annotate? As with subcategorization (section 3), we propose annotating the closest fit to the context. In figure 1, for instance, we annotate the correctly-used *have* as ROOT—appropriate for this tensed finite verb, but not for its alternative category of base form verb (depending upon one's definition of ROOT).

We leave open to what degree *context* is defined by syntax, semantics, discourse, or even extra-linguistic factors. For a non-native case, consider (4), an example where the morphological form is ambiguous between base form verb and non-third person singular present tense. Neither exactly matches the context, but as the head of the entire sentence, the tensed form could take precedence, as sentences syntactically require tense, leading to an annotation based on the morphological form of a non-third singular present verb.

(4) This first year **have** been wonderful ... (Díaz-Negrillo et al., 2010)

Importantly, the occurence of this phenomenon has led us to define morphosyntax to include some degree of contextual information, as was done with subcategorization. While we choose to base decisions on context, one may use other properties to disambiguate morphosyntax, e.g., taking the base form as more "basic" in (4), relying on an error taxonomy, etc.

There are further issues in defining morphosyntax which, due to space limitations, we only mention here. For example, returning to (3) and ignoring the space in *a bout* (=*about*), the dependency relation between *about* and *chooses* should be one appropriate for a verb, as this is the morphological form of *chooses*, but should the label be consistent with the head (in this case, a preposition, normally taking a noun phrase as its complement)?

To sum, annotating morphosyntax requires answering: 1) Is the analysis context-sensitive? 2) How does one disambiguate in context? Future work can also investigate: 3) How compatible with the head does a dependency relation need to be? 4) Is this dependency relation based more on lexical or categorical properties? With a contextually-influenced layer, we again overlap with distribution, which we now turn to.

5 Defining Distribution

We define a syntactic distributional slot as a position where a token with particular properties (e.g., singular noun) is predicted to occur, on the (syntactic) basis of its surrounding tokens. In

the constructed *Him sleep*, for instance, the presence of a verb (*sleep*) predicts a nominative noun functioning as subject to its left, while the presence of a third singular pronoun at the beginning of the sentence can be said to predict a third singular verb to the right. The pronoun predicts some set of properties (agreement) and the verb an orthogonal set of properties (case). Parts of those slots are satisfied, and parts are not.

As with subcategorization (section 3), we need to make precise the basis for the linguistic categories being used. Since all learners are learning the same L2, there are common aspects to their interlanguage development, in spite of the potential influence of the native language (L1) (Ellis, 2008). We thus use the L2 as a reference frame for the annotation, to define properties such as: "this verb requires a nominative subject to the left." While one might want to directly encode IL, it is not clear what terms like "subject" mean in such a case. Additionally, annotation reliability would be an issue for L1-based or IL-based annotation, as the same sentence can have different analyses depending upon the L1 (Gass and Selinker, 2008, p. 106).

Distributional syntax can be disambiguated by morphology to get a single layer, just as morphosyntax can be disambiguated by context. In *having an experience*, the slot before the noun could be either a determiner (DET) or a quantifier (QUANT), e.g., *some experience*. If this were purely distribution, we would need to mark both possibilities, in addition to noun modifier (MOD). The fact that the word *an* is present, though, leads to DET as the best relation.

Complements Looking at what drives distributional predictions, complements can be government-based or agreement-based. In a case of syntactic government, a head selects its dependents and determines specific properties which need to be true of them. In these cases, the definition of a distributional slot follows from the head's subcategorization (see section 6). For example, in (5), *with* selects for a prepositional object, governing the case of the object. Distributionally, then, *he* is in a prepositional object position, regardless of its actual form.

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

For cases of agreement, the head may not be the locus of agreement. Consider (6), where the subject-verb disagreement affects the forms of both tokens. In this case, the verb is the head, but the subject could be considered the source of the agreement features. This is why, instead of being head-driven, we speak of one token predicting another token's properties.

(6) **He play** by toys.

The exact treatment for cases of agreement depends upon defining the source of agreement in one's syntactic theory—as annotation depends upon the theory employed (Leech, 2004; Rambow, 2010). For most label inventories, there is no distinction for agreement, but if there were (e.g., SUBJ3s vs. SUBJp), one could choose to use the relation driven by the *dependent* (SUBJ3s), since the prediction works "backwards." This would directly contradict the head-driven subcategorization (SUBJpl), specifying that the L2 requires a verb which agrees. The interaction with morphosyntax—which would underspecify (to SUBJ or to nothing)—is then similar to the case of adjuncts, as in section 6 (see the discussion around *He runs quick*).

Adjuncts Adjuncts select for their heads (Pollard and Sag, 1994), yet at the same time, heads delimit the properties of adjuncts (cf. selective adjunction, Abeillé and Rambow, 2000).

Annotation is straightforward if the selective properties do not conflict. Unlike complements, adjunction cases are not mediated via subcategorization; we discuss them in section 6.

To sum, annotating distributional dependencies requires answering: 1) What is the basis of the categories? (L1? L2? IL?) 2) How does one disambiguate in context, specifically what is the role of morphology? 3) What information drives the distributional predictions? Our answers led us to conclude that subcategorization drives the predictions in part, but not in whole.

6 Annotation Redundancies

Consider again the trees in figures 1 and 2: 1) the *vroot* selects for one ROOT and finds two in the morphosyntactic tree; 2) *have* selects for a verbal complement (VC), not realized morphosyntactically; and 3) the head and label of *see* differs between the trees. But with a better understanding of distribution, note what the third mismatch means: *see* is in the distributional slot of a verbal complement in figure 2, defined by virtue of the subcategorization list of *have*. In other words, this mismatch is already in mismatch #2, where *have* selects for VC. Based on the treatment of complements and adjuncts, we are more inclined towards removing distributional dependencies. We briefly outline some cases which led to this conclusion here.

Complements In terms of argument structure, non-canonical constructions center around a mismatch between what is subcategorized for and what is realized (cf. consistency, completeness, and coherence (Bresnan, 2001)). In these cases, distributional dependencies require annotating more than is known from the evidence available in the sentence, as we will illustrate.

For mismatched requirements, consider (7), where a non-finite clause (*what success to be*) appears as the complement of *wondered*, where one would expect a finite clause.

(7) I wondered what success **to be**.

Morphologically, *to be* has non-finite marking and the clause is thus a non-finite complement (XCOMP), as shown in the left side of figure 4, assuming *to* is the head. The subcategorization selects for a finite complement (COMP), making for a clear mismatch.

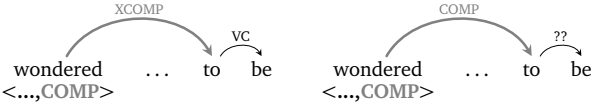

Figure 4: Morphosyntactic (left) and distributional (right) trees for a complement mismatch

Based on subcategorization predictions, in a distributional tree we use COMP as a label. The subtree, however, is unclear, as shown on the right side. If *to be* is in a finite distributional position, is *to* a finite modal verb with a verbal complement? Is *be* the finite verb with an extraneous *to* marking? Annotating subcategorization does not force such an internal analysis.

Missing arguments illustrate how subcategorization captures information distributional dependencies cannot. Consider (8), for example, where the learner wrote *shakes*. Neither verb has an object, but *shakes* here requires an object that *runs* does not. Both, however, have the exact same distributional trees (not shown), with only a SUBJ from the verb.

(8) As he saw it, at once he takes it and { **shakes** | **runs** }

This difference is easily captured via subcategorization, distinguishing <SUBJ> (*runs*) from <SUBJ,OBJ,> (*shakes*). Extra arguments work in a similar fashion. Missing heads (e.g., copulas) are more complicated, but the challenge is for all layers; space precludes a discussion here.

Adjuncts Non-native use of adjuncts are cases where distributional dependencies seem to be required, as subcategorization does not include adjuncts. We sketch some of the issues here.

Consider an adjective modifying a verb, as in the constructed *He runs quick* (cf. real examples like *It is quickly*). In this case, if we ignore the morphology of the dependent, the distributional layer would reflect the selectional properties of the head, encoding *quick* as a verbal modifier (JCT) of *runs*, whereas the morphosyntactic layer lacks a label which fits with both the head (e.g., JCT) and the dependent (MOD) (cf. (3)). With no adjunct subcategorization, the morphosyntactic layer does not convey that the L2 requires something like a JCT relation here.

This is not the entire story, though. First, once POS is taken into account, we have more information than an unspecified relation between *runs* and *quick*; namely, it is verb+adjective which has an undefined relation. Secondly, defining a distributional label makes more assumptions than it may at first appear. In this case, this is also an appropriate slot for CJCT (clausal adjunct) or XJCT (non-finite adjunct); it is only because *quick* is similar to *quickly* that we assume a label appropriate for adverbs. Working out which label to use when the morphology is not a totally valid piece of evidence requires more analysis (cf. (7)).

7 Conclusion & Outlook

We started with a proposal for learner language to annotate: 1) subcategorization, 2) morphosyntactic dependencies, and 3) distributional dependencies. By precisely defining each layer, we uncovered several questions that need to be addressed, including the degree of context to incorporate into subcategorization and morphosyntax in order to arrive at a single annotation. We suggested that it may be preferable to annotate subcategorization *instead of* distribution, as subcategorization is a source of distribution; such a decision prevents annotators from having to specify distributional trees in cases where they are indeterminate. Based on our ongoing annotation efforts, we have developed extensive annotation guidelines reflecting our decisions and examining various constructions, which will be made publicly available in the near future.

The decisions discussed here raise questions for the future, the foremost one being to definitively answer the questions raised about each layer here. Where, for example, does semantic evidence fit and what is the precise role of word order in defining each layer? We have only scratched the surface, and to carry out automatic analysis, for instance, will require a deeper look into the connections between different pieces of evidence. Secondly, how does such a division of layers of syntax bear on other non-canonical language use, such as web data, or the annotation of native language (cf. the discussion in Rehbein et al., 2012)? It is an open question as to whether the elucidation of layers for learner language can impact annotation schemes for syntax more broadly. Thirdly, there is a need to work out the exact connection between this annotation and the annotation of target hypotheses for learner language, building from annotation mismatches. Mismatches in annotation layers point to errors (Dickinson and Ragheb, 2009), an insight used for creating multiple parsing models for learner language (e.g., Dickinson and Lee, 2009).

Acknowledgments

We would like to thank Detmar Meurers, the IU CL discussion group, and the three anonymous reviewers for helpful feedback.

References

Abeillé, A. and Rambow, O. (2000). Tree adjoining grammar: An overview. In Abeillé, A. and Rambow, O., editors, *Tree Adjoining Grammars: Formalisms, Linguistic Analyses and Processing*, pages 1–68. CSLI Publications, Stanford, CA.

Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell Publising, Oxford.

de Haan, P. (2000). Tagging non-native english with the tosca-icle tagger. In Mair, C. and Hundt, M., editors, *Corpus Linguistics and Linguistic Theory*, pages 69–79. Rodopi, Amsterdam.

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

Dickinson, M. and Lee, C. M. (2009). Modifying corpus annotation to support the analysis of learner language. *CALICO Journal*, 26(3).

Dickinson, M. and Ragheb, M. (2009). Dependency annotation for learner corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*, pages 59–70, Milan, Italy.

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

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

Gass, S. M. and Selinker, L. (2008). *Second Language Acquisition: An Introductory Course*. Taylor & Francis, New York, third edition.

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

Hawkins, J. A. and Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1(1):1–23.

Krivanek, J. and Meurers, D. (2011). Comparing rule-based and data-driven dependency parsing of learner language. In *Proceedings of the Int. Conference on Dependency Linguistics (Depling 2011)*, Barcelona.

Leech, G. (2004). Adding linguistic annotation. In Wynne, M., editor, *Developing Linguistic Corpora: a Guide to Good Practice*, pages 17–29. Oxbow Books, Oxford.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Lüdeling, A., Walter, M., Kroymann, E., and Adolphs, P. (2005). Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*, Birmingham.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press.

Nagata, R., Whittaker, E., and Sheinman, V. (2011). Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1210–1219, Portland, OR.

Ott, N. and Ziai, R. (2010). Evaluating dependency parsing performance on German learner language. In *Proceedings of TLT-9*, volume 9, pages 175–186.

Pienemann, M. (1992). Coala-a computational system for interlanguage analysis. *Second Language Research*, 8(1):59–92.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. The University of Chicago Press.

Ragheb, M. and Dickinson, M. (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.

Rambow, O. (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.

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

Rehbein, I., Hirschmann, H., Lüdeling, A., and Reznicek, M. (2012). Better tags give better trees - or do they? *Linguistic Issues in Language Technology (LiLT)*, 7(10).

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

Sagae, K., Davis, E., Lavie, A., and an Shuly Wintner, B. M. (2010). Morphosyntactic annotation of childes transcripts. *Journal of Child Language*, 37(3):705–729.

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

Tetreault, J., Foster, J., and Chodorow, M. (2010). Using parse features for preposition selection and error detection. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 353–358, Uppsala, Sweden.

Thouësny, S. (2009). Increasing the reliability of a part-of-speech tagging tool for use with learner language. Presentation given at the Automatic Analysis of Learner Language (AALL'09) workshop on automatic analysis of learner language: from a better understanding of annotation needs to the development and standardization of annotation schemes.

van Rooy, B. and Schäfer, L. (2002). The effect of learner errors on pos tag errors during automatic pos tagging. *Southern African Linguistics and Applied Language Studies*, 20:325–335.