

Chapter 3

The CHILDES GR Annotation Scheme

One crucial aspect of producing useful automatic syntactic analysis for child-parent dialog transcripts is the definition of a suitable annotation scheme that targets the specific type of syntactic information needed by the child language community. To address this need, we have developed the CHILDES Grammatical Relation (GR) annotation scheme, described in this chapter.

3.1 Representing GRs with Labeled Dependencies

We represent syntactic information in CHILDES data in terms of labeled dependencies that correspond to grammatical relations (GRs), such as subjects, objects, and adjuncts. As in many flavors of dependency-based syntax, each GR in our scheme represents a relationship between two words in a sentence: a *head* (sometimes called a parent or regent) and a *dependent* (sometimes called a child or modifier). In addition to the head and dependent words, a GR also includes a *label* (or GR type) that indicates what kind of syntactic relationship holds between the two words. Each word in a sentence must be a dependent of exactly one head word (but heads may have several dependents). The single exception to this rule is that every sentence has one “root” word that is not a dependent of any other word in the sentence. To achieve consistency across the entire sentence (with each word having exactly one head), we make the root word a dependent of a special empty word, appended to the beginning of every sentence. We call this empty word the “LeftWall”³. Figure 3.1 shows the syntactic annotation of two sentences.

To describe the properties and constraints of syntactic structures defined in our annotation scheme, it is useful to think of these structures as graphs, where the words are nodes, and directed edges exist between each dependent-head pair, from the dependent to the head. A well-formed dependency structure must be connected. More specifically, from any word in the sentence (any node in the graph), there is a path to the root word (the single word in the sentence that is a dependent of the LeftWall). In addition, a well-formed dependency structure must be acyclic. Therefore, if we ignore the LeftWall and the directionality of the edges, a syntactic structure in this scheme is in fact a tree rooted at the root word.

One additional property that can be defined for dependency trees in our annotation scheme is projectivity. An in-order traversal of a projective dependency tree must list the words in the same left-to-right order as in the original sentence. In other words, a projective tree drawn above the sentence must have no crossing-branches. Syntactic structures in our scheme are *not* required to be projective. In fact, annotation of languages with free word order should produce dependency

³ We borrow the term LeftWall from Link Grammar (Sleator and Temperley, 1991), another dependency-based formalism. This is similar to the EOS word used in the definition of bare-bones dependency in Eisner (1996).

trees that frequently exhibit non-projectivity. However, English sentences rarely do, and most (if not all) of the recent work on lexicalized syntactic parsing of English (using constituent trees or dependencies) assumes projectivity (e.g. Rathnaparkhi, 1996; Eisner, 1996; Collins, 1996; Charniak, 2000; Nivre, 2004). Because in this thesis we are working exclusively with English, syntactic structures will be assumed to be projective, although this is not a constraint imposed by the annotation scheme.

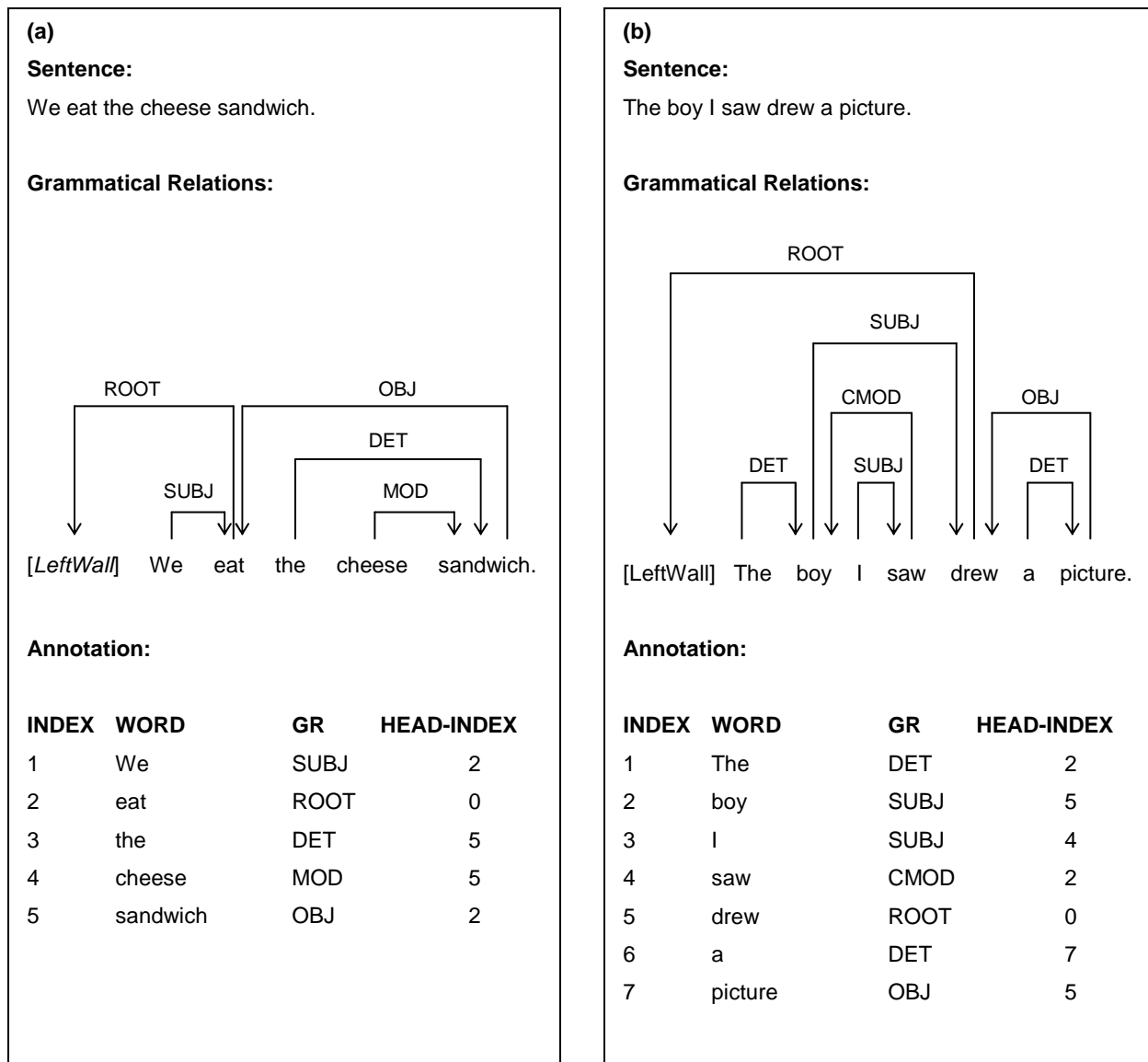


Figure 3.1 Sample sentences annotated with CHILDES GRs.

3.2 Specific GR Labels

The specific set of GR types (or labels) included in our annotation scheme was developed by using the GRs in the annotation scheme of Carroll et al. (2003) as a starting point, and adapting the set to suit specific needs of child language research. Sources of refinement included a survey of the child language literature (Fletcher & MacWhinney, 1995), a review of existing measures of syntactic development (MacWhinney, 2000), and input from child language researchers. Rather than just relying on our own experience and training in computational linguistics to determine a set of grammatical relations that addresses important research issues in child language, we solicited public input regarding this set of GRs via a posting to a mailing list with over 2,000 members who are active child language researchers (info-childes@mail.talkbank.org). A similar request for input appeared in the newsletter of the International Association for the Study of Child Language (IASCL).

Our final set of GRs is the product of two years of thoughtful development, paying close attention to the considerations mentioned above. Additionally, we note that a subset of these categories is sufficient to compute some of the most important measures of child language syntactic development – the Index of Productive Syntax, or IPSyn (Scarborough, 1990), the Developmental Sentence Score, or DSS (Lee, 1974), and the Language Assessment, Remediation, and Screening Procedure, or LARSP (Fletcher and Garman, 1988). An automated procedure for computing IPSyn using this GR annotation scheme is presented in Chapter 7.

A comprehensive list of the grammatical relations in the CHILDS GR annotation scheme appears below. Example GRs appear in the form $GR(\text{dependent}-di, \text{head}-hi)$, where GR is the GR label, dependent is the dependent word, di is the index of the dependent word (its position counting from left to right, starting at 1), head is the head word, and hi is the index of the head word.

- **SUBJ:** used to identify the subject of clause, when the subject itself is not a clause. Typically, the head is a verb (the main verb of the subject’s clause), and the dependent is a noun (or another nominal, such as a pronoun, or any head of a noun phrase). Clauses that act as subjects are denoted by $CSUBJ$ or $XSUBJ$, depending on whether the clausal subject is finite or non-finite (see below). Note that throughout the labeling scheme, in general, CGR denotes a finite clausal version of the relation GR , and XGR denotes a non-finite clausal version of the relation GR , where the relation GR may be a subject, adjunct, predicate nominal, etc.

Example:

Mary saw a movie.

SUBJ(Mary-1, saw-2)

- **CSUBJ:** used to identify the finite clausal subject of another clause. The head is typically the main verb of the matrix clause, and the dependent is the main verb of the clausal subject.

Example:

That Mary screamed scared John.

CSUBJ(screamed-3, scared-4)

- **XSUBJ:** used to identify the non-finite clausal non-finite subject of another clause. The head is typically the main verb of the matrix clause, and the dependent is the main verb of the clausal subject.

Example:

Eating vegetables is important.

XSUBJ(eating-1, is-3)

- **OBJ:** used to identify the first object of a verb. Typically, the head is a verb, and the dependent is a noun (or other nominal). The dependent must be the head of a required non-clausal and non-prepositional complement of the verb (head of OBJ). A clausal complement relation should be denoted by COMP or XCOMP (depending on whether the clausal complement is finite or non-finite, see below), not OBJ or OBJ2.

Example:

Mary saw a movie.

OBJ(movie-4, saw-2)

- **OBJ2:** used to identify the second object of a ditransitive verb, when not introduced by a preposition. Typically, the head is a ditransitive verb, and the dependent is a noun (or other nominal). The dependent must be the head of a required non-clausal and non-prepositional complement of a verb (head of OBJ2) that is also the head of an OBJ relation. A second complement that has a preposition as its head should be denoted by IOBJ, not OBJ2.

Example:

Mary gave John a book.

OBJ2(book-5, gave-2)

- **IOBJ:** used to identify an object (required complement) introduced by a preposition. When a prepositional phrase appears as the required complement of a verb, it is the dependent in an IOBJ relation, not a JCT (adjunct) relation. The head is typically a verb, and the dependent is a preposition (not the complement of the preposition, see POBJ below).

Example:

Mary gave a book to John.

IOBJ(to-5, gave-2)

- **COMP:** used to identify a finite clausal complement of a verb. The head is typically the main verb of the matrix clause, and the dependent is the main verb of the clausal complement.

Example:

I think that Mary saw a movie.

COMP(saw-5, think-2)

- **XCOMP:** used to identify a non-finite clausal complement of a verb. The head is typically the main verb of the matrix clause, and the dependent is the main verb of the clausal complement. The XCOMP relation is only used for non-finite clausal complements, not predicate nominals or predicate adjectives (see PRED below).

Examples:

Mary likes watching movies.

XCOMP(watching-3, likes-2)

Mary wants me to watch a movie.

XCOMP(watch-5, wants-2)

- **PRED:** used to identify a predicate nominal or predicate adjective of the subject of verbs such as *be* and *become*. The head of PRED is the verb, not its subject. The predicate may be nominal, in which case the dependent is a noun (or other nominal), or adjectival, in which case the dependent is an adjective. PRED should not be confused with XCOMP, which identifies a non-finite complement of a verb (some syntactic formalisms group PRED and XCOMP in a single category).

Examples:

Mary is a student.

PRED(student-4, is-2)

Mary got angry.

PRED(angry-3, got-2)

- **CPRED:** used to identify a finite clausal predicate of the subject of verbs such as *be* and *become*. The head of CPRED is the main verb (of the matrix clause), not its subject.

Example:

The problem is that Mary sees too many movies.

CPRED(sees-6, is-3)

- **XPRED:** used to identify a non-finite clausal predicate of the subject of verbs such as *be* and *become*. The head of XPRED is the main verb (of the matrix clause), not its subject.

Example:

My goal is to win the competition.

XPRED(win-5, is-3)

- **JCT:** used to identify an adjunct (an optional modifier) of a verb, adjective, or adverb. The head of JCT is a verb, adjective or adverb. The dependent is typically an adverb, a

preposition (in the case of phrasal adjuncts headed by a preposition, such as a prepositional phrase). Intransitive prepositions may be treated as adverbs, in which case the JCT relation applies, or particles, in which case the PTL relation (see below) applies. Adjuncts are optional, and carry meaning on their own (and do not change the basic meaning of their JCT heads).

Examples:

Mary spoke very clearly.
 JCT(clearly-4, spoke-2)
 JCT(very-3, clearly-4)

Mary spoke at the meeting.
 JCT(at-3, spoke-2)

Mary is very tired.
 JCT(very-3, tired-2)

- **CJCT:** used to identify a finite clause that acts like an adjunct of a verb, adjective, or adverb. The head of CJCT is a verb, adjective, or adverb. The dependent is typically the main verb of a subordinate clause.

Example:

Mary left after she heard the news.
 CJCT(heard-5, left-2)

- **XJCT:** used to identify a non-finite clause that acts like an adjunct of a verb, adjective, or adverb. The head of CJCT is a verb, adjective, or adverb. The dependent is typically the main verb of a non-finite subordinate clause.

Example:

Mary left after hearing the news.
 CJCT(hearing-4, left-2)

- **MOD:** used to identify a non-clausal nominal modifier or complement. The head is a noun, and the dependent is typically an adjective, noun or preposition.

Examples:

Mary saw a red car.
 MOD(red-4, car-5)

Mary saw the boy with the dog.
 MOD(with-5, boy-4)

The Physics professor spoke clearly.

MOD(Physics-2, professor-3)

- **CMOD:** used to identify a finite clause that is a nominal modifier (such as a relative clause) or complement. The head is a noun, and the dependent is typically a finite verb.

Example:

The student who visited me was smart.

CMOD(visited-4, student-2)

- **XMOD:** used to identify a non-finite clause that is a nominal modifier (such as a relative clause) or complement. The head is a noun, and the dependent is typically a non-finite verb.

Example:

The student standing by the door is smart.

XMOD(standing-3, student-2)

- **AUX:** used to identify an auxiliary of a verb, or a modal. The head is a verb, and the dependent is an auxiliary (such as *be* or *have*) or a modal (such as *can* or *should*).

Examples:

Mary has seen many movies.

AUX(has-2, seen-3)

Are you eating cake?

AUX(are-1, eating-3)

You can eat cake.

AUX(can-2, eat-3)

- **NEG:** used to identify verbal negation. When the word *not* (contracted or not) follows an auxiliary or modal (or sometimes a verb), it is the dependent of a NEG relation (not JCT), where the auxiliary, modal or verb (in the absence of an auxiliary or modal) is the head.

Examples:

I am not eating cake.

NEG(not-3, am-2)

Speak not of that subject.

NEG(not-2, speak-1)

- **DET:** used to identify a determiner of a noun. Determiners include *the*, *a*, as well as possessives (*my*, *your*, etc) and demonstratives (*this*, *those*, etc), but not quantifiers (*all*, *some*, *any*, etc; see QUANT below). Typically, the head is a noun and the dependent is a determiner. In cases where a word that is usually a determiner does not have a head, there is no DET relation.

Example:

The students ate that cake.

DET(the-1, students-2)

DET(that-4, cake-5)

- **QUANT:** used to identify a nominal quantifier, such as *three*, *many*, and *some*. Typically, the head is a noun, and the dependent is a quantifier. In cases where a quantifier has no head, there is no QUANT relation.

Example:

Many students saw three movies yesterday.

QUANT(many-1, students-2)

QUANT(three-4, movies-5)

- **POBJ:** used to identify the object of a preposition. The head is a preposition, and the dependent is typically a noun.

Example:

Mary saw the book on her desk.

POBJ(desk-7, on-5)

- **PTL:** used to identify the verb particle (usually a preposition) of a phrasal verb. Intransitive prepositions that change the meaning of a verb should be in a PTL relation, not JCT (see above). The head is a verb, and the dependent is a preposition.

Example:

Mary decided to put off the meeting until Thursday.

PTL(off-5, put-4)

- **CPZR:** used to identify a complementizer (usually a subordinate conjunction). The head is a verb, and the dependent is a complementizer. It is the verb (head of a CPZR relation) of an embedded clause that acts as the dependent in a relation involving the embedded clause and its matrix clause, not the complementizer (the verb is higher in the dependency tree than the complementizer).

Examples:

I think that Mary left.

CPZR(that-3, left-5)

She ate the cake because she was hungry.

CPZR(because-5, was-7)

- **COM:** used to identify a communicator (such as *hey*, *okay*, etc). Because communicators are typically global in a given sentence, the head of COM is typically the root of the sentence's dependency tree (the dependent of the ROOT relation, see below). The dependent is a communicator.

Example:

Okay, you can read the book.

COM(okay-1, read-4)

- **INF:** used to identify an infinitival particle (*to*). The head is a verb, and the dependent is always *to*.

Example:

Mary wants to read a book.

INF(to-3, read-4)

- **VOC:** used to identify a vocative. As with COM, the head is the root of the sentence. The dependent is a vocative.

Example:

Mary, you may not eat the cake.

VOC(Mary-1, eat-5)

- **TAG:** used to identify tag questions, where the tag is headed by a verb, auxiliary or modal. Tags of the type found in “this is red, *right?*” and “Let me do it, *okay?*” are identified as dependents in a COM relation, not TAG. The head of a TAG relation is typically a verb, and the dependent is the verb, auxiliary or modal in the tag question.

Example (other relevant GRs also shown, for clarity):

This is good, isn't it?

TAG(is-4, is-2)

NEG(n't-5, is-4)

SUBJ(it-6, is-4)

- **COORD:** used to identify coordination. The head is a coordinator (usually *and*), and several types of dependents are possible. The head coordinator may have two or more dependents, including the coordinated items. Once the COORD relations are formed between the head coordinator and each coordinated item (as dependents), the coordinated phrase can be thought of as a unit represented by the head coordinator. For example, consider two coordinated verb phrases with a single subject (as in “I walk and run”), where two verbs are dependents in COORD relations to a head coordinator. The head of COORD is then also the head of a SUBJ relation where the subject is the dependent. This indicates that both verbs have that same subject. In the case of a coordinated string with multiple coordinators, the COORD

relation applies compositionally from left to right. In coordinated lists with more than two items, but only one coordinator, the head coordinator takes each of the coordinated items as dependents. In the absence of an overt coordinator, the right-most coordinated item acts as the coordinator (the head of the COORD relation).

Example (other relevant GRs also shown, for clarity):

Mary likes cats and dogs.

COORD(cats-3, and-4)

COORD(dogs-5, and-4)

OBJ(and-5, likes-2)

Mary likes birds and cats and dogs.

COORD(birds-3, and-4)

COORD(cats-5, and-4)

COORD(and-4, and-6)

COORD(dogs-7, and-6)

OBJ(and-6, likes-2)

- **ROOT:** this is the relation between the topmost word in a sentence (the root of the dependency tree) and the LeftWall. The topmost word in a sentence is the word that is the head of one or more relations, but is not the dependent in any relation with other words (except for the LeftWall). This word is the dependent in the ROOT relation, and the LeftWall is the head.

Example:

Mary saw many movies last week.

ROOT(saw-2, LeftWall-0)

3.3 Related GR Annotation Schemes

Our GR annotation approach shares characteristics with other recent annotation schemes. Like Carroll et al. (2003), we use a rich set of GRs that provide detailed information about syntactic relationship types. As previously mentioned, the GR set of Carroll et al. was used as a starting point in the development of the CHILDES GR set. Like the annotation used by Lin (1998) for parser evaluation, and the scheme of Rambow et al. (2002) for annotation of a small treebank of spoken English, we use dependency structures instead of the constituent structures that continue to be thought of as the primary form of syntactic representation by many. We share Rambow et al.'s findings that dependency structures, compared to constituent structures, provide increased ease and reliability of manual annotation. However, our choice to use dependency-based grammatical relations is motivated primarily by the fact that much of the syntax-based child

language work that uses CHILDES data is predominantly GR-driven, and dependencies offer a natural way to encode such GRs.

In spite of the similarities mentioned above, our annotation scheme differs from those of Carroll et al. and Rambow et al. in important ways. The scheme of Carroll et al. does not annotate sentences with a full dependency structure, but rather lists a set of GRs that occur in the sentence as a set of propositions. By doing away with the requirement of a complete and consistent dependency structure, their scheme allows for n-ary relations (while our GRs are strictly binary) and greater flexibility in working with GRs that may not fit together in a global graph structure. In our framework, n-ary relations (with $n > 2$) must be represented indirectly, using combinations of binary GRs. Their set of GRs, although detailed, is meant for general-purpose annotation of text, and does not include specific pieces of information we have identified as important to the child language community. In addition, they distinguish between initial (or deep) GRs, and actual (or surface) GRs, while we report surface GRs only.

The scheme of Rambow et al., on the other hand, is dependency-based, like our CHILDES GR annotation scheme. However, their dependency labels are limited to seven syntactic roles (SRole and DRole features, which can have the values of subj, obj, obj2, pobj, pobj2, adj and root). These seven roles suffice for some applications, but do not offer the granularity needed for CHILDES annotation. In contrast, we have 30 distinct labels for GRs. Like Carroll et al., Rambow et al. annotate surface and deep relations.

3.4 Specific Representation Choices

While the CHILDES GR annotation scheme is suitable for the identification of specific syntactic structures of interest to child language researchers, we should stress that the annotation scheme is in no way meant to be a theory of syntax, nor is it intended to represent all possible kinds of syntactic phenomena. In fact, the choice of grammatical relations and how they are represented was motivated primarily not by specific linguistic theories, but by two practical concerns: the usefulness of the information contained in the annotations to child language research, and (to a lesser extent) automatic annotation of transcripts using syntactic parsing. In addition, the scope of the annotation scheme is purely syntactic. Although semantic roles are undoubtedly valuable, this annotation scheme makes no attempt to represent such information. Therefore, while we address the identification of syntactic relations such as subjects and objects, we do not address the identification of semantic relations such as agents and themes.

Our GRs typically assign content words as heads and function words as dependents. For example, nouns are the heads of determiners (forming a GR of type DET with the noun as the head and the determiner as the dependent), and verbs are chosen as the heads of auxiliaries (forming a GR of type AUX with the verb as the head and the auxiliary as the dependent). An exception to this rule is coordination, where coordinated items are dependents of the coordinator (for example, in “boys and girls” each noun is a dependent of “and”, forming two GRs of type COORD). When both words in a GR are members of lexical (content) categories, the direction of the relation follows common practice, where complements, adjuncts and modifiers are dependents of the words they complement or modify. For example, adjectives are typically

dependents of nouns, and adverbs are typically dependent of verbs, adjectives or other adverbs. Nouns are dependents of verbs when the noun is a complement (such as subject or object), but verbs can also be dependents of nouns, as in the case of relative clauses (where a clause modifies a noun, and the verb is the root of the dependency subtree representing the clause). In the case of prepositional phrases, the preposition is chosen as the head of the prepositional object (in a POBJ relation). By convention, vocatives and communicators are dependents of the main verb of their sentences.

In cases where a clause has a relation to another clause, the verb of the lower (subordinate) clause is used as a dependent. Thus, in the relative clause of figure 3.1(b), the verb “saw” of the relative clause is treated as dependent in a CMOD (clausal modifier of a nominal) relation with the noun “boy”. The other relations in this sentence are as expected: “The” and “a” are dependents in DET relations with “boy” and “picture”, “boy” is the dependent in a SUBJ relation with “saw”, and “picture” is in a OBJ relation with “drew”. Finally, “drew” is the root of the dependency tree for this sentence, and by convention is the dependent in a ROOT relation with the special word LeftWall, which we append to the beginning of the sentence.

A point worth noting is that, in general, only words that appear in a sentence can be participants in a GR as either a head or a dependent (an exception to this statement, specific to child language annotation, is presented in section 3.5). This is in contrast to the scheme of Rambow et al., where an empty word *e* can be added to the sentence in control structures or ellipsis. For example, in the sentence “I wanted to run,” Rambow et al. annotate the empty word *e* as the subject of run, while we do not annotate a subject for run.

3.5 Inter-Annotator Agreement

Inter-annotator agreement was measured by having a corpus of 285 words (48 sentences from a CHILDES corpus) annotated manually by two annotators, independently. The annotators had at least a basic background in Linguistics, and one was familiar with the annotation scheme. The other was trained for about one hour before annotating a separate 10-sentence trial corpus under close guidance. Annotation of the 285-word corpus took about 90 minutes. Annotator agreement was 96.5%, which is about the same as inter-annotator agreement figures for related annotation schemes involving dependencies and grammatical relations. Carroll et al. report 95% agreement, while Rambow et al., report 94% agreement. Because of the size of the corpora used for rating annotator agreement, these differences are not significant. Out of 285 possible labeled dependencies, there were 10 disagreements between the annotators. Of particular interest, four of them were disagreements on the attachments of adjuncts, and three of them were incorrect labeling (with correct dependent-head links) involving COMP, PRED and OBJ.