Annotation for Learner English Guidelines, v. 0.1

Markus Dickinson & Marwa Ragheb

June 9, 2013
# Contents

## Front matter

- 0.1 Notes for Researchers ................................................... vii
- 0.2 Acknowledgements ....................................................... viii

## 1 Getting Started

- 1.1 Quick Intro ................................................................. 1
- 1.2 General Principles ...................................................... 2
  - 1.2.1 Assumptions ......................................................... 3
  - 1.2.2 Benefit of the doubt ................................................ 3
  - 1.2.3 Semantics .............................................................. 4
  - 1.2.4 Evidence ............................................................... 5
  - 1.2.5 Underspecification .................................................. 6
  - 1.2.6 Mismatches .......................................................... 8
  - 1.2.7 Mismatches to define shortest distance ...................... 8
- 1.3 Label Inventories ....................................................... 10
  - 1.3.1 Dependency Relations ............................................. 10
  - 1.3.2 POS Tags ............................................................. 11

## 2 Initial Annotation Layers

- 2.1 Segmentation & Tokenization .......................................... 17
  - 2.1.1 Sentence segmentation ........................................... 17
  - 2.1.2 Tokenization ....................................................... 19
- 2.2 Lemma ........................................................................... 19
  - 2.2.1 Irregulars .............................................................. 20
  - 2.2.2 Misspellings ........................................................... 20
  - 2.2.3 Spelling vs. Morphology ......................................... 22
  - 2.2.4 Spacing issues ....................................................... 23
  - 2.2.5 Lowercase ............................................................ 24
  - 2.2.6 Anonymization ....................................................... 24
  - 2.2.7 Acronyms, unknowns, & foreign terms ....................... 25
- 2.3 POS .............................................................................. 25
2.3.1 POS mismatches .......................... 26
2.3.2 Defining distributional POS .............. 27
2.4 Lexical violations .......................... 31
  2.4.1 Lexical violations vs. Lemma changes .... 32
  2.4.2 Lexical violations vs. POS mismatches ... 33

3 Dependencies ................................ 35
  3.1 Morphosyntactic Dependencies ............. 35
    3.1.1 Inventory .................................. 35
    3.1.2 The *syntax* in morphosyntactic dependencies 43
  3.2 Where we differ from CHILDES ............... 46
    3.2.1 Possessives ................................ 46
    3.2.2 NJCT ........................................ 46
    3.2.3 Particles (PRT) ............................. 46
    3.2.4 Dependents of adjectives ................... 47
    3.2.5 Heads for verbal chains ................... 47
    3.2.6 Secondary Dependencies .................... 48
    3.2.7 Enumeration ................................ 49
    3.2.8 Coordination ................................ 49
    3.2.9 Ellipsis ..................................... 50
  3.3 Overview of annotation ....................... 51
    3.3.1 Verbs & Verbal relations ................. 51
    3.3.2 Nouns and Noun relations .................. 59
    3.3.3 Coordination ................................ 66
    3.3.4 Arguments vs. Adjuncts .................... 75
    3.3.5 Punctuation ................................ 78
  3.4 Subcategorization frames ...................... 80
    3.4.1 Grammaticality ............................. 81
    3.4.2 Determining subcategorization requirements 82
    3.4.3 Specific cases ............................. 86

4 A Variety of Dependency Constructions .......... 89
  4.1 Attachment decisions ........................ 89
  4.2 INCROOT ...................................... 90
  4.3 Extrapolation .................................. 91
  4.4 *wh*-words .................................... 93
    4.4.1 Displacement ................................ 93
    4.4.2 *wh*-questions .............................. 94
    4.4.3 Embedded clauses ........................... 94
    4.4.4 Relative clauses ........................... 96
  4.5 Prepositions vs. Complementizers ............ 97
Front matter

0.1 Notes for Researchers

The annotation scheme here is based on thinking that has evolved over several years, as captured in various papers, listed here. It is also part of an ongoing dissertation project at Indiana University. You can find pdf versions of the papers, as well as other information about the project, at [http://cl.indiana.edu/~salle/](http://cl.indiana.edu/~salle/).


The goal is to annotate syntactic and, to some extent, morpho-syntactic information, without necessarily encoding errors. The papers give more justification for this, but you can also read section 1.2 for more on what the annotation does and does not encode.
Information for citing these guidelines is:


And the BiBTeX entry is:

```bibtex
@TechReport{salle:13,
  author = {Markus Dickinson and Marwa Ragheb},
  title = {Annotation for Learner English Guidelines, v. 0.1},
  institution = {Indiana University},
  year = {2013},
  address = {Bloomington, IN},
  month = {June},
  note = {June 9, 2013},
}
```

It is important to note that, while we have made hundreds of decisions, these are not the only decisions one could have made. We hope that these guidelines are useful, not just to understand what the annotation means, but as a starting point for other annotation and analysis efforts. That is, similar to what we stated in [Ragheb and Dickinson](#) (2012), one of the most important contributions of these guidelines may be “to outline the questions which need to be addressed for grammatical annotation of learner language.”

As long as these guidelines are still in progress, we welcome feedback and discussion. Contact us at: mragheb@indiana.edu or md7@indiana.edu.

Our data will eventually be released, but as a) we are a small annotation effort, and b) we have had to take significant time determining what to annotate and what the annotation denotes—questions which form the core of a PhD thesis-in-progress—please bear with our slowness.

### 0.2 Acknowledgements

As can be gathered from the acknowledgments in the papers above, this work has been affected by many researchers in different areas.

We would also like to acknowledge our student annotators, who had the challenging task of annotating while decisions were still being made: Eric Benzschawel, Frank Linville, Shannon Manley, Lauren Swanson, Zachary Wampler, and Samantha Zimny.

The process was a little unusual, in that, for the most part, students were receiving course credit for annotating, and part of the credit was based on discussing syntax,
learner data, their decisions, etc. Thus, the annotators were required not only to annotate the data, but to reflect in great detail on it. This led to slower annotation, but a learning experience for all.
Chapter 1

Getting Started

1.1 Quick Intro

Before you get into the actual guidelines, here’s a quick idea about what the annotation is all about.

1. What are we doing?

We are adding syntactic annotation to texts written by learners of English as a second language. We add this information by use of an annotation scheme we developed, and which focuses on annotating linguistic properties that are present in any given sentence, without making too much interpretation about what the learner meant to say, or what the correct form should have been. To achieve this end, our annotation scheme adds several pieces of linguistic information about each word, based on its context in the sentence, and based on the rules of English (the target language). It is important to note that we do not perceive non-native-like phenomena as “mistakes” or “errors”, but rather, as unique features of the learner’s developing language system (what is often called interlanguage). We are using dependency grammar to mark syntactic relations between words in a sentence—e.g., one word is the SUBJECT of another word.

2. What does the annotation look like?

The most important part of the annotation represents word-word relations. You’ll be using a graphical user interface (GUI) to mark these relations, as well as other linguistic properties, so we hope the annotation will be fairly intuitive. You’ll be using Brat to annotate (see appendix A.1), and you can see an example annotated sentence in figure 2. In this case, the word I is a first person nominative pronoun (PPIS1), functioning as the SUBJECT of have, marked as a present tense verb (VH0). Every word has
CHAPTER 1. GETTING STARTED

Figure 1.1: Example annotation

at least one grammatical function in the sentence (adJunCT, OBJect, MODifier, etc.), and you can see that there is an extra “word” added at the beginning (ROOT) to be able to mark the ROOT of the entire sentence.

3. How difficult is it to annotate dependency relations?

Unsurprisingly, sentences range from very easy to very hard. Complexity of syntax and learner proficiency, among other properties, lead to dramatic changes in difficulty.

If a sentence is native-like, then it should be relatively straightforward to annotate. It will simply be a question of understanding what the syntactic properties of the sentence are. In other words, you will need to be (or become) familiar with properties such as raising, coordination, wh-displacement, and so on. Constructions in native-like English can be rather complex, so when that happens, there may be some difficulty in determining how to apply the annotation—or in whether the annotation is even defined enough to cover the case in question. In such situations, ask for help, and we will make sure you understand the proper annotation and/or define a new annotation practice.

If a sentence has any non-native-like feature, then the real fun begins!

The next chapters will go into the details of the annotation scheme, after outlining some general principles.

1.2 General Principles

It helps to read the general principles before delving into the annotation scheme, so you can get the spirit of the annotation; understanding what we’re trying to accomplish will help you in cases where the annotation is not entirely clear.
1.2. **GENERAL PRINCIPLES**

### 1.2.1 Assumptions

There are a few guiding principles to use while annotating, and you should probably know these forwards and backwards:

1. Give the learner the benefit of the doubt. When annotating, if there is no way to disambiguate an analysis based on context, choose the analysis that leads you to:
   
   (a) fit the sentence in best with the overall context
   
   (b) posit a fewer number of mistakes/errors/deviations from target form(s)

   The learner may or may not have intended this possible analysis, but we are using the heuristic of assuming that the learner is correct rather than not. See section 1.2.2 for more details.

2. Try to assume as little as possible about the intended meaning of the learner. You will have to use context to disambiguate at times, but in general, we are looking to annotate *(morpho-)*syntactic information and not semantic information. This is one of the most challenging principles to follow, and more on this point is covered in section 1.2.3.

3. As much as possible, annotate the language “as is”. In other words, try not to think in terms of errors, but in terms of linguistic evidence (see section 1.2.4).

4. When you cannot determine the properties of a construction, consider whether it would be appropriate to underspecify the linguistic contents, so as not to say too much (section 1.2.5).

5. If an analysis cannot be determined by any of the above principles, also consider whether one grammatical form is more “primary” than another. By primary, we mean that it is the preferred dictionary form.

   For example, *have* can be a base form verb (dictionary form) or a present tense verb, and *deer* can be a singular noun (dictionary form) or a plural one. In lieu of other evidence, prefer the analysis which is based on the dictionary form.

### 1.2.2 Benefit of the doubt

In giving the learner the benefit of the doubt, we try to assume that the text is more well-formed than not. As mentioned above, this means that we try to fit a sentence into the context and assume as few errors as possible (see section 1.2.6 for more on errors).

Specifically, do the following: 1) fit the sentence into the context, if possible; 2) if not possible, at least assume that the sentence is syntactically well-formed (possibly ignoring the semantics—see section 1.2.3); and 3) if that fails, at least assume the word
is the word it appears to be (i.e., do not substitute another word in for the word which is present).

Let’s take the example of (1) to see how this plays out in practice. This sentence follows the sentence *I think that my life is mine* and thus most likely has some intended meaning of: “In Korea, however, what parents think is very important.” But if we annotate based on this (possibly) intended form, we will have to posit a number of errors, including a misformed possessive and a missing word (*what*), completely restructuring the sentence.

(1) In Korea, however Parents ’s think is very important.

Since strict topic coherence leads to many errors, we see if we can nonetheless obtain a well-formed grammatical analysis staying relatively within the bounds of the context. If we view ’s as a minor orthographic issue (see section 2.2) and however as a lexical issue (see section 2.4), then we can annotate this along the lines of *how parents think*, with the string Parents ’s together as a subject of think and however as a normal adjunct of think.

We will ignore the technical details of this analysis for now, but note that: a) the sentence becomes more or less well-formed once orthographic issues are glossed over, and b) we have to ignore the analysis based on our preferred intended form, in order to obtain the more grammatical analysis (see also section 1.2.3).

Giving the learner the benefit of the doubt also means that if something is grammatical, it is marked as such, even if it sounds somewhat non-native. That is, we are not concerned with marking every non-native property, but rather with marking the grammatical properties. In (2), for instance, the phrase *a banktrucptcy area* (misspelling aside) sounds less natural than something like *areas related to bankruptcy*. There is nothing grammatically incorrect here, however, so it is treated like any normal noun phrase.

(2) The most important aims of my life are to be a skillful and experience lawyer in a banktrucptcy area and to make my family more happy.

We will see the benefit of the doubt (i.e., “assume well-formedness”) heuristic play out in a number of ways, including selecting a lemma (section 2.2), distinguishing broken coordination structures from enumerated lists (section 3.3.3), dealing with “missing” elements (section 5.1), etc.

1.2.3 Semantics

It is important to understand that the annotation is not intended to fully capture semantic relations, i.e., meaning. While meaning can help disambiguate, there are cases where a syntactically well-formed structure (or better-formed structure) does not correspond to
a well-formed semantics or to some “intended” semantics. In these cases, we annotate the syntax.

This was illustrated above in (1), but consider also (3) (the next sentence in the same essay). It is relatively clear that admit means something like ‘was admitted’. However, that is not what we observe here; what we observe is an active verb (see also section 1.2.4), and we can provide an analysis for that active verb, including a subject (I) and an indirect object (to korea university). Ignoring the semantics of the particular verb, this is a reasonable analysis.

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

In the case of what i find my own way, the meaning is less clear, and thus it is even more important to steer away from the intended meaning and annotate the clause based on the fact that i find my own way is a valid sentence, and, with what bearing some function such as a complementizer, a clause can serve as a complement of decide (see discussion in section 1.2.7).

To take another example, in (4), one does not usually study degrees. However, study takes a noun object, so we would not mark anything unusual here, except for the fact that degree (cf. degree) requires a determiner.

(4) then, I study law degree.

Since we are less concerned about intended meaning and gloss over some non-native properties, there are many types of oddities we do not capture. If there is unusual word choice, or pragmatically-odd constructions, for example, but the sentence is syntactically well-formed, it will be annotated as if there are no problems. In (5), for instance, the phrase in each one of the spaces may sound odd, but is a valid prepositional phrase and is annotated as such.

(5) In this moment of my life, I have differents goals in each one of the spaces.

1.2.4 Evidence

The annotation is based on breaking linguistic properties down into separate pieces of evidence. Consider part-of-speech (POS), normally a concept summing over different kinds of properties. In the sentence I moused over the icon, for example, the word moused is a verb because: a) mouse(d) is the kind of word which can be a verb (lexical evidence), b) moused has an -ed verbal suffix (morphological evidence), and c) moused occurs in the standard position for a verb in English (distributional evidence).

For native speakers, the evidence generally all points to a single analysis, but this is not true with learners. Consider makes in (6): if forced to decide between a third person singular present tense verb (VVZ) and base form verb (VVO), it is nearly impossible to do so, because the morphological evidence (the -s suffix) and the distributional evidence (after can) conflict.
(6) Tin Toy can makes different music sound.

NP1 NP1 VM VVZ JJ NN1 NN1 Y
NP1 NP1 VM VV0 JJ JJ NN Y

This is why we posit multiple layers. In this case, the morphological evidence points to a third person singular present tense verb (VVZ), but the distributional evidence indicates that this position is most appropriate for a base form verb (VV0), as a modal verb selects for base form verbs. Thinking in terms of evidence means being very precise about what exact property you are annotating.

Most importantly, our dependencies prioritize morphological evidence over distributional, as discussed throughout section 3.1. For example, in (7), the word see has conflicting evidence. The morphological form is the base form or the non-3rd person singular present tense (VV0 either way); distributionally, the position is of a dependent of have, i.e., a past participle (VVN). While the two POS layers will capture the mismatch between morphology and distribution, our dependencies will be based on the morphological POS.

(7) I have see a movie.

Figure 1.2 shows the dependency tree, where the relations are based on the surface form of the tokens and the morphological POS tags, in this case VV0—and not VVN. Thus, we find two ROOT relations instead of a relation between have and see (see section 5.2.1 for a fuller discussion of this particular analysis).

Figure 1.2: Example dependency tree where the POS evidence conflicts

1.2.5 Underspecification

One feature of the scheme that follows from avoiding making claims about intended meaning is underspecification. This means we do not encode more information than what is visible from the text. For example, in the lemma field, if the word is a foreign word, misused in place of a regular English word, we leave the lemma underspecified.
For (8), for instance, while words like *sofa* have a straightforward lemma (*sofa*), the lemma for *caches* is left blank.

(8) Finally, Tim **caches** under a sofa.

This applies to any of the annotation layers, but we will go over a few examples for distributional POS tags (see section 2.3.2), as in most situations, it is straightforward to simply remove information from the tag. In (9), for instance, it is not totally clear whether *rabbit* is in the position for a singular noun (NN1) or a plural noun (NN2). Given the uncertainty, an underspecified distributional POS tag of NN is appropriate.

(9) In the park, they have a lot of litter animal such as colorful of bird, pretty squarel and shy *rabbit*.

In other cases, almost nothing can be said about the tag, and so we underspecify to ‘.’, essentially giving no information. In (10), for example, the distributional POS tag of *at* is not clear at all, as it seems to have no function in English. Thus, effectively, no POS tag is given.

(10) He begins to walk and **at** to run.

For dependency relations, we generally use the underspecified category ‘.’ when the two POS do not generally form a relation. In (11), for instance, we find a connection between *After* and *to start* (itself headed by to). This gives us a relation between a preposition/subordinating conjunction (ICS) and the infinitive marker (TO). But there is no dependency relation which fits that description, so we annotate it as ‘.’.

(11) **After/ICS to/TO start**, I want to tell . . .

Bear in mind two competing principles here: 1) we prefer using underspecification as little as possible, so as to be able to say something about the linguistic contents of a word or phrase, but 2) we do not want to impose a stronger analysis on the data than the evidence warrants (section 1.2.4).

Consider (12) in this context, where there are two odd constructions centered around *loudly* and *outside*. In the former case, *loudly* is an adverb appearing after a copula, and there is no relation which fits this, thus leading to an underspecified _ . In the latter case, on the other hand, *outside* seems like an adverb—illicit for a dependent of a preposition—but it also has a nominal use, e.g., on the outside. Thus, we have enough reason to mark it as a POBJ.

(12) Before I study in Bloomington, I could not sleep a lot every Sunday because there was very *loudly* at *outside*.

\[1\] Just to finish this example: the noun *outside* should select for an unrealized determiner (see section 3.4), and the preposition *at* will be marked as a lexical violation (see section 2.4).
1.2.6 Mismatches

The properties we are annotating are linguistic properties and not errors. In other words, you are not mapping each sentence to a target form; you are trying to describe what is actually there in the sentence.

It is worth noting, though, that because evidence conflicts, you are, in some sense, capturing many (though not all) errors. For example, in (13), the morphological tag of have is VV0, while the distributional will be VVZ.

(13) Everyone have plan in their life and try to get it.

There will also be mismatches between the dependencies that are realized and the dependencies that a word is looking for, encoded via subcategorization (see section 3.4). In (14), for instance, the word degree (cf. degree) is a singular noun which requires a determiner (DET). There will be no DET annotated in the tree, and the mismatch between what is in the tree and what the word subcategorizes for indicates some issue.

(14) then, I study law degree.

Errors arise only from mismatches in the types of properties we capture. For example, in (15), there is nothing wrong with the POS categories, and we can form a proper dependency tree, as in figure 1.4. The noun house selects for a DET, and it finds one (an), while own serves as a proper MODifier of house.

(15) . . . the second [goal] is to buy an own house until II twenty seven years old JJ . . .

In some of these cases, we mark lexical violations (see section 2.4)—as we would here for an—but if something is odd and you do not mark any mismatches, do not fret. As long as you follow the linguistic evidence, you are okay.

1.2.7 Mismatches to define shortest distance

While we do not base the annotation on mismatches or errors, they can be a useful way to help determine the most likely annotation, given that we are giving the learner the
1.2. GENERAL PRINCIPLES

benefit of the doubt (section 1.2.2). Mismatches arise when one layer of annotation contradicts another layer, and, with the goal of giving the learner the benefit of the doubt, we aim for the fewest number of mismatches.

Consider (16), where there is a potential issue with verb tense: visit is present (habitual) tense, while met is past tense. One way to annotate this would be to mark the morphological POS tag for met as VVD (past tense) with the distributional POS tag as VV0 (present tense). This would be one mismatch.

(16) I say partly because most of us visit a lot of other towns, countries, met a lot of people.

But an alternative annotation is to treat met as a misspelling of meet (since the difference between e and ee is a “reasonable” spelling error; see section 2.2.2). In this case, both POS tags are VV0, giving zero mismatches. We thus prefer this second analysis.

As a more complicated example, consider again (17), the same as (3) from section 1.2.3.

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

There are a number of possible analyses for handling what i find my own way, some of which we list here:

- Analyze what as an extraneous word with no clear function (1 “mismatch,” ignoring evidence) and with a missing auxiliary (e.g., would) (1 mismatch, between subcategorization and dependencies, positing unknown information).

- Analyze what as a type of infinitival marker (1 mismatch between morphological POS and distributional POS), with i as an extraneous word (1 mismatch between subcategorization and dependencies, ignoring evidence).

- Analyze what as a complementizer, albeit a lexical violation (1 “mismatch,” see section 2.4) and the clause as a valid, even if odd, sentence.

This last analysis, regardless of whether or not it captures the intended meaning (section 1.2.3), still posits the fewest number of mismatches between different pieces of evidence and fits in with the syntactic evidence. Generally speaking, if you are sticking close to the syntactic evidence, this should go hand-in-hand with the fewest number of layer mismatches.
1.3 Label Inventories

This is simply a listing of all the labels you’ll be using, with brief definitions.

1.3.1 Dependency Relations

See section 3.1.1 for more complete descriptions of these labels.

1. SUBJ - subject
2. CSUBJ - finite clausal subject
3. XSUBJ - non-finite clausal subject
4. OBJ - object
5. OBJ2 - second object
6. IOBJ - indirect object
7. LOC - locative
8. COMP - finite clausal complement
9. XCOMP - non-finite clausal complement
10. PRED - predicate
11. CPRED - finite clausal predicate
12. XPRED - non-finite clausal predicate
13. POBJ - prepositional object
14. VC - verbal complement
15. JCT - adjunct
16. CJCT - finite clausal adjunct
17. TAG - tag questions
18. XJCT non-finite clausal adjunct
19. NJCT - nominal adjunct
20. MOD - nominal modifier
21. CMOD - finite clausal modifier
22. XMOD - non-finite clausal modifier
23. DET - determiner
24. QUANT - quantifier
25. PQ - post-quantifier
26. NEG - verbal negation
27. SRL - serial verb
28. CPZR - complementizer
29. POSS - possessive
30. PRT - particle
31. COM - communicator
32. VOC - vocative
33. TOP - topicalization
34. INCROOT - incorrect root
35. TRANS - transition
36. PUNCT - punctuation
37. ROOT - root
38. NAME - proper name element
39. DATE - date element
40. ENUM - enumeration
41. COORD - coordination
42. MCCOORD - modificatory coordination
43. CCC - coordinating conjunction
44. APPOS - appositive
45. ELL - elliptical relation
46. _ - unspecified

1.3.2 POS Tags

We base our listing here on Sampson (1995, p. 105–120). The tagset, as we use it, is organized hierarchically. Use as specific of a tag as possible when tagging. For example, if you are examining a singular common noun, use NN1, but if you cannot tell plurality, use NN; if you cannot even tell commonness or properness, use N.

For tags which list specific words, note that those words can be ambiguous (e.g., DD1 for *that* as a determiner, pronoun, or qualifier, while CST for subordinating conjunction). The SUSANNE guidelines provide more information. Note, though, that many multi-word expressions in SUSANNE are handled by us by treating them as separate words (see section 4.12).
CHAPTER 1. GETTING STARTED

- VROOT - dummy ROOT of sentence
- A - articles/possessives
  - APPG - possessive pronoun (her, its, their, my, our, his, your)
  - AT - the
  - AT1 - no, a, an, every
- BTO - in order introducing infinitive
- C - conjunctions
  - CC - coordinating conjunction (and, and/or, plus, &, or, nor)
    * CCB - but as coordinating conjunction
  - CS - subordinating conjunctions
    * CSA - as as subordinating conjunction
    * CSN - than in all uses
    * CST - that as subordinating conjunction
    * CSW - whether in all uses
- D - (non-article) determiners
  - DA - own (gen.), former, latter, same, selfsame
    * DA1 - much, little
    * DA2 - many, few, several, fewer, fewest
    * DAR - more, less
    * DAT - most, least
  - DB - all, half
    * DB2 - both
  - DD - yon, yonder, somewhich, enough, some, any
    * DD1 - that, either, this, neither, another, each
    * DD2 - these, those
    * DDQ - what, which, whose, whichever, whatever, whichever, whatsoever
- EX - existential there
- F - non-English terms (including formulae, foreign words, etc.)
- GG - genitive/possessive ‘s or ’
1.3. LABEL INVENTORIES

- I - prepositions
  - ICS - preposition or subordinating conjunctions (considering, notwithstanding, like, after, before, ere, since, until, till, but, except, save)
  - II - general preposition (including as, by, per, to), plus: for, of, with, without

- J - adjectives
  - JA - adjective used only predicatively (e.g., ablaze, unable)
  - JB - adjective used only attributively (e.g., chief, entire, major), including other, only, & very (e.g., the very/only thing), including comparative & superlative forms (e.g., lesser, utmost)
  - JJ - general adjective (including -ed & -most forms)
    * JJR - comparative adjective (including elder, further)
    * JJT - superlative adjective (including eldest, furthest)

- LE - pre-coordinators (both, either, neither)

- M - numbers
  - MC - cardinal number (e.g. two, 3)
    * MC1 - one as a numeral
    * MC2 - plural of cardinal (e.g., ones, twos)
  - MD - ordinal number (e.g., fourth, next)
  - MF - fraction

- N - nouns
  - ND1 - direction (e.g., north, N)
  - NN - common noun (including nouns with unclear plurality, e.g., the arts, or irregular plurality where singular=plural, e.g., sheep); mass nouns which can have plural readings (e.g., fish, measles) are tagged with NN
    * NN1 - singular common noun
      Usually, if the automatic preprocessing has marked the noun as NN1, keep it as such (e.g., snow), vs. marking it as NN (e.g., Chinese)
    * NN2 - plural common noun
  - NP - proper noun
    * NP1 - singular proper noun, including geographic names, etc.
    * NP2 - plural proper noun
CHAPTER 1. GETTING STARTED

* NPD - day of the week (e.g., Tues.)
  - NPD1 - singular day of the week (e.g., Tuesday)
  - NPD2 - plural day of the week (e.g., Tuesdays)
* NPM - month name (e.g., Oct.)
  - NPM1 - singular month name (e.g., October)
  - NPM2 - plural month name (e.g., Octobers)

• P - pronouns
  – PN - general pronouns
    * PN1 - anybody, anyone, anything, everybody, everyone, everything, naught, nobody, nothing, somebody, someone, something, one, so (cf. do so)
    * PNQ - interrogative pronoun (whom, who, whoever, whomever, whosoever)
    * PNX - oneself
  – PP - personal pronouns
    * PPG - hers, theirs, mine, ours, his, yours
    * PPH - subject/object pronouns
      - PPH1 - it
        - PPHO
          > PPHO1 - her, him
          > PPHO2 - them
        - PPHS
          > PPHS1 - she, he
          > PPHS2 - they
    * PPI
      - PPIO
        > PPIO1 - me
        > PPIO2 - us
      - PPIS
        > PPIS1 - I
        > PPIS2 - we
    * PPX
      - PPX1 - herself, itself, myself, himself, yourself
      - PPX2 - themselves, ourselves, yourselves
    * PPY - you
1.3. LABEL INVENTORIES

- **R** - adverbs
  - **RA** - nominal adverb / postnominal adjectives (e.g., ago, since, etc. ff., respectively, else, a.m./p.m., galore, whatever, whatsoever, apiece, each, B.C.)
  - **RG** - qualifier (e.g., very, mighty, as, quite, too, about, around, circa, under, over, some, rather, so, indeed, enough, how, however)
  - **RL** - adverb of place/direction (e.g., aside, downtown, elsewhere, here, there, downstairs, upstairs, somewhere, anywhere, everywhere, nowhere) – see list on p. 121-122 of Sampson (1995).
  - **RP** - adverbial uses of across, down, in, off, on, out, over, through, up, about (including about to)
  - **RR** - general adverb (e.g., fast, skillfully, including: far, long, otherwise, yet, only, any, so & enough as clause adverbs; also, namely and other single words marked by REX in the SUSANNE original set)
    - **RRQ** - interrogative or relative adverb (how, when, whence, where, whereabouts, whither, why, whereby, whereupon) and wh-...-ever adverb (however, wherever, whenever)
    - **RRR** - single-word comparative adverb other than more or less (e.g., better, closer, etc.)
    - **RRT** - single-word superlative adverb other than most or least (e.g., best, closest, etc.)
  - **RT** - temporal adverb (again, hereafter, overnight, then, now, today, tomorrow, tonight, yesterday)
- **TO** - infinitival to
- **UH** - interjection (e.g., blimey, hello, please, well, yes)
- **V** - verbs
  - **VB** - be verbs
    - **VB0** - be
    - **VBD** - was/were
    - **VBG** - being
    - **VBM** - am
    - **VBN** - been
    - **VBR** - are
    - **VBZ** - is
  - **VD** - do verbs
CHAPTER 1. GETTING STARTED

* VD0 - do
* VDD - did
* VDG - doing
* VDN - done
* VDZ - does

- VH - have verbs
  * VH0 - have
  * VHD - had as past tense
  * VHG - having
  * VHN - had as past participle
  * VHZ - has

- VM - modal verb (could, might, should, would, can, dare, may, must, need, shall, will) - including ought, used as modal catenative (cf. ought to ...)

- VV - all other verbs
  * VV0 - base form verb
  * VVD - past tense verb
  * VVG - present participle verb (including going as catenative)
  * VVN - past participle verb (including bound as catenative)
  * VVZ - third person singular verb

- XX - not
- Y - punctuation
- ZZ - symbols
  - ZZ1 - singular letter of the alphabet, including when spelled out (e.g., Eta)
  - ZZ2 - letter of the alphabet with plural inflection (e.g., esses)

- _ - underspecified tag
Chapter 2

Initial Annotation Layers

In this chapter, we give an overview of the different layers of annotation, prior to annotating dependencies: lemmata (section 2.2), part-of-speech (POS, section 2.3), and lexical violations (section 2.4).

You can view lemmata and POS as providing a platform for dependency annotation: each (morphosyntactic) dependency relation should be based on the lemma and the morphological POS tag which are given. Thus, it helps to first annotate/correct these fields before annotating dependencies.

Before examining these layers, we first outline changes you might need to make to the files regarding sentence segmentation and word tokenization.

2.1 Segmentation & Tokenization

You will be annotating a whole essay at a time, meaning that you will have access to the surrounding context. Before actually applying any annotation, you should first read through the file to check sentence segmentation and word tokenization. Fixing segmentation (and, to some extent, tokenization) issues is easier if it is done before any other annotation.

If you find any sentence or word which needs adjustment, contact a project leader immediately, so that they can fix it before you annotate.

2.1.1 Sentence segmentation

Learners do not always use punctuation consistently, leading to run-on sentences and sentences which are split that should be joined. We generally bias towards not re-segmenting, but to make annotation easier, in some cases we would like to correct any mis-segmentations which majorly affect grammaticality, so that annotation is on a more strictly intra-sentential level. We want to preserve as much of the learner’s sentence
segmentation as possible, but if you can only form a proper tree by re-segmenting, then we need to do so.

In (18), for instance, the automatic preprocessing treats these two sentences as two separate sentences. Without combining the sentences, however, it is very difficult to make a full clause out of the weather. Thus, giving the learner the benefit of the doubt, we want to combine these two sentences into one. Similar to misspelling diagnostics below, we do this because it is only a typographical issue, i.e., it is simply a period in the wrong spot.

(18) a. Sometime I do n’t like it because it ’s very complecated and in my country the weather .

b. It ’s very hat .

Likewise, you will find sentences that are actually several sentences with several roots. In some cases, it can be easier to annotate if we first split them into several sentences.

Having outlined this, however, an important point is that we mainly want to re-segment when grammaticality is at stake and the entire analysis is changed. Your default should be not to re-segment.

As an example of why we do not encourage re-segmentation, consider (19), in particular the clause in (19b). Since this clause does not form a full sentence with a verb, it is tempting to re-segment it, with either the previous or the following sentence.

(19) a. Lost , my country has too many people , there already every plac has person .

b. Whenever in the morning or in the evening .

c. You can n’t find a place own yourself .

We do not re-segment it, however, for a variety of reasons. First, it is not clear whether the clause goes with the previous (19a) or following (19c) sentence. Secondly, and more importantly, each of those two sentences are complete sentences, so their analyses would not change significantly, unlike in (18). Finally, although not a full sentence with a verb, (19b) is a relatively complete thought, the type of thing which could serve as a statement within a normal discourse.

Thus, our analysis is to use INCROOT, applying it to whatever can be determined as the most head-like element, in this case Whenever. See section 4.2 for how to define the INCROOT of a sentence and section 5.1.1 for INCROOT cases in the context of missing heads.

In some cases, there are clearly multiple verbs that cannot otherwise be connected—i.e., you have a run-on or a comma splice. For (20), for instance, there are two sentences connected by a comma. These will be treated in keeping with “broken” conjunction
Whenever in the morning or in the evening.

Figure 2.1: An incomplete/incorrect ROOT (INCROOT)

structures, as described in section 3.3.3 under Missing conjunction, and do not need re-segmentation.

(20) They taught me usual things, these were not grammar.

2.1.2 Tokenization

As with sentence segmentation, the automatic tokenization done before annotation is rather crude. Thus, for example, you might find something like Dr and . as separate tokens when they should be one token.

In general, we do the following:

1. Punctuation is treated as a separate token, unless it is part of an abbreviation.

2. Contractions are split into their component units, e.g., don’t becomes do and n’t.

3. Multi-word expressions (MWEs)—e.g., in spite of, New York—are treated as separate units, and our syntactic representation links them (see section 4.12).

If any token(s) appear(s) not to conform to the standards above, re-tokenize. The annotation tool generally lets you re-tokenize - see section 2.2.4 for how to deal with spacing issues. If there is a tokenization problem you cannot fix, you may need to contact a project leader.

2.2 Lemma

The lemma field is simply a way to record the lemma (≈ dictionary form) of each token. Any inflectional morphology is removed: thus, a verb like made is annotated with the lemma make, and a noun like children becomes child. (Note that to is a separate word, having its own lemma, so verb lemmas are like make and not to make.) In the case of misspelled words, if the misspelling is influenced by the way it is pronounced, we write the target spelling of it; for example, flexsible becomes flexible.
CHAPTER 2. INITIAL ANNOTATION LAYERS

Note that we only lemmatize inflectional morphology. There are only a handful of inflectional morphemes in English (plus irregular forms): noun plural (-s), adjective comparative (-er), adjective superlative (-est), verb third person singular present tense (-s), verb past tense (-ed), verb past participle (-ed/-en), and verb present participle (-ing). Derivational morphemes, such as the noun -er or -tion endings should be kept in the lemma. For example, the lemma of bigger is big, while the lemma of writer is writer. In the case of some phonetic variations, such as an for a, we normalize to the more basic form, e.g., a for both.

Note one detail with this criterion of focusing on inflectional morphology: some morphemes can be either inflectional or derivational. In (21a), for instance, the -ing morpheme is an inflectional verb ending, meaning that the lemma of traveling is travel. In (21b), however, the use here is nominal; thus, the -ing ending is a derivational morpheme added to a verb to change it to a noun, and the lemma is traveling.

(21)  
   a. We were traveling a lot last year. (constructed example)  
   b. I have been to U.S.A for traveling when I was 14 years old.

In general, we wish to stay as close as possible to the surface form. If the word is unrecognizable or ambiguous, underspecify the lemma slot in that case. Only use this as a last resort, however.

One point that will become important as you annotate more lemmas is that lemma annotation is indicating what the abstract form of the word is. In other words, you are determining what word (including its general word sense—see Government and agreement in section 2.3.2) is being used in a given sentence. Once you have determined the lemma, much of the rest of the analysis becomes easier, as your choices for annotation are more limited.

2.2.1 Irregulars

As a minor point, remember to convert irregular forms into their proper lemma, e.g., were becomes be, geese becomes goose, best becomes good, etc.

We treat modal verbs (e.g., could) as full-fledged forms; e.g., the lemma of could is could, not can.

2.2.2 Misspellings

As mentioned above, we allow for the correction of misspellings, when such misspellings are phonetically or typographically-driven (but not semantically-driven). Many of these cases are ones which native speakers also would make, e.g., confusing affect and effect. Consider the misspellings in (22): it is clear that there are simple spelling mistakes with ballence, organe, and dameged, and thus they can be lemma-annotated as balance, organ, and damage.
2.2. LEMMA

(22) . . . like body that will **loss** its **ballence** when one of two **organe** get **dameged**, a country also can **loss** its **ballence** when part of its people live in sadness.

Trickier is the case of **loss**. In this case, it seems like **loss** may be a misspelling of **lose**, yet **loss** is a valid word in its own right. One diagnostic you can use is to ask: is this a phonetically-driven or typographically-driven mistake which a native speaker might make? Tying that in with the heuristics from section 1.2.1, one can decide as follows:

- Because of the non-obvious rules for vowel pronunciation in English, **loss** is an easily-misspelled variant of **lose**, even for native speakers.
- Giving the learner the benefit of the doubt, it is preferable to have a verb in this location, giving a well-formed grammatical tree.

This last point is the most important: our POS and dependency annotation are based on the lemma field. Thus, if we annotate the lemma as **lose**, then the tree will follow in a straightforward manner. If we annotate it as **loss**, then we have a more complicated structure to determine.

In many cases, there is a debatable lemma, but it does not affect the POS or dependency annotation. Consider **merjors**, **arquisisiton**, and **cares** in (23a): 1) **merjors** is fairly clearly a phonetic rendering of **mergers** and so the lemma is **merger**; 2) **cares** is fairly clearly too far away from **cases**, so the lemma remains **care**; and 3) **arquisisiton** is debatably close to **acquisition**, especially since no other similar English word exists. Regardless of the distance, though, note the nice property that all are nouns, and thus POS and dependency relations will be relatively easily handled.

(23) a. My majors are bankruptcy, corporate reorganizations, and **merjors** and **arquisisiton cares**.

b. After to start, I want to tell that this **excersice** is very important in the life, no only as a **prison**.

We see similar types of cases in (23b), where **excersice** is a phonetic rendering of **exercise**, but **prison** is less clear. The vowel difference (e vs. i) is a simple orthographic difference, but the transposition in letters is questionable. We deem this too far away to annotate as **person** (i.e., the lemma is annotated as **prison**), but note again that it makes little difference for the rest of the annotation. We will return to **no** in a moment.

We have stipulated a phonetic similarity, in order to handle cases such as (24a). In this case, some annotators might be tempted to lemma-annotate **something** as **some**. But there is more than phonetic distance here, so one should annotate it as **something**.

(24) a. A baby crawl around the room, and found **something** interesting toyes.

b. ... I decide **what** i find my own way.
Likewise, in (24b), it may be tempting to lemma-annotate *what* as *that*, but this is more than phonetic distance. In both cases, changing the lemma would force a particular analysis of the data, which *something* and *what* on their own do not necessarily warrant.

Note that this is the same issue for *no* in (23b). Although *not* would give us an XX tag and a natural negation of *only*, what we observe is *no*, an AT word. (The relation between *no* and *only* will thus be underspecified—see section 1.2.5). It is close, but not close enough.

Sometimes the question of phonetic distance is complicated because multiple forms are within that distance, or a misspelling leads to a different correctly-spelled word. In (25), for instance, the word *met* could be a misspelling of *meet* since it is simply a question of similar vowels. If the form really is *meet*, then it properly coordinates with either *come* or *study*; if the latter case (*study ... meet*), then *meet* is clearly a form which would give a well-formed clause. However, the following sentence, in (26), discusses the people that the person met (past tense). Combining the whole context of the essay with a bias towards not changing forms that appear, we base the annotation on *met* here. In either case, the lemma is *meet*, but the morphological POS will differ, depending on what form it is, being either VVD (past tense) or VV0 (non-third present tense).

(25) I come to us from Kazakhstan to study English and *met* a lot of people from different countrys.

(26) Most of us have different views on the same things.

Note in passing the easier case of *us* in (25), which should be lemma-annotated as *U.S.*, a simple change due to typographical conventions.

### 2.2.3 Spelling vs. Morphology

Misspellings are confusable with incorrect morphology. To take one example, consider (27). In this case, we can either lemma-annotate as *different*, by removing the (non-appropriate) -s ending, or as *difference*, by virtue of a phonetic similarity.

(27) In this moment of my life, I have *differents* goals in each one of the spaces.

Thus, the two main possible analyses here are:

1. Morphological issue: Annotate the lemma as *different*. This will lead to a morphological POS tag (NN2) vs. distributional POS tag (JJ) disagreement (section 2.3.1), because the -s ending morphologically indicates a plural noun.

2. Phonological/Spelling issue: annotate the lemma as *difference*. This will also lead to a POS disagreement, because now the form is morphologically NN1, but again distributionally JJ.
In some sense, because both posit a POS mismatch, the analyses are equivalent in terms of number of errors/mismatches (see sections 1.2.6 and 1.2.7). But another piece of evidence we have is that goals is plural, giving a preference for different as the lemma, i.e., different is matching the noun in plurality. Furthermore, by annotating the lemma as different, then the learner is given more credit for selecting the appropriate word in this context.

Note that the morphological POS is weighted towards the suffixal information, as opposed to the stem (which is clearly adjectival in this case once annotated as different). It is the -s ending is the reason we pursue a noun analysis. Also, the pseudo-agreement with goals is also one reason that we will annotate the POS as a noun (NN2), as opposed to a verb (VVZ), as the -s ending can, in general, also be a verbal ending.

2.2.4 Spacing issues

Note that the rules for misspellings also apply to mis-spaced items, as in (28) and (29), featuring an unnecessary space and a missing space, respectively.

(28) I had a problem a bout chooses my car . . .
(29) The tin toy had adram and aacordion.

These are simply typographical conventions, so we want the final tree to reflect analyses for about and a drum, again giving the learner the benefit of the doubt. (It is less clear whether to treat aacordion as having an a as its lemma or not; since there is an extra a and a better analysis with it, we favor adding it.)

The question is: how do we do this technically? In both situations, you will need to adjust the tokenization within Brat. In the case of an added space, remove any annotation covering the two words (e.g., a and bout) and then drag the cursor to highlight both words at once (i.e., a bout). You should then be able to annotate the two words as if they were one, as in (30).

As for missing spaces, you will need to annotate within a word. Thus, start by removing any annotation for the word as a whole and instead highlight the relevant individual parts. In (31), for example, the first part of the word is lemma-annotated as a and the second as drum; crucially, they are treated as two separate tokens, which can then be annotated with separate POS and dependency relations (in this case a determiner relation between the two tokens). (We use a plus sign here, but in the Brat annotation tool, this is achieved by highlighting each part of the word as needed; see appendix A.1.)

(30) a_bout/about
(31) a/a+dram/drum

The rules for orthographic distance still apply for spacing issues. In (32), for instance, the lemma for here by can be annotated as hereby (one token), but not whereby.

(32) I had a problem a bout chooses my car . . .
(29) The tin toy had adram and aacordion.

These are simply typographical conventions, so we want the final tree to reflect analyses for about and a drum, again giving the learner the benefit of the doubt. (It is less clear whether to treat aacordion as having an a as its lemma or not; since there is an extra a and a better analysis with it, we favor adding it.)

The question is: how do we do this technically? In both situations, you will need to adjust the tokenization within Brat. In the case of an added space, remove any annotation covering the two words (e.g., a and bout) and then drag the cursor to highlight both words at once (i.e., a bout). You should then be able to annotate the two words as if they were one, as in (30).

As for missing spaces, you will need to annotate within a word. Thus, start by removing any annotation for the word as a whole and instead highlight the relevant individual parts. In (31), for example, the first part of the word is lemma-annotated as a and the second as drum; crucially, they are treated as two separate tokens, which can then be annotated with separate POS and dependency relations (in this case a determiner relation between the two tokens). (We use a plus sign here, but in the Brat annotation tool, this is achieved by highlighting each part of the word as needed; see appendix A.1.)

(30) a_bout/about
(31) a/a+dram/drum

The rules for orthographic distance still apply for spacing issues. In (32), for instance, the lemma for here by can be annotated as hereby (one token), but not whereby.
While *whereby* could make for an easier syntactic tree, it is not clear that it is a correct interpretation, and it requires a change in pronunciation.

(32) ... the education system is worst and the educational level of the students is too low, *here by*, the government helps students to get knowledge ...

In some cases, it might help to say the sentence out loud to determine whether it really is a phonetic issue. Repeating (33) (= [1]) from section 1.2.2, the orthography may make you miss the fact that *Parents’s* can be read as *parents*. In that case, there is a more or less reasonable tree to construct. (Note that if you tried to annotate *Parents’s* as a possessive structure, you would have to deal with the fact that *think* is not a noun form.)

(33) In Korea, however *Parents’s* think is very important.

Hyphenated forms In example (34), we see a hyphenated form which could be written as a non-hyphenated form by some (native or non-native) writers. Do we treat this as one, two, or three words?

(34) two-year-experience

In general, we prefer staying true to the original hyphenated form, i.e., not creating separate tokens. We do advocate breaking it apart in cases where a native speaker would not use (or is not supposed to use) hyphens. For *two-year-experience*, for example, writers of English are generally supposed to include a hyphen for *two-year* (because neither word is an adjective, though together they function as one), but not between that adjectival form and the noun *experience*. Thus, we treat this as two tokens: *two-year* and *experience*.

2.2.5 Lowercase

As a minor point, lemmas are in lowercase, except when the token is a proper noun. That is, sentence-initial words which are not proper need to be lowercased for the lemma. If it is not clear whether a noun is being used as a proper one or not, bias towards lowercase. For more on proper nouns, see section 3.3.2.

2.2.6 Anonymization

You will occasionally find tokens like *BNAME* or *SNAME*. These are abbreviations which were used to keep the data anonymous, e.g., *BNAME* would replace a name starting with *B*, like *Bridget*. Leave them as, e.g., *BNAME* in the lemma field and annotate like a proper noun (section 3.3.2).

---

1In this case, the token *here by* will be be an unspecified dependent of *gets*. 
2.3. POS

2.2.7 Acronyms, unknowns, & foreign terms

Acronyms are treated as regular words and not decomposed into their component parts. Thus, for example, the lemma of MBA is MBA.

Note that if you do not know an acronym or a term in general, you are free to use Google and try to figure out what it refers to. In (35), for instance, the annotation of Korea University depends upon whether you are aware that Korea University is the name of a university or not. (It is.)

(35) So when I admit to Korea University, I decide what I find my own way.

Sometimes words are being in a way which does not match any English word, whether because the word is foreign or because the misspelling is severe enough that there is no word within a reasonable distance from the form. In (36), for example, the word enterprenerius is too phonetically different from any English word, so the lemma is marked as UNKNOWN. Note that you should write out UNKNOWN, as opposed to leaving the field blank, so as to make clear that there is annotation here.

(36) ... I will create my own office with the objective to advice enterprenerius ...

In passing, also note that advice can have a lemma of advise here, as c and s are reasonable phonetic substitutions.

You will sometimes find foreign terms used as clearly-demarcated and properly-used foreign terms, as in (37). In such cases, each lemma is the word itself (no morphological analysis of a foreign term is needed), and the analysis will follow a fairly straightforward pattern for all cases (see section 3.3.2 on Compounds: foreign terms).

(37) the Instituto Del Progreso Latino center (constructed example)

2.3 POS

Part-of-speech tags (POS) are encoded using a simplified version of the SUSANNE tagset (Sampson, 1995). The full tagset we use is outlined in section 1.3.2. More detail for annotating a POS tag is given in the SUSANNE guidelines, in particular chapter 3 of Sampson (1995). Note that we use a “mapped” version of the SUSANNE tagset, ignoring any lowercase letters in the POS tags and also collapsing many cases. Thus, for nouns (NNL), we use the tags NN, NN1, and NN2, but do not distinguish, e.g., NN1 from NNL1 or NN1m from NNL1ux. As you will be using the tags given in the user interface, this is easily enforced.

We split the POS into 2 layers: one for morphological (or, morphosyntactic) information, and the other for distributional information (see section 2.3.2). In example (38), for instance, makes has the morphological form of 3rd person singular present tense
(VVZ), but, in its position following can, its use is as a baseform verb (VV0). (This labeling was shown in (6).)

(38) Tin Toy can makes different music sound.

Before you work with the data, we automatically tag the tokens with the POS tags, so you should check and see whether the assigned tag is the correct one, and then see if one, or both, of the tags needs to be changed, either because the tagger was wrong or because there is some novel, non-native-like use. If it is non-native, there is often a need to change at least one of the tags.

If it is not clear how a token is being used, you may have to underspecify a POS tag, usually a distributional POS tag. For example, consider the word sound in (38) above. The morphological form is singular (NN1), but the learner may be using this form as either a singular or plural noun (NN1 or NN2). Therefore, the distributional tag is underspecified (NN) (see section 2.3.2).

When underspecifying, include as much information as possible, i.e., as can be deduced. In this case, for example, NN is preferable over N or over the completely uninformative tag . The fact that the user interface is organized hierarchically may help in this respect (see section A.1).

Your underspecifications are limited by what is in the tagset. For example, in (39), it may not be clear in context what tense the learner intended to use distributionally.

(39) The child follow him.

With the current set of tags, we annotate this as VV (verb) instead of VV0 (baseform), VVD (past tense) or VVZ (third person singular present tense). However, you may want to state that, even if the tense cannot be determined, the verb is tensed, e.g., creating a tag VVT. Report any cases where you could not give as much information as you would have liked.

2.3.1 POS mismatches

As mentioned, you will often find POS mismatches. This happens when the morphological evidence conflicts with the distributional. Importantly, the lemma is the same; it is usually just a question of inflectional morphology. This will become important in distinguishing cases where POS mismatches are utilized from cases where there is a lexical violation (section 2.4.2). As mentioned with (39) above, in many cases you will wind up with an underspecified tag, often for a noun or a verb, since the morphology does not match the context.

To take a different case, in (40), we have the adjective full occurring in a normally nominal slot, after a preposition. Thus, its morphological POS tag is adjective (JJ), while its distributional tag is noun (NN).
Second to live with my family in full of love, I would like to have enough time with my wife here.

2.3.2 Defining distributional POS

Distributional POS tags are perhaps the most difficult annotation layer. We define a 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. The difficulty lies in pinpointing exactly what is predicted, and so we outline how to select a distributional tag when it is distinct from the morphological tag.

1) Use syntactic predictions

In making a prediction, try to stay rooted in syntactic predictions. In (41), for example, it is clear that place has something odd about it, due to the fact that there is no determiner and no plural marking. Semantically, it is most likely that the learner is referring to a singular place (since most people are aiming to live in only one place with their spouses), but syntactically, this is a perfectly valid plural position.

(41) Also, I need more natural and friendly place to live with my wife . . .

Secondly, when we speak of predictions, we are talking about categorical constraints, not specific lexical constraints. Consider example (42), where an appears nonnatively before experience. The distributional slot before a noun is appropriate for a determiner, so we may say that the POS tag is a determiner (AT1), even though lexically experience does not need it. The non-nativeness is encoded by the subcategorization annotation for experience, which does not select for a determiner (see section 3.4).

(42) Graduating from commercial and accountancy Faculty in Thailand and having an experience working as Product Movement and Inventory at ExxonMobil ,

An important point here is that you are annotating the words as they appear; experience may not select for a determiner here, but there certainly is one present.

Finally, remember that the whole context of a sentence plays a role in determining what the correct distributional analysis should be. Consider (43), where we have to decide what the distributional tag of one is. In particular, is it singular (MC1) or plural (MC2)? On the one hand, there is no determiner, which seems to make it a plural context. But on the other hand, verbs like switch seem to require parallelism of its dependents—i.e., if there is only one lifestyle in the first dependent, then it follows (syntactically) that the second dependent is also singular.

(43) . . . I realized we have to switch our wasteful lifestyle to more efficient one .
In this case, because the evidence conflicts, use the underspecified MC. Distributional POS tags require you to consider evidence from the whole sentence (e.g., parallel structure) and more local evidence (e.g., lack of a determiner). It is tempting to use MC1, to posit fewer errors (since it would match the morphological tag), but that goes beyond giving the learner the benefit of the doubt and gives more than they deserve.

2) Start with the morphological tag

A rule of thumb, especially when underspecifying a POS tag, is to try to stay as close to the morphological POS tag as possible. Consider (43) again: in this case, the numeral (one) appears in a strange context. Note that this context could include pronouns (tags starting with P) or nouns (tags starting with N), among possibly others—but we first try to fit a distributional tag which is consistent with the morphological one. Thus, MC or MC1 would be better choices than NN or NN1, even though such tags are also appropriate in this context as the object of a preposition.

Looking at (44), we have a potentially more challenging case, that of determining how what is being used distributionally. Again, multiple tags could apply here, such as PN (cf. something) or NN (cf. ways). But what is morphologically DDQ, so if D tags (determiners) can work here, then tag it as such. Indeed, some determiners (e.g., much/DA1) are appropriate here, even if what is not. Thus, D is the distributional tag.

(44) Although I’ve developed my business skill in marketing for past seven years, I found what to further improve my ability and deepen my communication skill.

In general, then, an appropriate way to assign a distributional POS tag is to first try alternate tags within the same main POS category and if nothing specific works, try “backing off” from the morphological tag. Only after this fails, seek other POS tags.

3) Dealing with conflicting predictions

As mentioned, distributional evidence can conflict; we outline some ways to resolve conflicts here, when two words provide differing evidence about each other.

Two-word conflicts: Government

Syntactic government is when the form of one token is determined by another. In these types of cases, the distributional POS follows from subcategorization (see section 3.4). For example, in the constructed (45), the preposition with selects for a prepositional object, governing the case of the object. Distributionally, then, he is in a prepositional object position (PPHO1), regardless of the fact that morphologically it is a subject pronoun (PPHS1). Distributional tags are rarely this cleanly defined, but this should help you get a sense of what distribution refers to.

(45) I must play with he.
This works similarly for adjuncts, even though they are not subcategorized for. In the constructed (46), for example, quick is an adjective form (JJ) modifying a verb. As it is modifying a verb, the distributional tag adverb (RR), appropriate for a verbal head.

(46) He runs quick

Two-word conflicts: Agreement For cases of agreement, the head is not always the source of agreement. Consider the constructed (47), where the subject-verb disagreement affects the forms of both tokens. In this case, the verb is the head, but the subject is what we consider the source of agreement.

(47) He sleep.

On the level of POS, then, annotate sleep with a distributional tag of a third person singular verb (VZ), while leaving He as is (PPHS1). For determiner-noun disagreements (e.g., these cat), generally treat the noun as the source of agreement, in lieu of any evidence to the contrary. Thus, in this example, these would be distributionally singular (DD1), as the adjacent token cat predicts its properties.

This heuristic is only “in lieu of evidence to the contrary.” If the context indicates properties of the noun, then the noun may be where the distributional POS tag is changed. In (48), for instance, it is not clear whether aspects is being used in a singular or plural fashion. Thus, we underspecify that distributional POS tag to NN, while the distributional tag of another is still the singular DD1. In general, we only try to modify one POS tag when there is an agreement conflict.

(48) ... to obtain such knowledge brings me another aspects to do my jobs.

Non-agreement can be trickier when a word is in an ambiguous or unclear context. Singular noun forms, for example, often occur without articles and without plural markings. In (49), for instance, community occurs in a position where the lack of a determiner would, in some sense, predict a plural form. However, it is hard to argue for a particular tag based on the lack of something, i.e., to reason from non-evidence.

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

In this case, start with the singular form (NN1), since that is what is present. The lack of a determiner opens up the possibility of a plural form (NN2). Without further evidence, allow both possibilities and underspecify, giving NN.

One further note on this example, relating back to the discussion of lemma annotation (section 2.2): community has a sense in which it can be used as a mass noun (e.g., We were building community by having dinner together), but this is a rarer sense and not the one in use here. Although you generally have to more or less shut off semantics in your mind (see next paragraph), try to stay within the context of the entire writing,
especially when an analysis involves a rare usage. The question of the sense of community is one of determining which abstract word object is being used here, just as with lemmas. As stated in section 2.2, once you determine which word (including its sense) is being used, the rest of the analysis follows fairly cleanly.

The question for this example of community is one of word sense, but if you are questioning the meaning resulting from a change in form, then you will need to refrain from thinking about the intended meaning. In (50), for example, place is a singular morphological form (NN1) and, similar to (49) above, appears in a context where plural (NN2) might be appropriate. By the same reasoning as for (49), we wind up with the underspecified NN.

(50)  Also, I need more natural and friendly place to live with my wife . . .

There is a difference from the previous examples, in that, the writer seems to be speaking about one and only one place where he wishes to live with his wife. However, that would commit us to a particular intended form (without changing the word sense), when the syntactic context is ambiguous. Thus, stick with NN here.

**Multi-word conflicts: Coordination**  With coordination, there are surrounding conjuncts to look to for evidence. Consider (51), where the adjective commercial (cf. commercial) conjoins with the noun accountancy.

(51)  Graduating from commercial and accountancy Faculty in Thailand . . .

Consider accountancy here: on the one hand, as a noun, it is a perfectly acceptable modifier of the noun Faculty, yet on the other hand, its adjectival conjunct commercial predicts that accountancy be adjectival—which is still compatible with Faculty. Thus, we might consider an adjective distributional tag for accountancy. However, by this exact same logic, we could deem commercial to have a noun tag.

But in this particular sentence, the conjunction is odd, not ungrammatical. Indeed, in many cases, adjectives and nouns are able to coordinate with no problems, as illustrated in (52). Thus, we simply mark the distributional POS for accountancy in the same fashion as the morphological POS (NN1), and likewise for commercial (JJ). (See also the section on Coordination of unlikes in section 3.2.)

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

In some cases, there may be a more major violation of conjunct parallelism. In those cases, since we treat the first conjunct as the ancestor of the other one, the descendant conjunct can be distributionally marked as the same category as the head. For ungrammatical cases parallel to this one, the second word (cf. accountancy) would be distributionally marked as the POS tag of the first word, because the first word is the head of the coordinate structure. (If other contextual factors are at play, a different conjunct may be selected as the more predictive one on the basis of that context.)
2.4 Lexical violations

Lexical violations are to be annotated when the sentence is ungrammatical (not semantically anomalous), and the only way to mark this is by marking that the word choice is wrong. This generally means that the rest of the annotation is fine, but there is still an issue with the sentence sounding syntactically grammatical. In other words, when there is something ungrammatical, but no other level of annotation captures it, a lexical violation might (if word choice is the issue).

As we will see in chapter 3, dependency relations are generally restricted by having two categories which can have such a relation; for example, a verb and a noun can have a subject (SUBJ) relation, whereas a verb and an adverb cannot. Sometimes, the categories are perfectly valid for a particular relation, but the specific lexical item is a problem. For example, in (53), we see a case where the preposition with is an odd choice to modify the noun interest (*cf.* in). Categorically, there is no issue, as a preposition is a perfectly valid modifier (MOD) of a noun, but with is not the right lexical choice.

(53) The baby had no more interest with the tin toy.

For lexical violations such as these, simply check the LEX box, indicating that, while everything else in the tree may be okay, the particular lexical item is odd.

As another example, consider (54), where the adverb kindly is an odd choice to modify strong, even though (certain kinds of) adverbs can in general modify adjectives. The LEX annotation seems appropriate here, too.

(54) That made me really kindly strong.

There are times when there appears to be a lexical violation, but it is a whole string that sounds odd, and it is hard to pinpoint which exact word is the source of the problem. In (55), for instance, the phrase spend a good time sounds ungrammatical, but it is not clear which word is causing the oddness. In such cases, we mark the head of the phrase, in this case spend, as a lexical violation. Perhaps not coincidentally, in this particular case a different verb (e.g., have) would have resulted in a well-formed sentence.

(55) A visitor will spend a good time in Bloomington.

Remember that the litmus test is whether the string is ungrammatical or not, not whether it is a semantically correct word. In (56), for instance, the word prison is clearly semantically odd, but it is not a syntactically ungrammatical string.

(56) After to start, I want to tell that this excersice is very important in the life, no only as a prison.

Furthermore, recall that it is a specific word choice which is the problem; presence or absence of a word is generally a different issue (at least for arguments; see below). The
anomalous presence or absence of a determiner (an argument) is not a lexical violation, in our sense, because that has to do with whether a category is being used correctly and thus will be handled via subcategorization (see section 3.4). However, if the wrong lexical determiner is used (e.g., the instead of a), then there is a lexical violation.

These principles can be seen in (57), with the two prepositions about and on, which are treated differently. In the former case, we have a preposition appearing with agree, which is normally allowed—but the phrase about me is anomalous here. It is debatable whether about or me is the true source of the anomaly, so we mark it on the head, about. The preposition on occurs as an argument and has a different analysis: influenced selects for an OBJ (indicated within the angled brackets—see section 3.4), but here finds an IOBJ. It is not (just) the specific preposition which is incorrect, but the fact that there is a preposition at all.

(57) ... I agree about me that my country’s help and cooperation influenced directly on my knowledge and my personality

Figure 2.2: Lexical violation vs. Subcategorization violation

The only time presence or absence makes a difference is for adjuncts which are extraneous, and their presence makes the sentence ungrammatical. In such cases, we can mark lexical violations. For more on this, see Extraneous adjuncts in section 5.2.2.

2.4.1 Lexical violations vs. Lemma changes

Lexical violations seem to be similar to normalizing a word to its lemma form (section 2.2), but there are some crucial differences. The lemma is the first issue to consider in annotating, while lexical violations are essentially the last one. To annotate a lemma, you are trying to determine what the word actually is; thus, you change a word only when it is a phonetic or orthographic problem. For example, deal might be changed to deer in a particular context, as in (58).

(58) I can see rabbits, squarrow[↦squirrel], deal[↦deer], or beautiful red birds.

For lexical violations (marked by LEX), you have already determined what the lemma is, and you are then determining whether the lemma is the correct element to
use in this (syntactic) context. In this example, the words *deer* and *squirrels* are perfectly acceptable, so LEX is not marked. Often, you will find that a corrected lemma means you do not mark a lexical violation (though, it is possible).

### 2.4.2 Lexical violations vs. POS mismatches

Likewise, when compared to POS mismatches, lexical violations are a last resort. The two relevant definitions are:

- If a word features different morphological and distributional evidence, then annotate a POS mismatch. This generally means that there is a problem with the morphology of the word.
- If a word is the wrong word in some other way, annotate a lexical violation.

If there is a POS mismatch, then do not annotate a lexical violation. Lexical violations are defined to be used only when the category of the word is correct—and when the morphological and distributional POS tags conflict, then the category is no longer correct.

In (59), for instance, the verb *study* is in the wrong tense, present when it occurs in a past tense context. This is a simple morphological issue: the morphological POS tag is VV0, while the distributional one is VVD (past tense). There is no need for a lexical violation here.

(59) So at first, I *study* economics.

(60) Everybody *have* plan in their life and *try* to get *it*.

Likewise, in (60) (=(13)), *have* is morphologically VV0, but distributionally VVZ (third person present tense), and the same goes for *try*. In this case, the subject *Everybody* drives the agreement. With a similar use of underspecification, *plan* is morphologically NN1, but distributionally NN, reflecting its unclear plurality.

The uncertain singular/plural word raises an issue for *it*, as it likely refers to *plan*. Since there is literal agreement between *it* and *plan*, we do nothing special, tagging both the morphology and distributional as PPH1. If, however, *it* appeared in a context where it oddly may refer to either a single individual or a group (cf. *them*), then one could underspecify the POS tag of PPH1 (*it*) to PPH.

Other cases involve something other than inflectional morphology, often derivational morphology. Still, if a POS mismatch is all that applies, it is all that we annotate. In (61), the word *swim* is morphologically a verb (VVO), but it occurs in a context where a noun (NN) or a gerund (VVG) is expected. In this case, we annotate it as a POS mismatch only. Since the items it coordinates with are gerunds, we mark VVG.

(61) Besides, I like *swim* and running, readings.
The same principle applies for the example back in (30), for full in the phrase *in full of love*. In this case, there is a JJ (morphological) vs. NN (distributional) mismatch, and no lexical violation is annotated.
Chapter 3

Dependencies

This chapter covers the dependencies which form the core of the annotation (section 3) and also subcategorization frames (section 3.4), which represent the dependencies that are selected for by words, but not necessarily realized. These are the primary layers of annotation to focus on.

3.1 Morphosyntactic Dependencies

Our most important annotation is for morphosyntactic dependency relations for each token/word. These relations are based on: 1) the lemma annotation and 2) the morphological POS tag. (The distributional POS tag is, in some ways, linked with the subcategorization annotation, as described in section 3.4.) In other words, it is important to recognize that these dependencies are rooted in morphology.

Arrows are drawn from the head to the dependent, and the dependency relations are from the CHILDES inventory (Sagae et al., 2004, 2007). We allow for words to have more than one head (so-called secondary dependencies, see section 3.2.6), but this is not something you have to worry too much about. In general, though, make sure your arrows go in the proper direction: from the head and to the dependent. For example, a SUBJ arrow goes from a verb head to its noun subject.

3.1.1 Inventory

We list here the dependency relations which we use, taken from the CHILDES manual. We try to make these guidelines self-contained, but if you need more information, you can consult the paper on the annotation (Sagae et al., 2010) (http://people.ict.usc.edu/~sagae/docs/sagae-etal-jcl2010.pdf) and/or chapter 12 of the CLAN manual itself (MacWhinney, 2000) (http://childes.psy.cmu.edu/manuals/CLAN.pdf), to make determinations about how the labels are used. Importantly, though,
CHAPTER 3. DEPENDENCIES

note that we have a few important, systematic deviations from CHILDES usage, as discussed in section 3.2.

We list the relations here, as listed in the manual, where our notation is used as follows:

- **in bold**: relations we added to CHILDES

- **strikethrough**: relations from CHILDES we decided not to use

- **plain font**: relations which are selected for, i.e., should appear on subcategorization lists (see section 3.4). In other words, these are argument labels (see section 3.3.4).

- **italicized**: relations which are not selected for, i.e., should not appear on subcategorization lists (see section 3.4). In other words, these are adjunct labels (see section 3.3.4).

**Predicate-head relations**

All of the predicate-head relations are ones which can be subcategorized for.

1. SUBJ (subject)
   
   SUBJ is a typical (non-clausal) subject.

2. CSUBJ (finite clausal subject)
   
   CSUBJ means that the subject is itself a finite clause (not that the subject occurs in a finite clause), as in this constructed example, where *likes* (heading the finite clause) is the CSUBJ of *is*.

   (62)  [That he **likes**<sub>CSUBJ</sub> tea] is well-known.

3. XSUBJ (non-finite clausal subject)
   
   For XSUBJ, the subject is itself a non-finite clause (not that the subject occurs in a non-finite clause), as in this constructed example, where *To* (heading the non-finite clause) is the XSUBJ of *is*.

   (63)  [**To**<sub>XSUBJ</sub> like tea] is a refined quality.

4. OBJ (object)
   
   Note that OBJ is for the first argument of a verb, regardless of how it alternates. In (64), for example, we can see that the first argument noun after *brings* is the OBJ.
3.1. MORPHOSYNTACTIC DEPENDENCIES

(64)  a. ... to obtain such knowledge brings $me_{OBJ}$ another aspects$_{OBJ2}$ to do my jobs.

   b. ... to obtain such knowledge brings another aspects$_{OBJ}$ to $me_{IOBJ}$ to do my jobs.

5. OBJ2 (second object)

   Distinguishing OBJ2 from IOBJ can be seen in (64): OBJ2 is the second object for a ditransitive verb, whereas IOBJ introduces prepositional phrases serving as arguments to verbs.

6. IOBJ (indirect object)

   This is for prepositional phrases serving as arguments to verbs (see section 3.3.4).

7. LOC (locative)

   Note that LOC is only used for arguments, e.g., for the last argument of put (e.g., I put the eggs away). If there is a location-specifying phrase which is an adjunct (e.g., I ran to the store), treat it as JCT (see below).

8. COMP (finite clausal complement)

   COMP indicates a complement, or argument, of a verb which is a finite clause. In (65), for example, the clause headed by will is a COMP of believe.

   (65) I believe that these 2 goals will make my life fantastic ...

9. XCOMP (non-finite clausal complement)

   XCOMP indicates a complement, or argument, of a verb which is a non-finite clause (to-phrase, -ing phrase). In (66), for example, the clause headed by to is an XCOMP of like.

   (66) I would like to further promote energy-efficient electric appliances ...

10. PRED (predicate)

   PRED, CPRED, and XPRED occur mainly after copula (be-form) verbs and similar verbs (e.g., become), though other verbs may license them, too (e.g., the second argument in make me happy). Otherwise, they parallel OBJ (noun/adjective), COMP (finite clause), and XCOMP (non-finite clause).

11. CPRED (finite clausal predicate)

12. XPRED (non-finite clausal predicate)
13. POBJ (prepositional object)

P OBJ is simply the prepositional object: for example, in full of love, the noun love is a POBJ dependent of of.

14. VC (verbal complement)

VC generally goes from a modal or auxiliary verb (or to) to the following verb form, e.g., developed is a VC dependent of have in have developed (see section 3.2.5).

Argument-head relations

One thing to clarify is the dictum in CHILDES that for these relations “the arguments (rather than the predicates) serve as the heads” (MacWhinney, 2000, p. 185), as it depends upon what is deemed the argument and whether that includes semantic information. We focus on syntactic heads and dependents, and we reserve the term argument to refer more specifically to dependents which are generally obligatory (see section 3.3.4).

15. JCT (adjunct)

We treat JCT as an adjunct for any non-noun modification: verb-adverb (e.g., quickly as a JCT dependent of ran), adjective-adverb (e.g., very as a JCT dependent of poor), adverb-adverb (e.g., really as a JCT dependent of quickly), and so forth.

16. CJCT (finite clausal adjunct)

CJCT is for a non-finite clause modifying a verb, generally modifying a whole sentence, as in (67) where came is a CJCT of did.

(67) When I came to Bloomington for living, I did not understand English ...

17. XJCT (non-finite clausal adjunct)

XJCT is for a non-finite clause modifying a verb, generally modifying a whole sentence, as in (68) where To is an XJCT of have.

(68) To fulfill the dreams, I have to take three actions.

18. NJCT (nominal adjunct)

NJCT is used for a post-modifying preposition of a noun. Technically, this label could sometimes selected for, sometimes not, as nouns can have complement prepositions like of. However, we treat NJCT as not being subcategorized for, in order to simplify an otherwise difficult decision. In this moment of my life, for example, of is an NJCT of moment, regardless of whether it is an argument or an adjunct.
3.1. MORPHOSYNTACTIC DEPENDENCIES

19. **MOD** (nominal modifier)

MOD is used for nominal (pre)modifiers, generally adjectives (e.g., *different* is a MOD of *goals* in *different goals*). Note that we treat this as strictly nominal modification, as compared to JCT, which applies to all non-noun cases.

20. **CMOD** (finite clausal modifier)

CMOD is the label for relative clauses modifying nouns. The head verb of the relative clause is a CMOD dependent of the noun. In *the boy who lives in the bubble*, for example, *lives* is the CMOD dependent of *boy*. See section 4.4.4.

21. **XMOD** (non-finite clausal modifier)

XMOD is the same as CMOD, but for non-finite modifiers: the head is the noun, and the dependent is the *to* of the clause modifying the noun. This is illustrated in figure 3.1.

![Figure 3.1: An example of an XMOD structure](image)

22. **DET** (determiner)

DET is the determiner relation, e.g., *a* is a DET dependent of *toy* in *a toy*.

23. **QUANT** (quantifier)

QUANT is similar to DET, but applies for quantifiers (numbers, *every*, *no*, etc.). The section on Determiners and Quantifiers in section 3.3.2 distinguishes these labels.

24. **PQ** (post-quantifier)

PQ is for a noun and a following quantifier, e.g., *we both PQ*.

25. **AUX** (auxiliary) (see section 3.2.5)

25. **NEG** (verbal negation)

Note that this is verbal negation, such as that between *n’t* and *run* in *it does n’t run*. For nominal negation, e.g., *no one*, we use QUANT. For adjectival and adverbial negation, we also refer to usual categories, such as JCT.

26. **INF** (infinitival particle) (see section 3.2.5)
CHAPTER 3. DEPENDENCIES

26. SRL (serial verb)  
   SRL is for verbs which come in sequence, essentially involving only come or go, i.e., come hang_{SRL} with us. The second verb is an SRL dependent of the come or go verb.

27. CPZR (complementizer)  
   The relation between a head verb and a complementizer like that or which (see section 4.5 among others).

28. POSS (possessive) ... This marks possessive structures, with the possessive marker (‘s, ’) being the head, as illustrated in figure 3.2.

\[
\text{DET} \quad \text{POSS} \quad \text{DET} \\
\text{my parents} \quad \text{opinion}
\]

Figure 3.2: A possessive structure

29. PRT (particle) ... This marks the relation between a verb and its particle, e.g., up as a PRT of call in call up. More on distinguishing PRT from JCT (for prepositions/adverbs) is in section 3.2.3.

30. TAG (tag questions)  
   This is used for verb-verb relations when a tag question is asked; the tag verb is the dependent, as shown in this constructed example:

\[(69) \quad \text{He went home, did}_{TAG} \text{ n’t he?}\]

Root linkage

31. COM (communicator)  
   COM marks filler words, or communicators, like hey, should they appear in the text. We do not expect this to be used much.

32. VOC (vocative)  
   VOC refers specifically to words or phrases referring to someone being addressed, for example, John in John, I’m only dancing. As with COM, we do not expect this to be used much in written form.

33. TOP (topicalization)  
   Since our data is written, we do not expect to use these three relations very often. However, we leave them as options, in case they should ever become necessary.
Also, we use TOP only when an argument is repeated (e.g., Tapioca, there is no tapioca); if the argument is not repeated, simply give it the label it would otherwise receive (e.g., OBJ for Tapioca, I like.)

We find a learner example in (70), where goals is a TOP dependent of catch. The pronoun them already functions as an OBJ, so TOP is what is needed, as this is a completely grammatical sentence. (See the Missing conjunction part in section 3.3.3 for cases where two objects appear to exist, but ungrammatically.)

(70) So, My goals I catch them.

34. INCROOT (incorrect root)

We use INCROOT not just for constructions such as predicates with missing copulas, but also for things such as non-finite verbs heading sentences (see section 4.2). For example, in The toy afraid, the predicate afraid is the INCROOT of the sentence.

35. TRANS (transition)

TRANS is a more specific case of JCT, attaching to the head of the entire sentence. It is used for words that are discourse connectives, often as transitions between sentences (e.g., First, Also, However), and it is also used for phrases. For example, if a sentence begins with After all, connecting the sentence to the discourse, we would treat After as a TRANS dependent of the sentence’s main verb since After is the head of that phrase. Note, too, that TRANS can occur within the middle of sentence, for example, if however or for instance appear mid-sentence, they are still TRANS relations.

We also mark items which are traditionally coordinating conjunctions, such as and, as TRANS, when they occur at the beginning of a sentence. Examples distinguishing TRANS (discourse-connecting) from JCT (non-discourse-connecting) are given in (71).

(71) a. TRANS: For this reason I need to obtain financial aid.
   b. TRANS: And I think my English skill is not good.
   c. TRANS: For other side, the past semester I was teaching ...
   d. JCT: In the future, I will want to be ...

Cosmetic relations

36. PUNCT (punctuation)

See section 3.3.5 for how to attach punctuation.
CHAPTER 3. DEPENDENCIES

37. **RDP** (right dislocation punctuation)

37. **VOCP** (vocative punctuation)

37. **ROOT** (root)

ROOT is used for the overall head of the sentence, typically the verb. You will generally only have one ROOT relation, and if you use INCROOT, it generally precludes using ROOT (and vice versa).

Series relations

38. **NAME** (proper name element)

NAME relations are for proper names and are generally headed by the rightmost word, e.g., *John* is a NAME dependent of *Smith* in *John Smith*. See section 3.3.2 for more details.

39. **DATE** (date element)

As with NAME, DATE relations are generally headed on the right. For *March 1, 2001*, for instance, *March* and *1* are DATE dependents of *2001*.

40. **ENUM** (enumeration)

The enumeration label is for lists of things (e.g., *apples, bananas, pears*). We treat enumerations as right-branching, i.e., the leftmost item is the head, and the chain works rightward (see section 3.3.7). This can be difficult to distinguish from broken coordination structures: see the section on *Missing conjunction* in section 3.3.3 for more details.

41. **CONJ** (preceding conjunction)

CONJ is in some sense replaced by CCC; see below.

41. **COORD** (coordination)

COORD is the label used between *and* and the word conjoining with it. For example, in *cats and dogs*, the relation between *and* and *dogs* is COORD, with *and* as the head.

Note that COORD only applies to coordinating conjunctions. If a subordinating conjunction (e.g., *because*) is involved, you will likely be using CPZR. Furthermore, it only applies between the conjunction (e.g., *and*) and the final item in a coordination structure; MCOORD is used if there are more than two elements in a coordination structure.

42. **MCOORD** (modificatory coordination)

MCOORD is used to relate two items in a coordinated list, when there are three or more total items. For example, in *cats, dogs, and antelopes*, the relation
between *cats* and *dogs* is MCOORD, with *cats* being the head. See sections 3.2.8 and 3.3.3. (Specifically, see the part on Missing conjunction for distinguishing MCOORD from ENUM.)

43. **CCC (coordinating conjunction)**

CCC is used for the relation between an item in a coordination and the conjunct. For example, in *cats and dogs*, the relation between *cats* and *and* is CCC, with *cats* as the head. The dependent of CCC will always be a conjunction. See sections 3.2.8 and 3.3.3 for overall coordination structures.

44. **APPOS (appositive)**

APPOS accounts for appositions, such as in (72), where *raings (and string)* is an APPOS dependent of *things*, as well as intensifying reflexives (e.g., *the clown itself*).

(72) thir are two things besaid the Toy, raings and string of beads.

Other relations

45. **ELL (ellipsis)**

This label marks the relation between two elements that normally would not have a relation, but do so due to missing material. See section 4.10 for more.

46. **_(unspecified)_**

This (last-resort) label is used when the evidence cannot point to a given label. See section 1.2.5 for more information on when to underspecify.

3.1.2 The syntactic in morphosyntactic dependencies

In order to annotate, it helps to know where the annotation is coming from. When we speak of morphosyntactic dependencies, we are talking about syntactic functions which are determined largely on the basis of morphology. This is not strictly the case, as we make clear next, but it should be your starting point.

Incorporating distribution

While we base dependencies on morphology, there is often more than one morphological possibility (i.e., words are morphologically ambiguous), so the general policy is to annotate the morphological dependency which is the most contextually-relevant. In

---

1This is essentially the same as APP, added to the CHILDES inventory after we had developed most of our label inventory.
other words: the morphosyntactic layer of dependencies (and also POS tags) incorporates some degree of distributional information. We can state the heuristic as such:

- **Policy:** Starting with the morphological form of a word, annotate the dependency tree which best fits the context and which leads to the most well-formed tree (cf. “give the learner the benefit of the doubt”).

In annotating the closest fit to the context, consider (73), where the morphological form is ambiguous, between base form verb and non-third person singular present tense (in this case, both have the POS tag VV0). Neither one is an exact match to the context, but as the head of the entire sentence, the tensed form takes precedence, as sentences need tense. Thus, we annotate the tree based on the morphological form of a non-third person singular present tense verb. In this case, the morphosyntactic dependency relation is ROOT because ROOT applies to third singular or non-third singular verbs. (The non-nativeness is then handled by marking the distributional POS as VVZ, to agree with year.)

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

The priority is still on morphology, though, and on putting together a well-formed tree, if the morphological information warrants it. Consider **dull** (cf. **doll**) in (74), where the preposition **with** occurs in what could be considered a post-positional slot. Morphologically, the **dull** is compatible with being the object of **playing**, while at the same time compatible with being the prepositional object of **with**. Crucially, one of these analyses (prepositional object) leads to a complete well-formed tree, while the other one (verbal object) does not. We advocate the prepositional object analysis here, using the following heuristic: when there is more than one possible morphosyntactic analysis, select the one which is the best fit, i.e., leads to the most well-formed tree.

(74) Baby playing **the dull** with.

Note that we are de-emphasizing the actual word order here, since morphosyntax is based largely on morphology.

**Selecting a label**

In terms of annotating on the basis of the morphology of the dependent, consider (75), ignoring the space in *a bout* (=about). The dependency relation we use should be one which is appropriate for a verb, as this is the morphological form of **chooses** (VVZ=third person singular present tense verb).

(75) I had a problem a bout **chooses** my car . . .
What label to use is complicated, however. The problem here is that *about* can really only have one type of dependent, namely prepositional object (POBJ), but this is not compatible with a verb like *chooses*.

In general, the guideline is:

- **Policy:** Select a label which is appropriate for both the head and the dependent (specifically: appropriate for the POS of the head and the POS of the dependent).

We thus have a few possible solutions for non-native morphological forms, outlined as follows:

1. If there is a label which is “generic” and covers the native and non-native cases, use that label. This happens for verbs which do not agree with the subject but are still ROOT, as with (73).

2. For certain widely-used cases, we may have a specific label. Currently, there is one: for non-finite verbs heading a sentence, we use the label INCROOT. In (76), for instance, the non-finite head of the sentence, *living*, is a INCROOT dependent of the virtual root (ROOT).

   (76) One day in Indiana, *living* for the cousin and his married, born in Chile, and studying in Chile Ingenerie

3. In many cases, we will have to leave the label unspecified, as there is no label that fits both the head and the dependent. This is what we do for (75), as shown in figure 3.3.

   ![Figure 3.3](image)
   
   **Figure 3.3:** Non-specification of a label when the head and dependent categories have no label appropriate for both

**Category-based definitions** One note on the idea of using labels which are appropriate for the head and the dependent is that the labels are appropriate for the category of a word (and not the specific lexical item). In (77), for instance, *the street* is not an appropriate object for the verb *fainted* because this is not a transitive verb. However, verbs can obviously take objects, so we annotate the relation as OBJ. In this case, the oddness is captured by the fact that the subcategorization of *fainted* does not select for an object.
46  
CHAPTER 3. DEPENDENCIES

(77) She fainted the street.
(78) The baby had no more interest *with the* tin toy.

Likewise, in (78), *with* is lexically an odd choice after *interest*. However, prepositions naturally occur here, so an NJCT label between *interest* and *with* is appropriate. In this case, since *with* is not subcategorized for, the oddness is captured through a special lexical violation marker (see section 2.4).

3.2 Where we differ from CHILDES

If you look at the CLAN/CHILDES documentation, it is important to note that we differ from the CHILDES scheme in a few respects, regarding possessives (3.2.1), the use of NJCT (3.2.2), particles (3.2.3), dependents of adjectives (3.2.4), verbal chains (3.2.5), secondary dependencies (3.2.6), enumeration (3.2.7), coordination (3.2.8), and ellipsis (3.2.9).

3.2.1 Possessives

As mentioned in section 3.1.1, we use POSS for possessive structures, whereas they are handled using a MOD relation in CHILDES. The possessive marker is the head of the possessive phrase. For example, in *my parents’ opinion*, the marker ’ is a POSS head of *parents*, and a DET dependent of *opinion*, as shown back in figure 3.2.

3.2.2 NJCT

It is not entirely clear how to handle noun-modifying prepositions in CHILDES which are complements. We define NJCT to be used for both argument and adjunct prepositional modifiers of nouns, as these can be hard to distinguish (see section 3.3.2).

3.2.3 Particles (PRT)

We use PRT (particle) for the case when a preposition-like word follows a verb, but binds more tightly to the verb than to the noun. In *look up the number*, for example, *up* is a PRT dependent of *look*. You can see that the meaning of *look up*, in this instance, is not derived in a straightforward way from the meaning of its parts.

We follow the diagnostics in Santorini (1990) for distinguishing prepositions from particles (see IN vs. RP, p. 10–11). Most prominently, if you substitute a pronoun for the noun (phrase), the ordering is critical: in this PRT case, you would have *look it up* and not *look up it*, whereas for the JCT relation in *look down the road*, you get *look down it*, not *look it down*. Particles in general are more flexible in where they appear.
3.2. WHERE WE DIFFER FROM CHILDES

An example of a verb-particle structure is in figure 3.4. Note that the noun is an object of the verb, not of the particle.

![Diagram of a verb-particle structure](image)

Figure 3.4: An example of a PRT structure (constructed example)

3.2.4 Dependents of adjectives

It is not entirely clear what the relations should be for adjectives in CHILDES and whether there is a distinction between arguments and adjuncts (which we generally distinguish, see section 3.3.4). It can be difficult to distinguish arguments from adjuncts, so we elect to use the labels JCT, CJCT, and XJCT for all dependents of adjectives (cf. the treatment of NJCT for nouns, section 3.2.2). An example of XJCT is in figure 3.5.

![Diagram of an non-finite clause dependent on an adjective](image)

Figure 3.5: An example of an non-finite clause dependent on an adjective

Even though it is XJCT, an adjunct relation, note that we “raise” the SUBJ, i.e., establish a SUBJ between the embedded verb gain and I. See section 3.3.1 for more on raising.

3.2.5 Heads for verbal chains

Because we are annotating morphosyntactic dependencies, and not semantically-based ones, we deviate from the CHILDES scheme in terms of what the head of a sequence of auxiliaries and verbs is. We treat the first auxiliary (or infinitive marker) as the head of the subsequent verb, assigning a VC (verbal complement) label, as shown for the heads would and to in figure 3.6. (VC is thus a replacement for both the AUX and INF labels used in CHILDES.)

Note, too, how subjects are handled here: auxiliaries are treated essentially as raising verbs (see section 3.3.1). Thus, both the local verb (i.e., would) and the contentful verbs
(like, promote) have I as a subject. This follows the general guideline of annotating an argument relation if:

1. The head is a raising (or control) predicate/verb, and the argument relation is local (i.e., introduces no crossing branches).

2. The head is a main/content verb, and the subcategorization requirement is such that it requires the argument, even if it has been raised.

We can see this played out in a chain of verbs, as in figure 3.7, where would and promoted have I as a subject. Note that, for linguistic thoroughness, it would be correct for the other three verbs in figure 3.7 to take I as a subject (likewise for to in figure 3.6); we simply do not require such annotation, for simplicity’s sake.

3.2.6 Secondary Dependencies

Dependency annotation often assumes that each dependent has only one head, and this is the case in the CHILDES scheme. As we have seen above, we allow for a word to have more than one head. A word having two heads should only be used in restricted cases, namely those where subcategorization is at stake. Up above in figure 3.15, for example, gain selects for a subject (SUBJ) and an object (OBJ); in order to make sure that a learner is given appropriate credit for properly realizing the subject, such a relation is annotated.
3.2. WHERE WE DIFFER FROM CHILDES

3.2.7 Enumeration

CHILDES treats enumerated lists as left-branching; we treat them as right-branching, as illustrated in figure 3.8. This will put enumerated lists more in line with coordination (see next section). Enumerated lists can be difficult to tell apart from coordinations missing a conjunction; we discuss this below in the Missing conjunction part of section 3.3.3.

![Figure 3.8: Treatment of an enumerated list (constructed example)]

ENUM has been used in CHILDES to handle appositives, but as mentioned in the label inventory in section 3.1, we use the more specific APPOS for this.

3.2.8 Coordination

We use a right-branching analysis for coordination, introducing a new label in the process, CCC, for the relation between the conjunction (e.g., and) and its conjunct. In figure 3.9, for example, the coordination is between toy and baby, so toy serves as the prepositional object (POBJ); and is the CCC of toy; and finally baby is the final coordination (COORD) element. As discussed in section 3.4, and subcategorizes for COORD, whereas nothing selects for CCC.

![Figure 3.9: Treatment of basic coordination]

If items are coordinated which share an argument, each one of them takes that argument individually. In figure 3.10, for example, walk and run both have He as their subject (SUBJ) (as begins also does, this being a case of raising).

As a side note, note that the relation between to and at is JCT. This is because infinitive phrases can take prepositional modifiers; just because there is no prepositional object does not change the fact that an infinitive marker (TO) and an preposition (II) can
be in a JCT relation. (The non-nativeness is captured by *at* having an underspecified distributional POS tag and by it having a subcategorization list containing POBJ, which is not realized; see section 3.4.)

### 3.2.9 Ellipsis

The CHILDES scheme defines a number of labels which are “used for marking ellipses, omissions, or deletions”, such as DET-POBJ for the relation between *up* and *the* in the cut-off phrase *climb up the*. We handle such cases differently, namely in two ways:

- In cases where something is missing, this is handled either via: a) the mismatch between subcategorization and realized dependencies; b) lexical violations; or c) unspecified dependencies (⊥), as would be the case for the relation between *up* and *the* above. See section 5.1 for more on missing elements.

- In the case where there is a grammatical sentence with omitted material (ellipsis, verb gapping, etc.), we generally use the special label ELL. In (79), for example, the head of the subordinate clause *beginning* is the type of clause which modifies a verb (XJCT), but here it seems to modify a *principal*. In some sense, *a principal* is an ellided version of *I want to be a principal*. Thus, we mark *beginning* as an ELL of *principal*. See section 4.10 for more on this type of situation.

(79)  I want to be a teacher and, **beginning** in 2020, a principal. (constructed example)
3.3 Overview of annotation

3.3.1 Verbs & Verbal relations

Dependency relations

Some of these points were covered above in section 3.2.5 on verbal chains. In general, verbs are annotated as follows:

1. Main verbs are the head of the sentence and attach to the virtual root (ROOT), e.g., loves in *He loves chocolate* is a ROOT dependent of the ROOT token.

2. If there are auxiliary verbs or an infinitive marker, they head the main verb. For example, in *They will go home soon*, *will* heads *go* in a VC relation. *They* is the subject of *will* and of *go*.

3. In the case of complex sentences, the main verb heads the subordinate verb, as defined in the CHILDES scheme. For example, in *We like her because she is funny*, the subordinate verb *is* is a dependent of *like*.

Raising & Control

We treat raising predicates (e.g., *seems*) and control predicates (e.g., *expects*) identically. As shown in figure 3.11, the subject of the “higher” clause (*seems*/*expects*) is the same as the subject of the “lower” clause (*win*).

![Figure 3.11: Raising and control treated identically (constructed example)](image)

Raising was first discussed surrounding figure 3.6. Remember that raising can involve not only sharing of subjects, but also so-called subject-to-object raising. In the constructed (80), for instance, *someone* is both the subject of *to help* and the object of *found*, as shown in figure 3.12.

(80) I found *someone* to help me.

(81) I found *what* to further improve my ability …
This is useful to keep in mind for learner examples like (81), where what (a potentially noun-like token) appears in a possible subject-to-object raising position. Thus, we can annotate it as in figure 3.13. Since what is a strange choice for a declarative subject, leading the sentence to sound ungrammatical, it is also marked as a lexical violation (see section 2.4), indicated here with an asterisk.

Another example of this is in (82), where we make (my) life the “raised” subject of the infinitival phrase, as shown in figure 3.14. Semantically, this is questionable, but it is a perfectly valid syntactic structure, following the notion that a verb such as like licenses subject-to-object raising.

(82) I would like my life to be successful in career and to be rich.

VC vs. (X)COMP

The VC label may at first seem highly similar to COMP (finite clausal complement) or XCOMP (non-finite clausal complement), and indeed they are similar. VC is used for a range of following verbal forms (base, past participle, passive participle, progressive) and the dependent does not head a clause (cf. sentence), but rather a verb phrase. VC is used specifically when:

- the following verb does not realize its subject locally; and
- the following verb is not to
3.3. OVERVIEW OF ANNOTATION

Another way to look at it is that VC is used when AUX or INF would have been used in the original CHILDES annotation scheme; i.e., the head of the VC relation should always be an auxiliary (forms of be, may, could, etc.) or infinitive marker (to). COMP and XCOMP are used as defined in the CHILDES documentation.

Copulas

Copulas are handled as defined in CHILDES, but we mention them here, so as to make properties such as raising clearer. In figure 3.15, the copula 'm takes a subject and a predicate. The subject is generally the item which precedes the copula, and the predicate (PRED/CPRED/XPRED) the item which follows it.

Note that the predicate itself does not take a subject (though, in some syntactic theories it would); we only have verbs take subjects. Copulas, though, always take a subject; thus, if we have an auxiliary or some other form of raising (see above), the copula will still take a subject, as in figure 3.16 where being is the copula. If annotating a sentence like John seems to be happy, the verbs seems and be would have John as a subject (SUBJ).
CHAPTER 3. DEPENDENCIES

I was being eager to gain knowledge...

Figure 3.16: Treatment of a copula and its raised subject (constructed example)

VC vs. (X)PRED

Another confusion stems from a verbal, or verbal-like, form appearing after the copula. In (83), for example, are these verbs or adjectives? If verbs, then we have a VC relation stemming from *are*; if adjectives, then we have XPRED.

(83) a. We are learning
    b. We are socialized

Two diagnostics are worth mentioning (see others in the sections on JJ or VBG and JJ or VBN in Santorini (1990)):

1. What types of words can modify the verbal-like form? If it is a degree adverb, such as very or quite, then it is an adjective. Compare We are quite socialized vs. *We are quite learning.

2. What other verbs could be used here? If you can use become, feel, or similar verbs, then it is an adjective. Compare We became socialized vs. *We became learning.

In these cases, then, socialized is an adjective and thus a PRED of *are*, whereas learning is a verb and thus a VC of *are*. Since learning is a verb, it also takes We as a SUBJ, (whereas adjectives do not—see more under Gerunds/Progressives below).

Passives

As discussed in section 3.2.5, we treat auxiliaries as heads which take verbal complements. We do the same for passive verbs, even though the semantic interpretation is quite different. For example, the tree for (84) is given in figure 3.17 where changed is a VC complement of was, and both verbs take life as a subject (SUBJ).

(84) My life because of Bloomington was changed into lovely.
As a side note, note in this example that lovely is underspecified here, because there is a preposition and an adjective, which in general is not a POBJ type of element. With such a category mismatch, we leave it underspecified (see section 1.2.5).

Also note that if a by-prepositional phrase appears with a passive, we treat it as JCT, because it is not required and can be reordered with other adjunct phrases (see section 3.3.4).

**Verbal modifiers**

The way we annotate verbs, if there is one or more auxiliaries, it is not immediately clear where to attach adjuncts. In (85), for example, it is clear that the bolded clause headed by to modifies the verbal chain must be helpful, but which specific verb does it attach to: must, be, or helpful?

(85) to understand global environmental issues, intensive discussions with diverse classmates must be helpful.

We specify the following:

1. If the adjunct clause precedes the verb, then attach to the head (first/leftmost) part of the verbal chain, in this case must. This follows the principle of trying to avoid non-projectivity, when possible.

2. If the adjunct follows the verb and is within the clause, attach it to the contentful (rightmost) element, as shown in (86), where in (full of love) is a dependent of live.

3. If the adjunct follows the verb and is outside the clause, attach it to the head (first/leftmost) part of the verbal chain.
CHAPTER 3. DEPENDENCIES

(86) **to live** with my family **in** full of love . . .

For complements, on the other hand, always make them dependent of the word which is selecting for it, even if that introduces non-projectivity.

In point #2 above, we specify that post-verbal attachments to the contentful verb when the adjuncts are within the clause. In point #3, though, we say that if the adjunct is a global, or transitory, type of adjunct, then attach it to the head/leftmost verb. In (87), for instance, the adjunct **too** modifies **are** because it is an adjunct which specifically links clausal, or discourse, elements.

(87) ... because , the people are poor **too**.

As alluded to in (87), these rubrics also work for predicates, not just verbs. In (88), for instance, the prepositions **from** and **to** could attach to either **is** or **free**. Extending the specifications above (specifically, point #2), we have them attach to the adjective predicate **free**, as shown in figure 3.18.

(88) . . . getting knowledge is free , from primary school to university ( higher education )

![Figure 3.18: Attachment of prepositions to a predicate](image)

In terms of word order, the only remaining thing to take care of are modifiers which occur between an auxiliary and another part of the verb chain. In (89), for instance, there is the question of where to attach **rather**, to **would** or to **learn**? We generally favor attaching to the auxiliary, as adverbs in this position can usually take scope over the entire phrase. For example, *John is not happy* means *It is not the case that John is happy*. If the adverb clearly scopes over the content-ful verb, then attach to it; otherwise, attach to the auxiliary or modal (i.e., head/leftmost) verb.

(89) I glad to be in Bloomington and I would **rather** learn various thing more .

One minor note on verbal modifiers in general: certain types of nouns can function as verbal modifiers, so treat them as you would any other JCT. In (90), for instance, *(every) Sunday* is a JCT of *cutting*, even though it is a noun.

(90) ... cutting grass of yard **every Sunday**
Subordinate clauses

Subordinate clauses are handled quite differently than with coordinate clauses. Namely, the subordinating conjunction is a CPZR dependent of the subordinate verb, and the subordinate verb (assuming it is finite) is a CJCT dependent of the main clause verb. We can see this in figure 3.19 for (91), where When is a complementizer (CPZR) of walked, which is itself a finite clause adjunct of the main verb found.

(91) *When I walked on the road, I always found somebody smiled to me.*

![Figure 3.19: Treatment of a subordinate clause](image)

Gerunds/Progressives

Given our treatment of subjects, we need to make clear what happens with -ing forms of verbs (gerunds and progressives). If a word like attending is a true progressive, appearing after a form of be, then we treat it as a verb, where the progressive takes a subject. This is shown in figure 3.20. This treatment follows from the fact that progressives are full verb forms, with the subject always present.

![Figure 3.20: Treatment of a progressive form (constructed example)](image)

Gerunds do not feature a subject, however, and are thus treated differently. In (92), for example, neither Attending nor living takes a subject. They do, however, take any
objects which are necessary, as shown in the OBJ relation in figure \[3.21\] We can also see in this figure that the clause as a whole is an XSUBJ of *are*. Clauses headed by gerunds are non-finite clauses—just like clauses headed by participle forms—and thus participate any relations that non-finite clauses do (XSUBJ, XCOMP, XPRED, XJCT, XMOD).

\[(92)\] **Attending** the MBA program at Kelley and **living** in Bloomington are my current best choices ...

![Figure 3.21: Treatment of a gerund form](image)

Gerunds are verbs, not nouns, and they often occur with some function word preceding it, as in \((93)\). Since these are verbal cases, *by/Before/for* should be treated as complementizers, not as prepositions, as shown in figure \[3.22\] We know they are verbal because, as with verbs, they take direct objects and do not have determiners. (See also section \[4.5\] for distinguishing prepositions from complementizers.)

\[(93)\] a. I believe we will have great time in Bloomington by **attending** a lot of great events and parties ...
   
   b. Before **coming** to Bloomington, my wife was also a hard worker .
   
   c. Therefore everyone have big machine for **cutting** grass .

If it were *by the enjoying of great opportunities*, we would have a prepositional structure. Indeed, there are gerunds which are clearly nominal cases, and in such situations, you should treat them as nouns. In figure \[3.23\] for instance, **thinking** is a noun and is thus a prepositional object.

The distinction between progressives and gerunds also means that progressives subcategorize for a subject, while gerunds do not (see section \[3.4\].)
3.3. OVERVIEW OF ANNOTATION

Negation

For verbs, the negation marker (not, n’t) is generally a NEG dependent of the head verb. In (94), for example, n’t is a NEG dependent of can. Note that it modifies the modal, not the content verb; this is because you can paraphrase this sentence as It is not the case that you can find a place [to] own yourself and not as You are able to not find a place [to] own yourself.

(94) You can n’t find a place own yourself.

In general, the negation modifies the auxiliary or modal. This is true, even if you have a predicate, as in He is not happy, which could be paraphrased as either It is not the case that he is happy or He is unhappy. Only modify the content word (i.e., verb or predicate) if it is clearly the case that negation scopes over it.

3.3.2 Nouns and Noun relations

Pre-dependents: adjectives

Both nouns and adjectives may modify nouns, and in terms of structure, they are generally MOD dependents of the head noun. The relation between exchange/NN1 and stock/NN1 in stock exchange is MOD, just as the relation between aid/NN and financial/JJ in financial aid is also MOD. See the SUSANNE guidelines for distinguishing nouns and adjectives; in general, a word is generally either one or the other, regardless of context.
In “elliptical” contexts where there is no head noun, such as the red, the adjective is still an adjective (morphologically and distributionally), but functioning very much as a noun would, e.g., taking a determiner, able to function as the object of a preposition (e.g., in the red), etc. The tree for in the red is shown in figure 3.24 as an example.

![Tree diagram for in the red](image)

Figure 3.24: Adjective operating like a noun (constructed example)

**Pre-dependents: determiners & quantifiers**

Determiners and quantifiers can be hard to tell apart, but one of the main criteria is whether the determiner-like word is determining the quantity of the noun. Thus, words like some, few, and three are always quantifiers. (Note that no is also a QUANT when modifying a noun, whereas the verbal modifier not is generally NEG.)

Consider words like many and three. In the phrase many children (or three children), many is a QUANT dependent of children. For uniformity’s sake, we also treat it as QUANT in the phrase the many children (or the three children), with the as a DET dependent of children.

![Tree diagram for the many children](image)

Figure 3.25: Treatment of a numeral

There are also cases where the quantifier precedes the determiner, as in all the children, as in figure 3.26. As shown in the figure, we follow the CHILDES scheme in attaching both to the noun. What is important to note is that since nouns select for both quantifiers and determiners, you should list both on the subcategorization lists, for both orderings (see section 3.4).

In cases like all of the children, treat all as the head, with an NJCT preposition after it. Likewise, in all seven of the children, treat seven as the head (cf. adjectives as head, described in Pre-dependents: adjectives above).
3.3. OVERVIEW OF ANNOTATION

For situations like my all children, we have a quantifier appearing inappropriately after the determiner. Our solution is mark this, as before, with a DET and a QUANT, but not to select for QUANT, as in figure 3.27. (There are other possible ways to analyze this, but this is how we choose to do it.)

If a determiner or quantifier has a modifier, this is typically an adverbial of some sort and is labeled as a JCT dependent. (The same is true of adverbs modifying adjectives.) In (95), for instance, over serves as a JCT dependent of the quantifier eleven.

(95) First, I have practiced law for over eleven years.

We will briefly mention the quantifier such, since when it appears before a determiner, it can appear only with a specific one, a (such a night vs. *such the night). In such cases, for example (96), we treat such as a QUANT dependent of the noun, as in figure 3.28. If the determiner were the wrong one (e.g., the), we would mark the as a lexical violation (see section 2.4).

(96) ... look for such a lady

The word such of course has other realizations; in such ladies, for example, such is a QUANT dependent of ladies. In general, though, treat such as a QUANT, even when preceded by a determiner.

There are sometimes adjectives which appear to be quantifier-like, e.g., various, but these are MOD dependents of nouns. If it is not a number and not POS-tagged as
a quantifier, generally label it as a MOD dependent. In general, even if it is a non-
canonical pre-modifier of a noun, it is labeled as MOD, as long as it is not a determiner
or quantifier.

Finally, note that determiners can occur for other parts of speech, in limited cases. For
example, in *a little afraid*, *a* is a DET of *little*, which then serves as a JCT of *afraid*.

**Pre-dependents: possessives**

Possessives are handled by positing a POSS relation between the possessive marker (’
or ’s) and the preceding noun grouping, as shown in figure 3.29. The possessive marker
then serves as the determiner of a noun. Note that the structure would be essentially the
same if there were preceding material, e.g., *knowledge of those consumers ’ behavior*,
simply with more internal structure, in this case a DET relation between *consumers* and
*those*.

**Post-dependents: prepositions**

In the CHILDES manual, it is not entirely clear whether there is a distinction for prepo-
sitional phrases which appear post-nominally, between ones which are subcategorized
for (complements) and ones which are not (adjuncts). Although we could make such a
distinction, we will use NJCT for both types, and these will not be subcategorized for
(see section 3.4).
Post-dependents: clauses

Nouns can be modified by relative clauses (CMOD) and by non-finite clauses (XMOD). One type of XMOD, the infinitive clause, is slightly easier in our annotation, so we start with that. In figure 3.30, we see the clause to be promoted ... modifying the noun chances, and correspondingly, there is an XMOD relation between chances and to, the head of the clause.

![Figure 3.30: Example of a post-modifying infinitive clause](image)

Finite clause dependents of nouns (relative clauses) are usually more challenging, in that we link up a missing argument from the relative clause to the noun. More discussion is in section 4.4.4, but you can see a brief example in figure 3.31, where the head verb thinks is a CMOD of the noun that the clause modifies, man.

![Figure 3.31: Example of a post-modifying relative clause](image)

This is true also of non-finite clauses, where an argument can be clearly (and syntactically) linked to the embedded verb. In figure 3.32, for instance, the reduced relative verb, called, has a subject in the noun that it modifies, Organization. Just as with full relative clauses, then, called is an XMOD of Organization, while at the same time Organization is a SUBJ of called.

Compounds: general nominal compounds

For noun-noun compounds, we generally take the rightmost element as the head, and the other elements are MOD dependents of this head noun. (If it is a proper noun, then NAME is more appropriate than MOD; see the next section.) For example, in the noun-noun compound fossil fuel, fuel is the head, and fossil is a MOD dependent of fuel. The structure for a noun-noun compound will thus look virtually identical to the case where
CHAPTER 3. DEPENDENCIES

I will want to create a Non Gobernamental Organization called "NAME"

Figure 3.32: Example of a post-modifying reduced clause

an adjective modifies a noun. This rightmost heuristic could be violated if a noun to the left is clearly (semantically) driving the compound, but we expect mostly all compounds to be right-headed.

Compounds can have embedded structure: for example, in air conditioner factory, air modifies conditioner, and conditioner modifies factory. If you cannot tell what the structure should be, then annotate it "flatly," where the potentially ambiguous words modify the rightmost noun head. For example, in chief executive officer, it seems to make the most sense to attach both chief and executive to officer.

The policy applies also when an adjective is driving the formation of the compound, as in environment friendly in figure 3.33.

Compounds: potentially proper nominal compounds

Proper name compounds are treated in a similar fashion as regular compounds, but the NAME label is used instead of MOD. This only applies to premodifiers and not to prepositions. Consider figure 3.34 where we have a proper name: the previous noun Kelley is a NAME dependent of School, but the following prepositonal phrase of Business is treated as any other.

Proper names can also be embedded, as in figure 3.35. If you cannot determine the embedded structure, for proper names or regular compounds (see above), apply a flat structure, i.e., attach any ambiguous element to the head.

The POS tagging of proper nouns diverges from the syntactic annotation. The SUSANNE tagset tags words within proper names as determiners, adjectives, and common
nouns, if they have such common correlates. In *New York Central Library*, for instance, we have the tag sequence: *New/NP1 York/NP1 Central/JJ Library/NN1*. As the SU-SANNE guidelines say, “A word is tagged as a proper noun if the word is inherently a proper noun” ([Sampson, 1995](#) p. 87) and “any word ... listed in a published dictionary as a common word is tagged as a common word even if it is capitalized and identical to a proper name” ([Sampson, 1995](#) p. 88). In other words, for NAME sequences, you may still have common noun and adjective POS tags.

This property of POS tagging means that the POS tags can be applied, regardless of whether or not you know something is proper. A complication with learner language for the syntactic annotation, however, is that it is not always clear when a noun phrase is being used in a proper name fashion. Consider the bolded phrases in (97). Which of these phrases are proper and which are common? As we can see here, capitalization is but one piece of evidence for how the analysis proceeds and is not always reliable for making a final judgment.

(97) **Graduating from commercial and accountancy Faculty in Thailand and having an experience working as Product Movement and Inventory at ExxonMobil, I realized that I wanted to concentrate on "Supply chain Management."**

Our policy, then, is that in cases where properness is inconsistent and cannot be determined, to treat it as non-proper. This means:

1. annotating the lemma as lower-case instead of upper-case
2. using MOD dependency labels instead of NAME

If capitalization is consistent and a proper noun combination can be made, then let the evidence of capitalization lead you to posit a proper noun analysis. In (98), for
example, *Indiana State* can make a proper noun phrase, even if it is not a totally sensible one. Thus, the relationship between *State* and *Indiana* is NAME, with the rightmost *State* as the head. (The POS sequence is *Indiana/NP1 State/NN1*.)

(98) Bloomington is one of cities in Indiana *State*.

**Compounds: foreign terms**

There are times when a learner uses a term which is a foreign expression, as illustrated in (99). Although the expression can be analyzed, in Spanish in this case, into subcomponents, we do not require you to know anything about the source language. Instead, in such cases, following our default strategy for proper names, we treat the rightmost as the head and everything else as a NAME (or MOD, as appropriate) dependent of the rightmost/head word. This is shown in figure 3.36.

(99) the *Instituto Del Progreso Latino* center (constructed example)

![Figure 3.36: Treatment of a foreign term](image)

3.3.3 **Coordination**

As mentioned in section 3.2.8, we treat coordination as right-branching. Coordination is one of the most difficult constructions to annotate, and we outline how to handle several challenging cases here.

**Three-item coordination**

If there are three or more items in a coordination, we use a new label to handle the cases which come between the first conjunct and the conjunction. In figure 3.37, for instance, the relation between *toy* and *baby* is that of MCOORD (modificatory coordination), whereas that between *and* and *robot* is COORD. The main difference between COORD and MCOORD is that COORD is subcategorized for (e.g., by *and*), whereas MCOORD is an adjunct and is not subcategorized for.
Conjunct phrases

You will sometimes find instances of conjuncts being paired together, i.e., common pairs such as both . . . and or either . . . or; an example is given in (100). We treat these by making both the leading word in the conjunction phrase and treating it essentially as another conjunct, as shown in figure 3.38.

(100) I believe that these 2 goals will make my life fantastic from both professional life and private life .

Important, note that both can also function as a quantifier (QUANT) of a noun, in non-coordination contexts such as both people want change.

Non-traditional conjunctions

Coordinating conjunctions are usually and, but, and or, but if a token behaves like a conjunction, we treat it as such. In (101), for instance, the slash is conjoining the tokens actor and actress. Thus, we treat it like any other coordination, as shown in figure 3.39.

(101) ... each person is like an actor / actress ...

The more general point to make is that are coordinating conjunctions beyond the canonical set of and, but, and or. In (102), for instance, the word so is functioning as a coordinating conjunction, as it is linking two independent clauses.

(102) Luckily, my uncle live in indiana so i get to studying in USA .
CHAPTER 3. DEPENDENCIES

Coordination scope

It is not always entirely clear how a coordination scopes, i.e., what exactly it conjoins with. In cases where it is unclear, bias towards coorindating the smallest possible units. Often, though, a determination can be made, using context and our other heuristics. In (103), for example, there is the question of whether year coordinates only with month (since it lacks a determiner), or whether all four time periods coordinate. In this case, even though year does not have a determiner, it is a simpler analysis to coordinate it with all four time units, as this means that the coordination structure is perfectly well-formed. This means that year has an unrealized DET on its subcategorization, but the alternative would be to have day, wekk, and month+year form an ungrammatical coordination lacking a conjunction.

(103) first, I made a schedule that I have to do a day, a wekk, a month, and year.

In some cases, as in determining whether met coordinates with come or study in (104), the decision of what to coordinate is connected to the decision of what the lemma is. Once the lemma is decided (in this case, treating it as met), it is clear that how it coordinates (here, with come). See section 2.2.2 for the discussion of this case.

(104) I come to us from Kazakhstan to study English and met a lot of people from different countrys. Most of us have different views on the same things.

Coordination of unlikes

Sometimes learners coordinate things which are of different categories, as in (105). In these cases, it is difficult to know exactly how to annotate such a non-parallel structure, and the exact annotation depends upon how these items relate to one another. We will take them each in turn.

(105) a. commerical/JJ and accountancy/NN faculty
    b. for the cousin/NN1 and his married/VVN
    c. He running/VVG and eats/VVZ (constructed example)
3.3. OVERVIEW OF ANNOTATION

In (105a), either commercial faculty or accountancy faculty sound fine on their own, and the relation in both cases between the modifier and faculty is MOD. Thus, we can build a normal tree, as in figure 3.40, and indeed, in English, people can coordinate unlikes, so this structure can be used for native-like examples, as in (106). If it indeed sounds non-native-like to you, you can mark the distributional POS tag of the second item to be consistent with the POS of the first item, i.e., accountancy would be marked as JJ. In general, mark the clearly awkward item to be consistent with the other ones in a coordination structure; if that cannot be determined, then use the first item(s) to determine the distributional POS of the later items. (For more on distributional POS in such contexts, see section 2.3.2, especially around example (51).)

![Figure 3.40: Coordination of unlikes with the same label (here, MOD)](image)

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

In (105b), the situation is quite similar, though the non-native-like is much stronger. In this case, cousin, as a noun, is a clearly valid prepositional object (POBJ). The question is whether the category for married is also a valid POBJ—as described in Category-based definitions of section 3.1.2, the important question is not whether married is valid, but whether its category is valid. Because of cases like for the broken/VVN, it seems that VVN tags can indeed be prepositional objects (arguably, elliptically). Thus, we treat married as a part of the POBJ chain and annotate it as COORD, as usual. Again, the non-nativeness is captured through the distributional POS tag (not shown here).

![Figure 3.41: Coordination of unlikes with arguably the same label](image)

If we had found VVN to be invalid, we would have marked the relation between and married as _ . To see why, consider now (105c). In this case, we are attempting
to coordinate something which on its own would be INCROOT (cf. *He running*) with something which would be ROOT (cf. *He eats*). To capture this, we eschew the COORD label and mark each element with the label it would receive independently, as shown in figure [3.42].

Given that *and* only selects for COORD elements, this clearly indicates a problem.

![Coordination of unlikes with different labels](image)

**Figure 3.42:** Coordination of unlikes with different labels

In the case of (105b), if VVN had not been a valid POBJ, then there is no valid label to use, so instead of something like ROOT, we would resort to _. Indeed, in figure [3.42] INCROOT is a specific label for ROOTs which are not correct, essentially a more specific form of the generic _.

In a nutshell, then, we use COORD when there is nothing unusual about the coordination, even if there are unlikes coordinating. But when there is something strange, either use the top-level label that is appropriate (e.g., ROOT above) or mark a distributional POS tag consistent with a surrounding POS tag. In (107), for instance, we have an unusual coordination, but one which seems valid in English for native speakers. Thus, it is annotated as in figure [3.43] with no special marking.

(107) The American people in Bloomington are friendly, honest, and like to help people.

### Missing conjunction

If a conjunction is missing, the way to annotate is very tricky. We provide an overview of possible ways to annotate, but each case is unique, so be sure to understand the principles and apply them on a case-by-case basis.

With a missing conjunction, we sometimes have a situation where multiple elements satisfy the same requirements for a predicate. Consider the constructed example (108), where at least under one interpretation happy and lovely are being conjoined without a conjunction, though the relation between the words is not entirely clear.

(108) The queen was happy lovely. (constructed example)

\[^2\]In [Dickinson and Ragheb 2011], we advocate using COORD-ROOT for such cases, but for simplicity, we use only ROOT here. The COORD can be recovered from the subcategorization of *and*.
3.3. **OVERVIEW OF Annotation**

Crucially, the morphological evidence for both *happy* and *lovely* is the same: either one could be functioning as the predicate (PRED) of the main verb *was*. Thus, we annotate both as PRED, as in figure 3.44, since the evidence is consistent for both. (See section 5.1 for more on what to do with missing elements.) In this case, it is not entirely clear that the items are being coordinated (cf. a possible interpretation of *happy* modifying *lovely*) and no evidence of such (e.g., a comma or semi-colon), so the safest bet is to stick with only the evidence at hand and treat the items as unique PREDs of the verb.

![Figure 3.43: An unusual but valid coordination of unlikes](image)

Sometimes, however, it is clear that the words are connected or that you have a list of items. Then, we prefer to treat it as either as coordination (COORD) or as enumeration (ENUM), depending upon the resultant (un)grammaticality. In (109), for example, we have two verbs, and the comma gives a stronger indication that there is some coordination/enumeration-type relation involved. Thus, instead of multiple ROOTs, this could form an enumeration or be a coordination without a conjunction. If you can fully determine that the sentence is ungrammatical, because of the missing conjunction, then treat it as a coordination. However, if you can determine that no conjunct is required, or if it is not entirely clear, then give the learner the benefit of the doubt and mark it as an

![Figure 3.44: Possible missing conjunction, where the words are not clearly connected to one another](image)
ENUM relation.

(109) I was born there, finished school.

In this case, it seems that it is ungrammatical. The solution for this is in figure 3.45, where instead of using the adjunct MCOORD relation, we use the argument COORD relation between *was* and *finished*. This means that COORD is being used without being selected for, as only conjunctions like *and* subcategorize for COORD (see section 3.4).

![Figure 3.45: Missing conjunction, where the words are more clearly connected to one another](image)

If you have a string of items which is ungrammatical, treat all of them except the last one as MCOORD, and then make the last one COORD. This is illustrated in figure 3.46 for the sentence in (110).

(110) A lot of things have influence on people, things like hometown, tradition, family, friends.

![Figure 3.46: Missing conjunction, with a string of items](image)

On the other hand, if the sequence is grammatical, then mark it as ENUM. In (111), for instance, while there are issues with the usage of determiners, the enumeration structure is fine, as indicated by the *etc*. Thus, it is treated as a sequence of ENUM relations, as in figure 3.47.

![Figure 3.47](image)
3.3. OVERVIEW OF ANNOTATION

and I sing in church, street, station etc.

(111) in church, street, station, etc.

Figure 3.47: A proper ENUM structure

In some cases, the evidence is so unclear that the only thing to do is to mark a relation as 

In (112), for example, it is clear that swim and running coordinate (see section 2.4.2 around example (61), for more on handling their disagreement in forms). However, it is not clear how to treat readings: as a misordered MCOORD, a separate COORD, part of an enumeration structure, or even as an appositive (see section 4.8). Thus, we treat it as a dependent of running, as shown in figure 3.48.

(112) Besides, I like swim and running, readings.

Figure 3.48: An unclear element related to a coordination structure

Independent clauses As a special case of a type of missing conjunction analysis, we have two possible cases involving independent clauses: 1) properly-demarcated independent clauses (e.g., connected by a semi-colon); and 2) improperly-demarcated independent clauses (e.g., run-on sentences, comma splices). Paralleling the discussion above, properly-marked clauses are joined by ENUM, while improper ones are joined by MCOORD/COORD (with COORD being unselected).

An example of a run-on/comma splice is given in (113) (=20), with the corresponding tree in figure 3.49.

(113) They taught me usual things, these were not grammar.

This type of broken coordination structure should be used only when: a) the sentences are somehow connected, largely by some discourse cohesion (vs. a less likely
They taught me usual things, these were not grammar.

Figure 3.49: Multiple sentences connected by a (non-subcategorized) COORD

Figure 3.50: Non-constituent coordination example
selecting for DET relations. However, this allows our tree to maintain local relations and to make sure that every determiner is selected for only once.

![Diagram of tree structure with DET, MOD, COORD, DET, and CCC labels.](image)

**Figure 3.50**: Non-constituent coordination (DET + MOD)

In this case, the adjective, being more content-ful, is the item which selects for the determiner, paralleling cases where adjectives really can take determiners (e.g., *We’re in the red*). In cases where the more content-ful item cannot be determined, we prefer a structure which minimizes non-projectivity. In (116), for example, it is not clear whether the OBJ (Walt, Malcolm) is more primary than the OBJ2 (the salt, the talcum).

(116)  ... give Walt the salt and Malcolm the talcum \footnote{Steedman and Baldridge 2011}

To avoid crossing branches, we thus coordinate the salt and the talcum, as shown in figure [3.51]. In this case, talcum then takes Malcolm as an OBJ. If we tried to coordinate Walt and Malcolm, then the arc between give and salt would cross the arc between Walt and and.

![Diagram of tree structure with OBJ, OBJ2, OBJ, DET, COORD, and DET labels.](image)

**Figure 3.51**: Non-constituent coordination (OBJ + OBJ2)

### 3.3.4 Arguments vs. Adjuncts

Some labels depend upon being able to distinguish adjuncts from arguments (sometimes called complements). For example, if you see a preposition dependent upon a verb, it
is either IOBJ (argument) or JCT (adjunct). Here is a reminder of what the tests are for argumenthood; if you are unfamiliar with these, we can point you to the relevant syntactic literature. The same criteria are also used to determine whether something should be included on a subcategorization list, as these are only for arguments (see section 3.4).

These diagnostics are generally true, so if you face a difficult decision, try as many as you can to see what the majority rule is:

1. Iterability: Arguments cannot be iterated; adjuncts can be iterated.
2. Optionality: Arguments are generally obligatory; adjuncts are generally optional.
3. Lexical sensitivity: Arguments can be highly specific lexical forms, dictated by the head (e.g., depend requires on(to)).
4. Reordering: Arguments cannot be reordered; adjuncts can.
   Relatedly, arguments tend to be closer to the head than adjuncts.
5. Conjunction: Arguments tend to coordinate with arguments, adjuncts with adjuncts.
6. Pro-form replacement: adjuncts can follow pro-forms like do so (for verbs) or one (for nouns), while arguments cannot. In other words: pro-forms like do so must include the argument in what they refer to.

We will go through several examples below, but to take one example here, in (117) it is clear that with is dependent on live, but is it IOBJ (argument) or JCT (adjunct)? In this case, even though in full of love is oddly-formed, note that the sentence sounds equally grammatical by saying to live in full of love with my family (reordering) or by dropping the phrase, as in to live in full of love (optionality). These indicate that the phrase is an adjunct, making JCT the label.

(117) to live with my family in full of love

The distinctions we need to concern ourselves with are:

- IOBJ (arg.) vs. JCT (adj.)
- COMP (arg.) vs. CJCT (adj.)
- XCOMP (arg.) vs. XJCT (adj.)

Note that these are all verbal (or adjectival) relations. For nouns, we generally do not maintain an argument/adjunct distinction; e.g., NJCT, MOD, CMOD, and XMOD all apply to nominal dependents, regardless of whether they are arguments or adjuncts (see section 3.1.1). Let us now consider some different types of examples.
IOBJ vs. JCT

In (118), we have two prepositional phrases to consider, headed by from and to. Note how both phrases are optional (test #2)—i.e., the clause is grammatical if it ends at position—and how they can be reordered (test #4): change my position to finance from marketing. You can switch this to a coordination structure easily (test #5): change my position from marketing and to finance. And a do so replacement does not need to include either preposition (test #6): I will change my position from marketing to finance, and you will do so with whatever majors you want. Assuming these judgments are decent for you, they all point to both to and from being adjuncts of change, resulting in a JCT label.

To obtain various skills and experience, I will change my position from marketing to finance, because ...

Consider now (119), where we also find JCT. We know this because: about goals is optional (#2); it is not lexically sensitive (#3)—cf., e.g., we spoke regarding goals; other adjuncts can be inserted between these (#1/#4)—e.g., we spoke in confidence about goals; and do so does not have to include the preposition—e.g., we spoke heatedly about goals on Tuesday and they did so regarding finances the very same day (did so = spoke heatedly).

With my husband, we spoke about goals.

Note in this case how the conjunction test (#5) gives mixed results: e.g., we spoke about goals and on Tuesday sounds a bit odd. If some tests conflict or provide dubious information at best, then be sure to get the overall picture with multiple tests.

To see a case which is IOBJ, consider (120), where in is dependent upon focused. Although the in phrase is technically optional (#2=adjunct), note that the choice of in is odd, indicating some degree of lexical sensitivity (#3=argument). The relative oddness of ??I focus on fluctuating assymetry, and Maria does so on chemistry also indicates argumenthood (#6=argument). The reordering test (#4) is mixed, with ??I focused last year on fluctuating assymetry sounding semi-okay. Given all the evidence, there is a favoritism towards labeling as IOBJ.

My diploma work is focused in fluctuating assymetry.

Note how the IOBJ/JCT distinction is largely a syntactic one. In (121), for example, the as prepositional phrase distinguishes what seems to be a key semantic property, namely the instrument by which “them” are being used. Syntactically, however, if you do the tests, you will find that it is an adjunct (JCT). (e.g., inserting another adjunct in I used them each and every day as a general tool).

I used them as a general tool.

In general, if you are not quite sure, bias towards JCT.
COMP/XCOMP vs. CJCT/XJCT

Similar reasoning applies for verbal clauses which modify verbs, distinguishing between COMP (finite) or XCOMP (non-finite) for arguments and CJCT (finite) or XJCT (non-finite) for adjuncts. Other than the finiteness issue, the same reasoning applies for COMP vs. CJCT as for XCOMP vs. XJCT.

The distinctions here tend to be easier than for prepositions, as to-phrases for example are generally very clearly adjuncts or complements and required by a verb. Note in (122), for example, how to obtain financial aid cannot be moved to the beginning of the sentence, indicating its tightness with the verb and thus its argumenthood (the other tests concur).

(122)  
a. XCOMP: For this reason I need to obtain financial aid.
   b. XJCT: ... to win this financial aid ... I have been making investigation about that subject.

Adjectives

As mentioned in section 3.2.4, we use the same dependency categories for adjectives as for verbs, but only the non-argument ones: JCT, CJCT, and XJCT. Thus, in (123), the relation between good and for is JCT, while in (124), the relation between glad and to is XJCT. There is thus no need to perform any tests.

(123)  
It not only good for your body health, but also good for your mind.

(124)  
I glad to be in Bloomington ...

3.3.5 Punctuation

The general rule for punctuation, annotated as a PUNCT relation, is to attach each piece of punctuation to the (previous) head over which it takes scope. The question is usually in how to define the scope of the punctuation.

End of sentence markers  End of sentence markers include full stops (periods), question marks, and exclamation points. For these markers, our general policy is to attach them to the head of the entire sentence, i.e., the word which is the ROOT dependent. If for some reason there happen to be multiple ROOT elements, attach to the closer one. The label used in PUNCT.

Essentially the same policy is applicable for the attachment of colons and semicolons (i.e., to previous head), though the overall tree is an ENUM structure (see Missing conjunction in section 3.3.3). Note that if there is a conjoined sentence, the head that it takes scope over, given our definition of coordination, is the first conjunct.
Given that we can combine clauses into a single sentence, with a sentence-internal period (see section 2.1.1), there is the question of where to attach a sentence-internal period. In (125), for instance, there is a period in a strange position. However, in such cases, we still look for the head of the entire sentence—which potentially could come after the period. In this case, it is still do, for both periods.

(125) Sometime I do n’t like it because it ’s very complecated and in my country the weather . It ’s very hot .

**Sentence-internal commas** Commas are attached to the previous head over which it takes scope. In (126), for example, the comma attaches to the head of the whole subordinate clause; in (127), the comma attaches to the only previous word it can, Also.

(126) Although I ’ve developed my business skill in marketing for past seven years , I found what to further improve . . .

(127) Also , I need more natural and friendly place . . .

(128) . . . to take advantage of environment friendly resources , such as nuclear power , hydro power and wind power .

The important distinction here is that of what a comma scopes over, and this is dependent upon what items are being connected by the comma. Example (128) is a bit more challenging in determining what the commas scope over. Note, though, that there is a (appositive) relation between resources and such as, and there is also a coordination relationship between nuclear power and hydro power. These relationships indicate that it is the nouns which the commas scope over. A good rule of thumb is that, if possible, you do not want to create non-projective structures when attaching commas.

We can see this rule of thumb in (129), examining the comma between life and no. The phrase no only as a prison is a modifier of important (see section 6.10 for a full analysis of this sentence). This helps us to see that the comma is scoping over important, i.e., the importance is as a prison/person. If we tried to attach to want or tell, for instance, there would be crossing branches, since no only as . . . links to important.

(129) After to start, I want to tell that this exercise is very important in the life , no only as a prison .

The same principle applies when there are coordinate structures: the applicable head is then the first one, which heads the entire coordination. In (130), for example, the comma attaches to Graduating because that heads the entire subordinate phrase.

(130) Graduating from commercial and accountancy Faculty in Thailand and having an experience working as Product Movement and Inventory at ExxonMobil , I realized that I wanted to concentrate on “ Supply chain Management . ”
Quotation marks Quotation marks attach to the head of the item they are quoting, either to the previous or the following head, and given a PUNCT label. The same is true for parentheses (and dashes which are used similarly to parentheses). In example (130) above, for instance, the quotation marks both attach to Management. Note that since the period attaches to the head of the whole sentence, we have a non-projective structure.

![Figure 3.52: Partial tree showing treatment of quotation marks](image)

### 3.4 Subcategorization frames

One of the most important dependency annotations for each sentence is the subcategorization information, which is marked for each token which selects for a specific kind of other word. For example, verbs select, or subcategorize, for subjects (SUBJ); transitive verbs also select for objects (OBJ); singular count nouns subcategorize for determiners (DET) or quantifiers (QUANT); etc.

One of the issues we annotate subcategorization for learner language is that learner language often contains violations of subcategorization (sometimes referred to as argument structure). While dependencies show what is realized in a sentence, annotating subcategorization allows you to mark what is selected, whether or not it was realized, or realized correctly.

The selection of a word’s subcategorization frame is based on the normal usage of a word in English (i.e., what is grammatical, a point discussed below). Annotating subcategorization means that you are annotating the dependencies that a word selects for if it were to result in a grammatical sentence. For example, for some instances of give, you would annotate SUBCAT1=SUBJ, SUBCAT2=OBJ, SUBCAT3=IOBJ. We use the same labels for subcategorization as with dependencies (section [3.1.1](#)), but only for arguments, not for adjuncts (discussed more below). Note also that the subcategorization annotation is based on the lemma, i.e., whatever word you have already decided is present (section [2.2](#)).
Words are often ambiguous, and thus the subcategorization is based on the use of the word in the specific context of the sentence. For example, the word *house* can be a noun or a verb in English, so we list the subcategorization that fits the context of the sentence. In the constructed example (131), *house* is a singular noun and thus selects for a determiner, so the value of SUBCAT1 is DET.

(131) This is my **house**.

Note how this allows you to capture learner innovations. In (132), *house* requires a determiner, yet does not have one. The annotation is shown in figure 3.53, where *house* selects for a determiner in its subcategorization list. No determiner is realized, whereas *to* subcategorizes for a prepositional object, which is realized.

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

![Diagram](https://example.com/diagram.png)

Figure 3.53: Example showing one subcategorization that matches the realized dependency (*to*’s selection of POBJ) and one not (*house*’s selection of DET)

Returning to the issue of annotating only arguments, note in this example how MOD is not selected for by *house*, nor does *other* select for *house* in any way. In section 3.1.1, the dependencies which can be selected for are printed in regular font and those which are not subcategorized for are printed in italic font.

### 3.4.1 Grammaticality

For (morpho-syntactic) dependency annotation, you are annotating based on the evidence at hand, without too much regard for whether the sentence is grammatical or not. For subcategorization, you are to a greater extent concerned with grammaticality, as you are marking what should be in a sentence.

Always bear in mind, though, that you are not marking a correct sentence; you are marking what the verb is selecting for. In (133), for instance, the verb *Want* finds what appears to be a prepositional phrase after it (*to both of them*), making it IOBJ.

---

3In trees, we illustrate subcategorization using angled brackets, corresponding to a list. <X,Y>, for example, represents SUBCAT1=X, SUBCAT2=Y.
CHAPTER 3. DEPENDENCIES

Want, however, should be selecting for something that it can grammatically select for in English, like XCOMP or OBJ. In this case, OBJ seems intuitively closer to IOBJ, to we prefer that. In other words: the dependency between Want and to is IOBJ, and SUBCAT2=OBJ. But you do not have to specify where the OBJ should be realized or what the target form should be (which is not clear here, anyway). You simply state that one thing is found (IOBJ) and another thing is being looked for (OBJ).

(133) Cause I like work and hobby, I Want to both of them.

In general, we define grammatical as being “standard English”, of the type learned in English classrooms. There is some variation in what this might mean, but most cases are clear. Consider (134), where it seems like man is selecting for a DET which is not realized. Indeed, this is how we annotate the sentence, as shown in figure 3.54.

(134) I’m not boring man.

One could, however, annotate this sentence based on the more colloquial I’m not boring, man, with man as a vocative or interjectory form. However, this would be annotating the sentence on a rather non-standard and informatl form, so we do not advocate this particular analysis.

\[ \text{Figure 3.54: Annotating based on standard English} \]

3.4.2 Determining subcategorization requirements

The basics of annotating subcategorization may seem clear, but there are several issues to take note of: 1) A dependency being subcategorized for is not the same as a dependency being required; 2) Context must be used to obtain a single subcategorization annotation; 3) There is a strong relation between subcategorization and distributional POS; and 4) The criterion of pronouncability serves to distinguish semantic/logical arguments from the syntactic relations we subcategorize for.
Subcategorized ≠ Required

Items that are subcategorized for are arguments. But recall from section 3.3.4 that arguments are not necessarily always required; i.e., you cannot simply test for argument status by removing an item and seeing if it is no longer grammatical.

In fact, for many categories, you will simply always put them onto a subcategorization list. In section 3.1, we enumerate the dependency labels, where the non-italicized ones are arguments and the italicized ones are adjuncts; if the label is an argument label and is being used correctly, it goes in the SUBCAT annotation. For example, both DET and QUANT are the types of categories which are selected for (see section 3.3.2). In (135), we find enough as a QUANT dependent of time. Because we treat QUANT as an argument type of category, then you should automatically add it to the subcategorization of time if a QUANT relation is present and used correctly.

(135)  I would like to have enough time

If the QUANT relation is present, but is not being used correctly, then it will not be included. This is how we proposed to handle phrases like my all children, for example, as shown back in figure 3.27.

More on using context

As mentioned, you should annotate the subcategorization frame which best fits the context of a given sentence. In the constructed cases in (136), for example, handed would receive different annotations. In (136a), it is <SUBJ, OBJ, OBJ2>, while in both (136b) and (136c), it is <SUBJ, OBJ, IOBJ>. For (136c), 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, OBJ2>, where two elements would be wrong. This stems from the heuristic of giving the learner the benefit of the doubt.

(136)  a. Max handed his sister the wagon.
        b. Max handed the wagon to his sister.
        c. *Max handed to his sister.

As a side point, we do not treat order as important, so marking SUBCAT2=OBJ and SUBCAT3=OBJ2 is the same as SUBCAT3=OBJ2 and SUBCAT2=OBJ.

Note how the context of the whole sentence is important. In the constructed sentence (137), the first instance of School requires no determiner, but the second one does. Remember that you do not have to explain what exactly triggers the difference, but simply mark the subcategorization as appropriate.

(137)  a. School is fun
        b. School is on fire
Likewise, in (138), in this context, *arts* seems to require a DET, even though it might not in other contexts and even though plural nouns do not often require determiners.

(138) Truth be told, I’m loving music and *arts*.

This is part of the general principle that you must consider not just the immediate context, but the context of the entire sentence. Consider (139), in particular the phrase *a day* (the other temporal expressions work similarly). In this case, the subcategorization for *day* could be SUBCAT1=DET, which would match the local context of having a determiner, *a*, precede it. On the other hand, given the context of the entire sentence, where *day* modifies *do*, this is not accurate.

(139) first, I made a schedule that I have to do *a day*, a week, a month, and year.

In this context, *day* would more appropriately select for a quantifier, SUBCAT1=QUANT (cf. *every day*), or for nothing (cf. *by day*). Thus, we annotate it as SUBCAT1=, which says that the exact requirement is unknown, but that something is required. (In some similar contexts, one might lean more towards QUANT, and real labels such as that are preferred over _.)

Balancing the different contextual pieces of information can become very challenging for words which have a lot of potential subcategorizations and for situations with usage that is slightly odd. Consider the verb *serve*. This verb can take direct objects (OBJ) or second objects (OBJ2, cf. *they served us finger food*), or even no object (cf. *It’s your turn to serve*). Certain types of prepositional phrases (headed by *in* and *for*) are close to indirect objects, but are likely better treated as JCT dependents of an intransitive *serve*.

Now consider the learner instances in (140), where each *serve* verb has a JCT for-phrase following it. There is a strong sense *serve* would sound better without the *for*, i.e., as an object (OBJ) analysis (i.e., SUBCAT2=OBJ). Yet, syntactically speaking, *for* phrases are acceptable here (e.g., *they can serve for twenty years*).

(140) a. ... after that they can *serve* for their country and for their people.

b. we can live as brothers and sisters and love each other and *serve* for each other.

Such *for*-phrases, however, are only licensed with particular types of (temporal) nouns, i.e., this is more of a lexical type of property. Indeed, if we replace *for* with the complex preposition *on behalf of*, the sentence is grammatical. The solution is thus to annotate no subcategorization for OBJ, but instead to mark *for* as a lexical violation.

**Connection to distribution**

It may help to annotate subcategorization in tandem with distributional POS since what words select for (subcategorization) in large part determine how words are being used...
distributionally (distributional POS, section 2.3.2). In (141), for example, the issues of what *is* selects for and what *have*’s distributional POS tag is are intertwined.

(141) Bloomington is **have** very nice and flexible weather

In this case, *is* selects for SUBCAT1=SUBJ, SUBCAT2=VC (and not, e.g., SUBCAT2=PRED) because there is a following verb, and *have*’s distributional POS is an underspecified VH, appearing as it does after *is*, which can select for different verbal forms. In other words, the subcategorization of VC from *is* helps determine that *have* has distributional evidence of a verb form other than its morphological base form (VH0).

One difference between distributional properties and subcategorization is that subcategorization is lexically-based, while distribution is defined by where appropriate categories go. In (142), for instance, *experience* does not select for a determiner (DET), yet we can still say that *an* is in an AT1 distributional slot, as NN1 nouns (such as *experience*) predict the occurrence of a determiner to the left. In practice, this distinction typically only arises in contexts such as this one, with an extraneous word.

(142) having **an experience** working as Product Movement and Inventory at ExxonMobil

**Pronouncaability**

We include arguments on subcategorization lists only for items which, in some sense, are “pronounceable”. Consider (143), where the verb *Graduating* heads a subordinate clause that modifies the main clause. Such gerund uses do not take overt subjects—even though they can often be recovered from context—and therefore we do not annotate any SUBJ subcategorization. Typically, such non-finite verbs (gerunds and infinitives) head clauses that modify other words and do not (or cannot) require overt subjects.

(143) **Graduating** from commercial and accountancy Faculty in Thailand and having an experience working as Product Movement and Inventory at ExxonMobil, I realized that I wanted to **concentrate** on "Supply chain Management."

The verb *concentrate*, on the other hand, occurs within a **verbal chain** (see *Heads for verbal chains* in section 3.2.5). In these cases, the subject is always syntactically present (in this case, the second *I*), and so *concentrate* selects for SUBCAT1=SUBJ, which is realized here, the same subject as for *wanted*.

Note that predicates (e.g., XPRED) are cases where a subject is not selected, largely because there is often not a connection between the subject of the infinitive and the subject of the sentence. In (144), for example, the subject of *study* (which is an XPRED of *is*) is not directly realized in the sentence. Again, even if there is logically, or pragmatically, a subject for *study*, it is not something which is pronounced, thus not being subcategorized for, in our sense.
One of the greatest ways to realize my plans for life is to study business at Indiana University.

### 3.4.3 Specific cases

#### Interaction with coordination

When there is coordination, you often have several words subcategorizing for elements, which are then realized further away. To revisit the *Coordination* part of section 3.2, in a coordination structure, *and* selects for COORD (i.e., SUBCAT1=COORD), while CCC and MCOORD are not subcategorized for (i.e., are treated as adjuncts).

Once you know what words subcategorize for, the key with coordinations is to make sure that every item which selects for something realizes it (if indeed it does). An example is shown in figure 3.55, where *He* is selected for by three verbs, all of which have links to it. Note, too, how COORD is selected, and CCC is not.

As another example, in (145), the word *urge* has a subcategorization of: SUBCAT1=SUBJ, SUBCAT2=OBJ, SUBCAT3=XCOMP, all of which are realized (*I*, *people*, and *to (burn)*, respectively).

(145) ...I would like to further promote energy-efficient appliances, and *urge* people not to burn fossil fuel ...

Bear in mind that you should only link the subjects that the coordination structure permits. In (146), for example, *is* and *get* are coordinated because they are the only finite verbs that can be coordinated. In this type of finite verb coordination, *get* subcategorizes for a subject, and that subject is logically something like *MBA graduates*, but we cannot link *get* to *graduates* here, because the only available subject is the subject of the coordinated verb, *is*.
3.4. SUBCATEGORIZATION FRAMES

(146) According to a survey, the average salary of MBA graduates is 10-20% more than ordinary people in marketing and get more chances to be promoted to manager positions.

In this case, one could choose to link a SUBJ between get and salary, but the verb does not agree with salary, thus indicating that they may not go together. In this case, one may thus annotate SUBCAT1=SUBJ, but then give it no SUBJ.

Complementizers

First, note that “complementizer” refers to true complementizers, subordinating conjunctions, and also (in our usage) relative pronouns. In all these uses, complementizers (CPZR) are subcategorized for by their head verbs, if they are present or if they are otherwise required; this is akin to how determiners are selected by nouns. In (147), for instance, attending subcategorizes for CPZR, which is realized as by.

(147) I believe we will have great time in Bloomington by attending a lot of great events and parties, . . .

This may seem a bit odd at times, but it allows us to specify situations where a complementizer is definitely required or definitely not allowed. To take a few examples: in the constructed I asked you should leave, if asked takes a finite complement (COMP), such a complement must have a complementizer (cf. whether, how, etc.). Perhaps more clearly, relative clauses often require relative pronouns—which are treated as complementizers. In figure 3.56, for instance, the sentence is ungrammatical without who, and thus CPZR is needed on the list.

Figure 3.56: Complementizer treatment for relative clauses (constructed example)

There is at least one facet of this analysis which may be counter-intuitive, namely which word selects for the complementizer. It could be argued in this case that the noun
does, but to enforce consistency across different kinds of cases, we have the verb of the subordinate clause select for the complementizer.

There are also cases where a complementizer is forbidden, such as in the so-called that-trace effect. For example, *Who do you think (*that) hurt me?* is grammatical if *hurt* does not take a complementizer.

Note, too, in figure 3.56 how *person* is both the subject of *is* and the subject of *lives*. See section 4.4.4 for more on relative clause handling.

**Reminder: argument and adjunct tests**

Remember that you must distinguish arguments and adjuncts for this labeling. For example, the labels IOBJ and JCT are both used for prepositions which are dependents of verbs; the difference is that IOBJ is subcategorized for, while JCT is not. Recalling the distinction of arguments vs. adjuncts (section 3.3.4): if the tests point to a preposition being used as an argument, use IOBJ and include it in the subcategorization; if the tests point to an adjunct use, use JCT and do not include it in the subcategorization annotation.

For example, in *depend on it*, the verb *depend* requires a preposition, so include IOBJ in the subcategorization, while *hunt on Tuesday* has a JCT relation between *hunt* and *on*, which is not included with subcategorization.

This can be challenging for phrases such as *talk about X* (JCT) and *give X to Y* (IOBJ), so be sure to: a) try all the tests in section 3.3.4, and b) be consistent in your use of IOBJ and JCT with respect to whether they appear in subcategorization.
Chapter 4

A Variety of Dependency Constructions

In this section, we cover a variety of dependency constructions, which tend to be difficult to analyze linguistically (i.e., regardless of learner innovations). Learner innovations are handled more in chapter 5.

4.1 Attachment decisions

There are certain constructions that are always difficult to disambiguate, for native or non-native language. For example, preposition attachment is always a challenge—cf. the often-used I saw the man in the park with a telescope, which has at least five interpretations. If you cannot disambiguate the attachment, we favor an heuristic of attaching “high” (cf. Bies et al. 1995, p. 102). In this example, if you cannot tell whether it is the park, the man, or the seeing event which involves a telescope, then attach with a telescope to the word which covers more of the tree, i.e., saw in this example.

There are two things to keep in mind, however, which can be illustrated in example (148). The questions here are: 1) where does in my life attach (to made or series steps)? and 2) where does in Almaty attach (to made, series steps, or life)?

(148) I made my first series steps in my life in Almaty

The first point to know is that you should first try to disambiguate the attachment, by seeing which phrases must occur together. In other words, try moving each phrase around with and without the adjacent preposition. In (149a), we see that in my life cannot be found modifying the verb without the noun phrase present, whereas it is fine moving with the noun, as in (149b). But when we try to have in Almaty modify anything other than the verb made, it cannot, as in (149c).

(149) a. ??My first series steps, I made in my life in Almaty
CHAPTER 4. A VARIETY OF DEPENDENCY CONSTRUCTIONS

b. My first series steps in my life, I made in Almaty

c. ??My first series steps in my life in Almaty, I made

These movements indicate that

1. in my life modifies series steps (NJCT)

2. in Almaty modifies made (JCT)

The second point is to avoid non-projective structures. Thus, even though it is clear that in my life modifies series steps, it is not entirely clear which word it modifies. Both words are dependents of made (see section 5.2.2), so if in modifies series, there will be a crossing branch. Thus, we attach in (in in my life) to steps. This heuristic will often lead to annotating the closer possible head.

Similar to prepositions, conjunctions can introduce ambiguities. In (150), for instance, it is not clear whether the bolded clause conjoins with the main clause (Sometime I do n’t like it because it ’s very complecated) or only the embedded clause (it ’s very complecated). (Section 2.1.1 discusses the sentence segmentation here.)

(150) Sometime I do n’t like it because it ’s very complecated and in my country the weather . It ’s very hat

Again, if context cannot disambiguate, favor attaching high, i.e., coordinating with the main clause.

For more information on where to attach modifiers to verbs, see the part on Verbal modifiers in section 3.3.1.

4.2 INCROOT

In general, INCROOT is used when the head of the whole sentence is not a typical, valid head of the sentence (i.e., a finite verb). In these cases, you need to identify the most head-like element and make that the INCROOT. Usually, this is the item from which the meaning of the sentence is mainly derived. Often the INCROOT word is the word which is the “topmost” remaining word. In (151), for example, the leftmost in is, in lieu of better evidence, the head of this fragment.

(151) in powrful University in law .

One important point is that predicates are heads of fragments when a verb is absent. In (152), for example, there is no verb, but the preposition about forms the head of a predicate, thus making it the INCROOT.

(152) Some of thim about my studies and my job .
4.3 Extraposition

Extraposition, as it is commonly known, is not a major problem for dependencies, as it involves an element appearing farther to the right than expected. In *The person is old who lives here*, for example, the relations between *person* and *lives* is exactly the same as in figure 3.56 (i.e., CMOD), even though *lives* is now much more rightward. This is shown in figure 4.1.

For *it*-extraposition, we want to ensure some connection between the subject *it* and the extraposed *that*-clause which appears later. In (153), for instance, we see that *It* refers to *that learning more means success in the future*. We do not capture this relationship directly, but mediate it via the verb *is*. As shown in figure 4.2, the verb *is* selects for not only a subject and a predicate—as with other copula instances—but also an additional clausal predicate (CPRED).

(153) It is my parents’ opinion that learning more means success in the future . . .

If we had the corresponding non-*it*-extraposition case of *learning more means success ... is my parents’ opinion*, then we would simply have *means* be the CSUBJ (finite clause subject) of *is*.

This can happen with non-finite clauses, as well. For example, in (154), the verb ‘*s has three arguments: SUBJ (*it*), PRED (*important*), and XPRED (*to*).

(154) ... it’s quite important to get a MBA degree .
It is my parents' opinion that learning more means success in the future.
4.4 *wh*-words

In section 3.4 under Complementizers, we discussed the treatment of relative pronouns, an example of a word being displaced. Here, we examine other displacements more generally, including the use of *wh*-words and other long-distance dependencies.

4.4.1 Displacement

If you have a simple displacement, such as a topicalized noun, then you do not do anything special, but rather give the displaced element the grammatical relation it would have anyway, as illustrated in figure 4.3. The only special property is that you might wind up with a non-projective structure (i.e., crossing branches), as shown in the next section.

![Figure 4.3: A displaced (topicalized) element (constructed example)](image)

However, if the argument is repeated, there is a special notation. As mentioned back in section 3.1.1 for the TOP(ic) label (see example (70)), there are cases where a word is made a topic, but doesn’t fulfill an argument elsewhere. In these cases, a TOP label is used, as illustrated in figure 4.4. Crucially, the arguments for *catch* are already filled (by *I* and *them*), so this is not a displaced or topicalized argument, but rather an extra topic.

![Figure 4.4: A topic (≠ topicalized) element](image)
4.4.2 *wh*-questions

For *wh*-questions, we have the same idea of linking the *wh* to its a lower predicate. This is illustrated in figure 4.5. Here we see that, as with declarative sentences, both the auxiliary (*does*) and the main verb (*love*) take the same subject. The relations in general are the same as with declarative sentences, with *Who* serving as the object of *love*.

![Figure 4.5: A simple *wh*-question (constructed example)](image)

In some theories, there might be a direct relationship between *does* and *Who*, but we do not specify such a relation here. Note that the general analysis holds, regardless of the distance, e.g., the same OBJ relation is found in *Who did Kim say that Charles thinks Sandy loves?* Additionally, you can see that this introduces crossing branches, a point you will no doubt notice with most every *wh*-question.

4.4.3 Embedded clauses

*wh*-clauses functioning as an argument of a verb can be more challenging to annotate, though they start with the same principle. In examples like the constructed *I wonder who(m) Sandy loves*, for example, *who* is still an OBJ of *loves*, regardless of the rest of the analysis. In this case, we treat *loves* as a COMP of *wonder*, and essentially nothing more needs to be said. We refer to this as the clause analysis.

![Figure 4.6: An argument *wh*-clause](image)

But note the similar-looking, but differently-functioning phrase in (155). In this case, *What* is still going to be a dependent on the embedded verb, *am*, in this case
4.4. *WH*-WORDS

functioning as a PRED of *am*. And as for the relation to the higher verb, a verb like *made* can take a nominal or adjectival predicate (PRED) as a second argument (e.g., *made me a human*), so it seems like something in the *What I am* clause should be a PRED of *made*.

(155) My hometown made me **What I am**.

We thus annotate this as in figure 4.7, where *What* is the PRED of *made*. We choose *What* to serve in this role because *made* selects for nouns and noun-type things. Then, similar to relative clauses (see section 4.4.4), *What* takes the *I am* as a CMOD. We refer to this as the word analysis. Note that we do not put CMOD onto the subcategorization list of *What*, though one might argue that it is necessary in this case.

We list diagnostics for distinguishing these two analyses in table 4.1. In essence, you must determine the usual subcategorization of the verb and thus whether it is selecting for the *wh*-word or for an embedded sentence.

If there is a non-finite clause, you have the same tests. In (156), for instance, the non-finite *what to say* is in a slot where nouns can appear (*I know things*); perhaps slightly less predominant, non-finite clauses can also appear here (?*I know (enough) to finish the assignment by tomorrow*). This particular sentence can be rephrased with a relative, e.g., *I know the thing I (will) say* or *I know what I want to say*, giving an inclination towards the word analysis. Contrarily, the *wh*-replacement tests also work (*I know why/how much to care*), though this likely reflects the fact that *know* can takes lots of types of arguments (e.g., *I know enough*). Overall, the evidence seems more in favor of a word analysis, as shown in figure 4.8, the main difference with the previous analysis is the use of XMOD instead of CMOD.

(156) I know **what to say**. (constructed example)
CHAPTER 4. A VARIETY OF DEPENDENCY CONSTRUCTIONS

<table>
<thead>
<tr>
<th>Category</th>
<th>S non-S replacement</th>
<th>wh-replacement</th>
<th>Rephrasing: with relative</th>
<th>Solution:</th>
</tr>
</thead>
<tbody>
<tr>
<td>e.g., wonder whether I am happy, *wonder happy, *wonder a human</td>
<td>e.g., made me happy, made me a human, *made me whether/that I am happy</td>
<td>Yes, e.g., wonder why I care</td>
<td>Yes, e.g., *made me why I care</td>
<td>Table 4.1: Diagnostics for determining a wh embedded clause analysis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4.1: Diagnostics for determining a wh embedded clause analysis</th>
</tr>
</thead>
</table>

With non-finite clauses, you sometimes also have raising constructions (see section 3.3.1), though these are generally a subject of the embedded infinitive clause. The verb *know* can be a raising predicate (e.g., *I know there to be a bear in the room*), but *what* is the lower clause’s object and further it is a wh-word, two indications that it is not being raised. However, note that the subject (*I*) is indeed a raised argument, being both the subject of the upper verb (*know*) and the lower (*say*), as shown in figure 4.8.

4.4.4 Relative clauses

We will continue the discussion of wh-elements and similar (displaced) data with relative clauses. (See also section 3.4 under Complementizers). The analysis here is basically identical to the so-called word analysis from the previous section, where a noun takes a CMOD (or XMOD) dependent, while at the same time functioning as an argument of a relative clause verb.

Consider (157), where *I will have* seems to be modifying *my family.* (Note that there is a lexical violation for *my,* in this context where *the* is a better-sounding choice.)

(157) *I will prepare enough economic base to support my family I will have*
4.5. PREPOSITIONS VS. COMPLEMENTIZERS

The analysis is in figure 4.9, where we see that: a) the displaced *family* is the non-local OBJ of *have*, while at the same time being the local OBJ of *support*, and b) the relative clause, headed by *will*, is a CMOD modifier of *family*. (For those who care, there is a cycle here, i.e., starting with *family*, the arrows lead you back to where you started.)

![Figure 4.9: Relative clause treatment](image)

As discussed in section 3.4 under *Complementizers*, the analysis is the same for relative clauses starting with a relative pronoun (*who*, *that*, etc.), except that the verb takes an additional CPZR argument which is the relative pronoun.

4.5 Prepositions vs. Complementizers

One confusing thing about prepositions and complementizers is that they often appear to be fairly similar for words like *as*, yet the choice leads to completely different analyses. The rule of thumb is: if the following item is a noun phrase (and only a noun phrase), then make it a preposition (generally, II, though sometimes CS forms for, e.g., *as*) which takes a prepositional object (POBJ); if the following item is a finite clause, then make the word a subordinating conjunction (CS, CSA, CSN, etc.), which serves as a CPZR dependent of the verb. There are some minor exceptions to this, when words are being used comparatively (see below), but otherwise follow this rule. Some examples were given in the section on *Gerunds/Progressive forms* in section 3.3.1, particularly surrounding figure 3.22. (See also the section on *Complementizers* with respect to subcategorization in section 3.4.)

Note that in the prepositional analysis the word like *as* is the head, while in the complementizer analysis, it is the dependent. Consider (158). In this, *as* is a preposition because a noun phrase (*an assistant manager*) follows it; the tree is in figure 4.10.

(158) **As** an assistant manager of one of the largest utility companies in the world, I would like to ...

An example of the subordinating conjunction analysis is given in figure 4.11, where the entire clause is a subordinating clause, modifying the sentence (CJCT). In this case,
CHAPTER 4. A VARIETY OF DEPENDENCY CONSTRUCTIONS

As is selected for as a CPZR of the subordinate verb (*know*). Remember that this type of analysis applies even if there is no subject, as long as there is a verb, e.g., *for cutting grass* (see [93] in section [3.3.1]).

The only exception to the above rule involves comparative uses of words (section [4.6]), in which case we treat them as subordinating conjunctions all the time (CSA), following the definition of CSA in SUSANNE. This is true even if the comparison is made between noun phrases. However, we deviate from our usual use of subordinating conjunctions by having the conjunction select for a COMP argument.

We can see a comparative phrase in [159] and its corresponding tree in figure [4.12]. In this case, both instances of *as* are CSA. The first is a CPZR of the following head, while the second one—deviating from our normal usage—selects for and finds a following COMP phrase.

(159) As far as I know, studying is about the only one thing that is important for students . . .

More examples of comparatives are given in section [4.6].

While we are on the topic of *as*, there are some fixed forms, such as *as follows*, illustrated in [160]. The key point is to determine what *follows* is. We treat it as a verb that doesn’t need a subject, on a par with phrases like *as is known*. Thus, *as* is a subcategorized-for CPZR of *follows*, and *follows* is a CPRED of *are* in this case.

(160) Details of each plan are *as* follows
4.6 Comparative constructions

There are different types of comparative constructions

4.6.1 as X as

As discussed in section 4.5 for as X as constructions, we deviate from our normal treatment of clauses. The rules are:

1. X is the head, making different categories of X distinct, e.g., *as quick as* can be a PRED while *as quickly as* can be a JCT.

2. X takes the first as as a CPZR.

3. X takes the second as as a COMP. Thus, the subcategorization of X will be <CPZR,COMP>.

4. The second as takes a COMP, even if it is a noun phrase that follows. It is this treatment of the second as which does not follow from our normal treatment of subordinating conjunctions. This will make Xer than constructions relatively straightforward to adapt to.

Two examples are shown in figures 4.13 and 4.14.

4.6.2 Xer than

The treatment of as X as really involves two parts, the definition of the comparative, e.g., *as important*, and then the linking word to compare to another entity, e.g., the second *as*. For Xer than constructions, such as *bigger than* in (161), we can apply something...
CHAPTER 4. A VARIETY OF DEPENDENCY CONSTRUCTIONS

Figure 4.13: Treatment of *as* in a comparative clause, functioning as an adverb

Figure 4.14: Treatment of *as* in a comparative clause, functioning as a predicate

like the COMP-COMP analysis from above, but instead using JCT for the dependent of the adjective. This is shown in figure 4.15, where *bigger* takes *than* as a CJCT, and *than* takes a COMP argument.

(161) ... we moved again to other house *bigger than* the other ... 

Figure 4.15: Treatment of *Xer than* comparative

We use CJCT, so that you do not need to distinguish reduced clauses from non-reduced ones. Furthermore, CJCT is the adjective corollary of the verbal COMP.

4.6.3 Discontinuous comparatives

We modify the annotation scheme described above for cases where the comparative appears at a distance, as illustrated in figure 4.16. Instead of the *than* phrase attaching to the adjective (*bigger*), we have it attach to the noun (*house*), thereby maintaining a projective structure (i.e., no branches cross). In keeping with the COMP and CJCT
labels used above, we use CMOD here for the relation between the noun and than—CMOD, COMP, and CJCT all refer to following clauses, but CMOD is appropriate for nouns and COMP for verbs, and CJCT for adjectives.

Figure 4.16: Treatment of a discontinuous Xer than comparative (constructed example)

There are other cases where prepositions appear to modify comparative or superlative adjectives, but which do not involve the same comparative-type structures. In (162), for example, the in the world phrase seems to be restricting the meaning of largest, not companies. Such a connection is difficult to maintain, so we have the preposition attach to the noun, in a regular NJCT relation, with no special marking, as in figure 4.17.

(162) As an assistant manager of one of the largest utility companies in the world, ...

Figure 4.17: Treatment of a superlative with a discontinuous prepositional modifier

4.7 Purpose Clauses (cf. in order to)

A purpose clause is a subordinate clause (see section 3.3.1) which defines the purpose of an action and can usually be phrased or rephrased using in order to. In (163), for example, the phrase to get knowledge is describing the (potential) purpose of spending money.

(163) ... in my country no one needs to spend money to get knowledge ...
There are two things to note about this example: first, as with other subordinate clauses, we have to determine the attachment point. Since the clause defines what the purpose of spend is, we attach to spend, as an XJCT (since this is an adjunct non-finite clause). For more on where to attach clauses, see section 4.1. Secondly, note that—unlike other subordinate clauses—there is a link to the subject (no one). This distinguishes purpose clauses, in that there is often a long-distance link with the subject.

There are variants of this basic purpose clause, involving in order (or so as) and for phrases which specify a subject. We treat both kinds of clauses as JCT relations to the to phrase, as shown in figures 4.19, 4.20, and 4.21. When a for phrase is used, note how the SUBJ relation changes significantly.

Note that sometimes purpose clauses do not actually have subjects that can be linked to, but which are instead contextually-derived. For example, in (164), the initial subordinate clause is defining a purpose, as can be seen by being able to rephrase it with in order to make this dream come true. However, it is less clear what the subject of this clause is, since the main clause takes a different direction. In such cases, you do not need to mark the SUBJ, instead treating the clause as any other XJCT.
4.8. Appositives

Appositives are used for (noun) phrases which have no grammatical purpose other than to clarify the usage of another noun phrase. As mentioned earlier (cf. (72)), APPOS is used to notate appositives. These noun phrases are generally adjacent to one another. For example, in (165), the unit *traffic rules or nabor rules* clarifies what *these* refers to, and thus the head of that string, the first *rules*, is an APPOS of *these*, as shown in figure 4.22.

(164) To make this dream come true, it’s necessary to get a MBA degree ...

(165) For example, traffic rules or nabor rules, these are very different from Japan.

Appositives do not necessarily have to agree in properties, given that learners may have unclear structures. In (166), for example, the word *everything* seems semantically incompatible with *something*. However, *everything* as an APPOS of *something* is the best analysis here (where *something* (cf. *anything*) is also marked as a LEX violation).
(166) When I came to Bloomington for living, I did not understand English, social rule, or how to do something, almost everything.

We also use APPOS to denote the relation between a noun and a reflexive, when the reflexive is merely intensifying the noun. In (167a), for instance, myself is an APPOS of I, as it intensifies the meaning. In (167b), on the other hand, myself is not intensifying and is a regular object (OBJ).

(167) a. I myself care less for badgers. (constructed example)

b. I told myself to care less. (constructed example)

Additionally, we often use APPOS for what seem to be reduced clauses that modify a noun, as in figure 4.23. In this instance, responsible is an adjective describing the subject; the fact that you can add a subordinating conjunction (e.g., Although) and not a determiner (e.g., *the usually responsible (,) she ...) tells us that this is a clausal type of entity.

(Although) usually responsible, she forgot to do her homework

Note that if the sentence were Although she was usually responsible, she forgot ..., then the analysis would not be an appositive one. Instead, it would follow the analysis for subordinate clauses, as in section 3.3.1 where the subordinate verb is dependent on the main verb.
4.9 Parentheticals

Note that an appositive analysis (section 4.8) is often the right one for a parenthetical. In (168), for instance, the phrase higher education is in apposition to university, so we annotate it as in figure 4.24.

(168) ...getting knowledge is free, from primary school to university (higher education)

 Figure 4.24: Parenthetical as an appositive

4.10 Ellipsis

The CHILDES manual (sec. 12.2) defines several elliptical relations (AUX-ROOT, DET-OBJ, INF-XMOD, etc.), but, for the sake of simplicity, we collapse these into a single ELL relation. We see two examples of ellipsis in (169).

(169) a. I am a graduated Biologist actually an Ecologist.

b. In conclusion, I’d like to say that hometown has influence on us, but just partly.

In (169a), we have an appositive, where Ecologist restates Biologist. The adjunct actually, however, indicates that there is potentially more structure here, as actually should modify a verb. To indicate that there is an elliptical structure (cf. actually [I am] an Ecologist), we attach actually as an ELL dependent of Ecologist, as shown in figure 4.25. Note that it is the word which is missing its head which takes the ELL label, in this case actually.

Note that ELL is used only when nothing else applies. In (169b), for instance, the conjunction but requires something after it; in fact, it requires something clausal since it is coordinating with has. But COORD is already underspecified with respect to the nature of the element it coordinates with. Thus, we mark this sentence with a typical COORD relation, selected for by but, as shown in figure 4.26.
CHAPTER 4. A VARIETY OF DEPENDENCY CONSTRUCTIONS

Note that if the ellipsis results in an item missing, and the sentence is grammatical in context, then the only change we make is in the subcategorization. In (170), for instance, if we assume a context where the sentence is grammatical, then *can* does not select for a verbal complement (VC) as it normally would.

(170) Yes, we can. (constructed example)

4.11 Linguistic mentions (quotations)

Oftentimes, people mention, sometimes fully quoting, words or phrases; i.e., they are not using them, but making reference to them and speaking meta-linguistically about them. In (171), for instance, the writer makes reference to the word *partly* (used previously in the discourse), but is not truly using it.

(171) I say partly because most of us visit a lot of other towns, countries, met a lot of people.
4.12. MULTI-WORD EXPRESSIONS

In such cases, we take a strategy which creates a discrepancy between the POS tags and the syntactic layer. Namely:

1. Morphological (and distributional) POS = the inherent POS of the word, in this case an adverb (RR).

2. Dependency syntax = the function the word (or phrase) has in the sentence, in this case as an object (OBJ).

The tree for this particular case is shown in figure 4.27, where we can see that an adverb is functioning as an object. This is one of the few times we actually advocate a mismatch between morphological POS and the morphosyntactic tree.

![Figure 4.27: Linguistic mention](image)

4.12 Multi-Word Expressions

Although multi-word expressions (MWEs) are in some sense fixed phrases, we try as much as possible to decompose them into component parts for our analysis—especially since parts of a MWE may be missing or altered by a learner. We present several analyses of MWEs here, without much discussion. Note that, when no other relation seems to fit, we use JCT as the relation.

We do not include subcategorization here, but follow the standard practice of subcategorizing for arguments, e.g., POBJ and IOBJ, while not for adjuncts, e.g., JCT—even though the distinction is much less clear with MWEs.

according to

(172) According to a survey, the average salary of MBA graduates is 10-20% more than ordinary people in marketing field and get more chances to be promoted to manager positions.
a lot  Note that *a lot* can be a verbal modifier (cf. section \[3.3.1\]), or can head or modify a noun phrase (e.g., *a lot of chocolate*). In either case, we treat *lot* as a noun with *a* as its DET.

(173) Before I study in Bloomington, I could not sleep a lot every Sunday because there loudly at outside.

all in all

(174) All in all, I was surprised for everything happened in Bloomington.

and so on

(175) I believe we will have great time in Bloomington by attending a lot of great events and parties, going for a walk, visiting beautiful parks, and so on.

at first

(176) So at first, I study economics.
4.12. MULTI-WORD EXPRESSIONS

**Figure 4.30: MWE: all in all**

**Figure 4.31: MWE: and so on (cf. and so forth)**

**because of** The analysis for *because of* is given in figure 4.33 for (177). Note that when it modifies nouns, *because* is an NJCT dependent of the noun.

(177) My life *because of* Bloomington was changed into lovely.

**even though**

(178) Even though our country is very poor, destructed by civil wars but it does help students in this field because, the people are poor too.

**for example** The phrase *for example* (or *for instance*) is treated as a transition (TRANS) of the thing that it is an example of. In figure 4.35, *John* is the example, and *so for* is a TRANS of this word. We treat *for* as the head, with a POBJ dependent.

**not only (not only / but also)** For the complicated conjunctions *not only* and *but also*, treat the left item (*not, but*) as the head, with the second item as a JCT dependent. After that, the rest of the structure is akin to a *both/and* or *either/or* structure (see *Conjunct phrases* in section 3.3.3).

(179) But now, I can spend a lot of time *not only* with my classmates *but also* with my wife in one of the most beautiful cities in the United States!
So at first, I study economics.

Figure 4.32: MWE: at first

My life because of Bloomington ...

Figure 4.33: MWE: because of

Note that, for phrases which are not only used in a non-conjunctive manner, as in the constructed (180), we still treat not as the head, even though it is more debatable whether this is the linguistically correct structure. The tree is shown in figure 4.37, where you can note that not now does not select for a COORD, but is rather a JCT dependent of as in this case.

(180) This is very important in my life, not only as a person.

such as The analysis for such as, with as the head, is in figure 4.38, taken from (181). Note that for phrases like such people as them, such is a QUANT of people and as a NJCT of people. (Note that as could be a part of a CMOD structure, as in such people as they are)

(181) ...to take advantage of environment friendly resources, such as nuclear power, hydro power and wind power.

truth be told We essentially treat truth be told as a valid non-finite clause, modifying the main clause of the sentence.

(182) Truth be told, I’m loving music and arts.
4.12. MULTI-WORD EXPRESSIONS

Even though our country is very poor, ... but it does help students ...

Figure 4.34: MWE: *even though* (cf. *even if*)

... a friend, *for example*, John

Figure 4.35: MWE: *for example* (constructed example)

(N) years old In this construction, *old* is treated as the adjective head, with *years* as a modifier (JCT) of this adjective.

(183) I’m a lawyer and I am 25 years old.
spend ... **not only** with my classmates but also with my wife ...

Figure 4.36: MWE: *not only ... but also*

... important in my life, **not only** as a person

Figure 4.37: MWE: *not only*

resources, **such as** nuclear power

Figure 4.38: MWE: *such as*

Truth be told...

Figure 4.39: MWE: *truth be told*
4.12. MULTI-WORD EXPRESSIONS

Figure 4.40: MWE: \((N)\) \textit{years old}
Chapter 5

Learner Innovations

This chapter deals with learner-specific language features that will likely be encountered, sometimes requiring their own guidelines for annotation.

5.1 Missing elements

You are only annotating what is there in sentence. So, the question arises about what to do when there appears to be a missing element somewhere in the sentence. In general, we try to annotate based on evidence which is present, so sometimes the missing element leads to an odd structure where nothing is missing, but in a few cases, mainly dealing with missing heads (e.g., INCROOT), there are specific practices we follow which treat something as missing.

5.1.1 Missing head

Often, it appears as if a head is missing. We handle these cases in two different ways, depending upon if we are missing a verb/root or something else.

INCROOT and its cousins

One common case of a missing head is when we see a sentence with no finite verb. For example, in (184), the main verb *playing* in not finite. Thus, the sentence is without a head, since *vroot* looks for a finite verb.

(184) Baby *playing* the dull with.

In this case, we can make the non-finite element the root of the sentence, but give it the label of INCROOT as we see in figure 5.1.1.
This policy is true even if the INCROOT is not verbal. In (185), for example, with no copula verb, the predicate adjective *friendly* becomes the INCROOT of the entire sentence. Although in general we only allow attachments between categories which are compatible, we make an exception for missing root elements; thus, in this case, everything which normally attaches to the head of the sentence now attaches to *friendly*, including *Also* (TRANS) and *people* (SUBJ).

(185) Also, the people in it very *friendly*.

For embedded clauses, we do not have the INCROOT label available, so we often resort to the unspecified _category. However, following the treatment of missing copulas in the main clause, we treat the embedded clause as essentially a well-formed clause otherwise. In figure 5.3, for instance, the string *that very smart and acting* is treated essentially like a normal relative clause (see section 4.4.4), despite missing a copula.

One important point for these cases regards subcategorization. While we allow words to take arguments they normally wouldn’t (e.g., *friendly* takes a SUBJ in figure 5.2), subcategorization is still a lexical property. If the (type of) word does not normally subcategorize for a category, do not add it to the subcategorization. In these
5.1. MISSING ELEMENTS

There is many people that very smart and acting.

Figure 5.3: Missing head of an embedded sentence with an adjective predicate

cases, friendly has a SUBJ it does not subcategorize for, and smart takes a CPZR it does not subcategorize for.

Non-ROOT cases

For non-ROOT cases, you should be inclined not to posit missing heads, unless absolutely necessary, in line with giving the learner the benefit of the doubt. This generally leads to better-formed trees and fewer mismatches as well. An example is in (186), where Amircan (cf. American) is ambiguous: it could be a proper noun subcategorizing for a determiner it doesn’t find (cf. the American) or an adjective with a missing head (cf. American Airlines, American locations, etc.).

(186) ...my friend from Amircan help me ...

In this case, we choose the analysis that does not assume a missing head, but, rather, a missing determiner, as missing a head is a more severe violation of the language. (A “missing” determiner is only an issue of subcategorization.) Note also that we use the morphosyntactic reading over the distributional one: considering from would favor a reading akin to using the word America—but treating the word as America would violate our phonetic distance heuristic for determining the lemma, as discussed in section 2.2.

The annotation for this example may seem non-intuitive to you, but this just serves to illustrate that the annotation principles may sometimes lead you to a slightly non-intuitive solution. As much as possible, you should stick to the principles.

To underscore the idea that we want to shy away from missing heads, consider again example (187), which we discussed earlier. The phrase working as Product Movement and Inventory may at first be seen as missing a head, such as manager or any other job title. It could also be viewed as a “wrong” preposition (or perhaps semantically-odd noun phrase) being used, where another preposition would be a better fit (cf. working in Product Movement). We favor an analysis that does not assume a missing element.
Graduating from commercial and accountancy Faculty in Thailand and having an experience **working as Product Movement and Inventory** at ExxonMobil, I realized that I wanted to concentrate on “Supply chain Management.”

Since it is a lexical choice that is the problem here, we indicate this by noting a lexical violation on the preposition *as*. It is quite possible that the noun chain is the semantic oddity, but remember that when situations are unclear for entire phrases, you should mark the head of the entire unit as the lexical violation (see section 2.4), so *as* is the preferred choice here. Notice here that, aside from a lexical violation, the dependency analysis largely ignores the semantics of the prepositional phrase, focusing more on combining syntactic categories in appropriate ways.

There are some cases where it is unavoidable to posit a missing head. Consider (188), where a purpose clause (section 4.7) lacks an infinitive marker.

Also, I need more natural and friendly place to live with my wife in order **understand** each values and natures by enjoying great opportunities inside and outside of IU.

![Figure 5.4: Purpose clause missing to](image)

The core problem here is the lack of the infinitive marker *to*, so that is the only part of the analysis which changes. Consider what now happens: in some sense, the relation between *need* and *understand* is a combination of *XJCT* (as would be between *need* and *to*, were it present) and *VC* (between *to* and *understand*). Given such a combination, we annotate it as the underspecified _ (see section 1.2.5).

### 5.1.2 Missing argument

In some cases, learner sentences have missing arguments. For example, in (189), *destroying* needs an object, but there isn’t one in the sentence.

The toy was worried about **destroying**.
5.2. EXTRA ELEMENTS

In this case, ‘destroying’ has OBJ in its subcategorization list, but this dependency is not realized in the dependency tree. In general, this type of phenomenon will be caught by the mismatch between the subcategorization list of any head, which includes all obligatory arguments, and the actual, realized dependents for this head, where one or more may be missing.

5.2 Extra elements

Extra elements are those not selected for, but are still there in the sentence.

5.2.1 Extra head

As we mentioned before, positing a missing head in an analysis is the last resort, and, similarly, so is positing an extra head. Sometimes, though, this is necessary. We often see this when more than one verb has the form indicative of the head of the sentence. This can be seen in (190), where both am and agree are verbs with (arguably) present tense morphology, either of which could be agreeing with the subject I.

(190) Yes, I am agree with this statement.

Thus, the tree contains two ROOTs, as shown in figure 5.5. Note here how the TRANS word Yes is attached to the first ROOT, while the final punctuation is a PUNCT dependent of the second ROOT. In both cases, we are following the principle of attaching to the closer element.

\[ \text{Figure 5.5: Sentence with two ROOTs} \]

It can be more complicated when one of the verbs does not fully agree with the subject. In (191), for example, both verbs is and have can bear tense, making them ROOT candidates. A possible solution would be to attach both finite verbs to vroot and give them a ROOT label. However, is agrees with the subject, while have does not.
Thus, we assign the ROOT label to *is*, while for *have*, it takes an INCROOT label, as it is not a fully proper head in this context of non-agreement.

(191) Bloomington *is have* nice and flexible weather.

![Figure 5.6: Sentence with two ROOT-like labels](image)

The main idea, though, is that the form of *have* better fits with a INCROOT analysis than with trying to make it a VC of *is*, when its form is clearly not that of a verbal complement of *is*.

Another example that might be analyzed as having an extraneous element is in . . . *when I was ten years, I . . .*. It is possible to analyze *years* is extraneous, since *when I was ten* is perfectly grammatical on its own, just as it is possible to annotate a missing adjective head, as *when I was ten years old* is also grammatical. We definitely want to avoid the missing head analysis, but we also prefer to avoid positing extraneous elements, and so we assign *years* the dependency label of PRED, making it the dependent of *was*.

**Run-on sentences** For run-on sentences, and more generally missing conjunctions, see the section on *Missing conjunction* in section 3.3.3. Instead of using multiple ROOTs, we use a “broken” COORD structure, e.g., *tought and were* in (192) would be joined by a non-subcategorized COORD.

(192) The *tought* me usual things, these *were* not grammar.

### 5.2.2 Extra dependent

For arguments, extra dependents are also caught by examining mismatches between subcategorization frames and realized dependents, similar to missing arguments. Extraneous adjuncts, however, must be handled differently.
5.2. EXTRA ELEMENTS

Extraneous arguments

Some clear cases occur when the learner uses multiple subjects or determiners; an example is in figure 5.7, where we can see the noun *tasks* being preceded by two determiners. Both are valid determiners of the noun, so the deprel is DET between each determiner and the noun, as shown in figure 5.7. However, the subcat frame of *tasks* selects for only one of these, and this mismatch in number between the subcategorization list and the realized determiners points at the non-native aspect here.

![Figure 5.7: A noun with two determiners](image)

Such determinations can be more challenging in other cases, but remember that we are examining morphosyntactic evidence more than semantic. In (193), for example, the verb *made* finds a direct object (OBJ) in *series*. Crucially, *steps* is essentially in the same slot, looking like a nominal direct object. Regardless of what the “intended meaning” is, we mark this as having two OBJs. The subcategorization frame will require only one OBJ for *made* to have saturated its required argument, again pointing to non-nativeness.

(193) I made my first **series steps** .. in Almaty.

Extraneous adjuncts

As a case of an extraneous adjunct, consider (194), where it seems as if either *only* or *one* may be extraneous. Either one is a valid modifier of *thing*, but the occurrence of both seems redundant; thus, we can mark both as MOD dependents of *thing*.

(194) studying is about the **only one** thing that is important for students

Consider what is happening here: there is no semantic anomaly (the sentence makes complete sense), but rather the presence of an extraneous adjunct is leading to grammatical ill-formedness. This fits our definition of lexical violations (section 2.4). Thus, one of these two words, *only* or *one*, should be marked as a lexical violation.

In this case, the presence of *about* indicates that *one* is the extraneous word and is thus marked as a violation. If there were not this evidence and either word could be the extraneous one, we would mark the head *thing*, in keeping with the heuristic for
lexical violations of marking the head of strings when the specific anomaly cannot be
determined (section 2.4).

We can see extraneous adjuncts with punctuation, too. In (195), for example, the
comma is unnecessary between because and people, we also mark this as a lexical
violation, but still attach it as a PUNCT relation to the previous word (since the scope
is not entirely clear). Note that, in general, we try not to mark any problems with
punctuation, as that is a more stylistic issue, but in clear cases, such as this one, lexical
violations are appropriate.

(195) Even though our country is very poor, distructed by civil wars but it does help
students in this field because, the people are poor too.

Adjunct cases which are missing—such as a missing comma (e.g., after wars in
(195)—are simply not able to be captured in our annotation scheme. Outside of punc-
tuation, these should be extremely rare, however, as adjuncts by definition are generally
optional.

Note still that lexical violations are last resorts. If you can mark an underspecified
POS tag or dependency relation, mark that underspecification and do not mark a lexical
violation. In (195), for instance, consider the fact that we have both a subordinating
conjunction (Even though) and a coordinating conjunction (but), though only one is
required. Slightly different from the cases above, picking one word or the other will
completely change the analysis—as either a subordinate clause modifying a main clause
or two main clauses coordinating.

We use two heuristics here: treating the left (first) word as more important and
treating the less semantically-bleached word (Even though) as more likely to be used
in the analysis. In a sense, this means that it is less likely for Even though to be an
“accident.” In treating but as extraneous, we give it an underspecified relation to the
main verb (does). See section 6.5 for a full analysis.

Extra as Missing

There are cases where you must decide whether you have a missing element or an extra
element. In principle, there is no clear favoritism between one type of analysis or the
other, but you must use the general principles you know.

In (196), for example, it seems like the writer could have simply said then, I study
law, thus making degree (lemma=degree) an extraneous word. If this were the case, then
the sentence would be annotated as in figure 5.8, where law is a valid object (OBJ), and
then degree is either a separate OBJ or given some non-specific label.

(196) then, I study law degree.

Note, however, that as an object, degree needs to have a determiner (DET). In essence,
then, we would be saying that not only is there an extra word, but that extra word has a
missing word. By contrast, then, we can treat law degree as a phrase, and we then only have the missing word issue, as shown in figure 5.9. This is a much better analysis, as only one mismatch is present (an unrealized DET), and it uses the evidence as present in the words themselves. That is: we don’t treat degree as some unnecessary appendage, but see if it can be fitted in.

Importantly in this case, the meaning of studying (a) law degree is not entirely clear, but that is fine. Syntactically, it is a valid sequence, modulo the unrealized determiner, and that is sufficient.

5.2.3 Extra word with unclear function

While extra heads and arguments can more or less be fit into a total dependency tree, there are cases of extra words whose function is entirely unclear. In (197), for example, the word at seems to have no function in English. This is why the distributional POS tag is left unspecified, as discussed in section 1.2.5.

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

For dependency annotation, when we have unclear cases such as this preposition, we follow a few guidelines:

1. If the POS can attach to a neighboring POS, do so, most likely giving it an unspecified label. In the case of at in (197), a preposition can modify a verb phrase,
so attaching it to to is appropriate. It would also be appropriate to attach to run, but that introduces non-projectivity, so it is disfavored.

2. If there is some other evidence which provides a clue—phonological, how the word generally attaches, other instances of a phrase, etc.—then attach as that evidence leads. In (198a), for example, it is not clear how to attach A, but looking at more instances in the same essay, such as (198b), we see that a and partly tend to co-occur, meaning it should attach to partly, though with an unspecified function.

(198)   a. A partly it is because of our hometown .
        b. But I think hometown influenced on me a partly .

3. If the case is totally unclear, treat it similarly to punctuation (see section 3.3.5) and attach to the nearest head over which it could have scope, favoring larger chunks.

5.3 Word order

As we mentioned in other sections, we do not handle word order errors. In (199), for instance, Too and now are marked as adjuncts (JCT) of their verbs, despite being in the wrong location. Nothing special is notated in these cases.

(199)   a. Too I want to study French.
        b. He can’t see now nothing.
Chapter 6

Extended Examples of Difficult Cases

In this section, we’re simply going to provide some more examples, to discuss how we arrive at a final annotation.

6.1 Example 1: one, complement clause

Consider (200), where we find an odd use of one after the verb dreamed. The thrust of the string seems to be rooted in the finite clause (I can go to a typical American city), which seems to indicate that this clause, at least, is a complement (COMP) of dreamed, as shown in figure 6.1.

(200) When I was in my country, I dreamed one I can go to a typical American city.

![Figure 6.1: Example of a confusing use of one]

What then about the word one? While you might think of it as some sort of marker within the finite clause, note that it forms a valid object (OBJ) on its own, especially as dreamed can take objects. Thus, we annotate one as an OBJ, in addition to having a COMP. This may go against your intuition, but it is consistent with the practice of
extra arguments (section 5.2.2) and missing conjunctions (section 3.3.3). Remember to follow the annotation principles above any intuitions about intended meaning.

Crucially, the subcategorization only selects for one of the two realized post-verbal categories. We have it select for COMP (and not OBJ) because, as mentioned above, the thrust of what comes after the verb seems to be rooted more strongly in the finite clause than in the object, making it a better fit to the context (see section 3.4).

6.2 Example 2: lemma

For a second extended example, consider (201), where the word heart is problematic, since it is a noun occurring in a verbal context. While we could consider annotating the lemma as hear or heard, to match a possible intended meaning, this is too phonetically distant, given the prominence of a t ending forming a new word (see section 2.2). With access, on the other hand, there is no confusable word, and r’s seem to be easily dropped.

(201) I can heart the sound of stream access the stone.

\[ \begin{array}{c}
\text{ROOT} \\
\text{VROOT} \\
\text{SUBJ} \\
\text{DET} \\
\text{NJCT} \\
\text{POBJ} \\
\text{PUNCT} \\
\end{array} \]

\[ \begin{array}{cccccccc}
\text{I} & \text{can} & \text{heart} & \text{the} & \text{sound} & \text{of} & \text{stream} & \text{access} & \text{the} & \text{stone} \\
\text{PPIS1} & \text{VM} & \text{NN1} & \text{AT} & \text{NN1} & \text{IO} & \text{NN1} & \text{II} & \text{AT} & \text{NN1} \\
\text{PPIS1} & \text{VM} & \text{VV0} & \text{AT} & \text{NN1} & \text{IO} & \text{NN1} & \text{II} & \text{AT} & \text{NN1} \\
\end{array} \]

Figure 6.2: Example of a problematic lemma

With heart as the lemma, the POS tags can be deduced (NN1 morphologically, VV0 distributionally), but what then does the morphological tree look like? Since there is a conflict between the modal verb can and its dependent heart, we give it no label, in keeping with the principle of underspecification in the face of conflicts (see section 3.1.2). One could consider treating can as the main verb here (with a meaning of canning items), with two objects (heart, sound), but this reading is completely absent from the context.

\[ ^1 \text{One could consider annotating heart as morphologically a verb, on the basis of constructions like “I heart NY,” but we deem this too informal (see section 3.4.1).} \]
As a more minor issue, consider the attachment of *access*. It is not immediately clear whether it attaches to *sound* or *stream*. To disambiguate, we follow the tests in section 4.1 and see whether various phrases can be moved, as shown in (202). The crucial test is in (202b), testing whether the *access* phrase can move with *stream*, but not *sound*, and maintain the same meaning as the original. In this case, it seems to be acceptable, meaning we attach *access* to *stream*.

(202)  a. The sound of [the] stream across the stone, I can hear(t)
       b. [the] stream across the stone, I can hear(t) the sound of

6.3 Example 3: syntax vs. meaning

As an example illustrating where meaning can lead you astray, consider (203). The main clause could be interpreted along the lines of *I will join many social events (in order) to look for such a lady*. However, what we actually have, based on the words, is a coordinate structure, and thus it is annotated as in figure 6.3.

(203) Because beauty and attractive characters are very important for me, I will join many social events and look for such a lady.

![Figure 6.3: Example of prioritizing syntactic structure over meaning](image)

6.4 Example 4: lemma, word order, run-on

Sentence (204) has a number of features to note, starting with *plac* being orthographically similar enough to *place* to treat it as such. Additionally, even though *Lost* is a real word, different from *Last*, it is orthographically similar, as *a* and *o* are easily confused (sometimes even by transcribers). Furthermore, the discourse context to this point indicates that a final point is being made. Thus, we treat the lemma here as *Last*, allowing it to be a TRANS dependent of *has*, as shown in figure 6.4.
(204) **Lost**, my country has too many people, **there already** every place has person.

Looking at the sentence as a whole, we have a run-on, which we treat as a single sentence, since the learner wrote it as such and especially since the comma indicates a connection between them (see section 2.1.1). Following the treatment outlined in the *Run-on sentences* part of section 5.2.1, we treat the second verb (*has*) as a non-selected COORD dependent of the first.

We have two adverbs, *there* and *already*, neither of which sounds ideal in their current locations. Note, though, that the issue is primarily of word order (see section 5.3): if the sentence were *every place already has (a) person(s) there*, then it would be perfectly acceptable, as far as the adverbs are concerned. Our annotation, especially for adverbs, is not overly concerned with word order (remember that these are dependencies rooted in morphology, as discussed in section 3.1), so they are not marked in any special way, simply being JCT dependents.

The treatment of *person* has no special marking in the dependency tree, being a valid object (OBJ). Its distributional POS tag, however, is NN, representing the fact that it is not clearly singular (NN1) or plural (NN2) (see section 2.3.2), and it has SUB-CAT1=DET, as this singular common noun requires a determiner.

Finally, as a minor point, note that *too* is a JCT modifier of *many*, this being the way we handle modifiers of quantifiers.
6.5 Example 5: missing verb, double conjunction, unclear phrase

In (205), we have a missing verb in one clause (see section 5.1.1) and an unclear phrase (*something own itself*) in another. In general, the missing copula is handled as outlined in section 5.1.1: the predicate *big* takes the subject (SUBJ) *it*, but does not subcategorize for it, as shown in figure 6.5. (With the coordination, we could also have *famous* take a SUBJ—paralleling the analysis of coordinated verbs, shown in section 3.2.8—but this is not necessary.)

(205) Although it not big and famous, but it still has something own itself.

Figure 6.5: Unclear string

We do have *big* subcategorize for the CPZR, due to the fact that phrases without verbs, like *Although not big and famous*, are grammatical. With no verb, but with a subject, we have a non-finite clause modifying the main clause, and thus XJCT is used as the dependency relation between *has* and *big* (see both section 4.8 and 3.3.1 to distinguish appositive clauses from subordinate ones).

The next major issue is the double conjunction issue: both *Although* and *but* are used in the same sentence, when really only one is needed. Following a left-to-right heuristic—and the fact that coordinating conjunctions seem to be more semantically bleached—we treat *but* as extraneous, giving it an underspecified relation to the main verb (see section 5.2.2).

The most challenging question for this sentence is what to do with *something own itself*. There are at least three different possibilities, all of which are more or less consistent with the syntactic evidence:

1. Treat *own* as a misspelling of *on*, making *o(w)n itself* an NJCT of *something*. Furthermore, *o(w)n* is also a lexical violation, as this is not a proper preposition for this context (cf. *by, with*). Since this in essence posits two errors, we (heuristically) disfavor it.
2. Treat *own* as a verb. But this is even more problematic, as we have a dramatically reduced relative clause, and the meaning makes no sense in this context (*something which owns itself?*)

3. Treat *itself* as an intensifier of *something* (using APPOS, as described in section 4.8). The word *own* is then in its adjectival (JJ) use and modifies *itself* (cf. phrases like *my own self*). Since the use of *own* is not clear, the relation is underspecified.

Given that this last analysis essentially posits only one problem and forms a more or less sensible meaning, working with the words as they appear, we use it.

Interestingly, in this case, this learner did something similar earlier in the same essay, which could be given a different analysis. The similar sentence is in (206). The major difference here is that there is a perfectly sensible reading with *own* as a verb.

(206) You can’t find a place *own yourself*.

Both the verbal analysis and the intensifier analysis posit only one error (missing *to* vs. extraneous *own*), so they seem equally valid. Given the repetition of the same type of phrase in the same essay where the verbal reading is ruled out (and the fact that *You can’t find a place to own yourself* is perhaps slightly odd), we use the intensifier analysis here, too, i.e., *own* as *of* *yourself*. Since the intensifier *yourself* modifies a verb (*find*), it receives a JCT label instead of APPOS in this case.

### 6.6 Example 6: syntax vs. discourse, multi-ambiguity

For an example of a sentence whose syntactic analysis does not fit the discourse, consider (207). This comes in a context, after mentioning people smiling at the person, where it seems fairly clear that *all these [things] were things that made the person feel better.* However, there is no word *made* here, so as much as you might be tempted to analyze the sentence based on it, it is not the procedure we use.

(207) All *these me* felt better.

While the string *All these me* is challenging, note that you can start with the easy properties that *felt* is the ROOT of the sentence, with *better* as its predicate (PRED), as shown in figure 6.6. Furthermore, we know that *All* is clearly a QUANT, of either *these* or *me*. These leaves the analysis of *All these me*.

The problem is that *me* appears out of place (should be post-verbal), in the wrong case (should be nominative), and/or with the wrong modifiers (should take no determiners or quantifiers). To distinguish different kinds of analyses, we rely on a few points:
6.7. EXAMPLE 7: SYNTAX VS. MEANING

1. We are trying to annotate grammatical sentences which as much as possible fit into the discourse. Treating *All these me* as a string clearly does not fit the discourse. Annotating something which has a meaning akin to *All these [things] felt better to me* seems a shade better, though still not a good fit.

2. If *me* heads a noun phrase (i.e., heads *these* and/or *All*), there are two possible grammatical problems: *me* might take a quantifier, in addition to being in the wrong case for a subject.

3. Another possible analysis treats *me* as the SUBJ, with *All these* as some sort of TOPIC marker (cf. *For all these reasons, I felt better*), but this topic is not well-formed—in addition to *me* still being the wrong case.

4. If *me* is some type of dative-like entity, then *All these* is a proper subject, and *me* is simply in the wrong case. It is in the wrong position, but for our morphosyntactic annotation, we are less concerned about word order. This analysis is close to what we wind up, the only difference being that we cannot fully commit to an analysis with *me* as a proper (dative) argument of *felt*.

Thus, we choose to annotate *these* as the SUBJ (with *All* as a QUANT), while *me* is some unspecified dependent of *felt*. This seems to fit the evidence at hand without overpositing errors.

6.7 Example 7: syntax vs. meaning

Example (208) is a good case of where one has to shut meaning off while annotating. The issue is that the sentence intuitively probably means something like *I was surprised by everything happening in Bloomington*. Yet this meaning is not consistent with what we observe in the syntax.

Example (208) *All in all, I was surprised for everything happened in Bloomington*
What we actually observe syntactically is a subordinate clause for everything happened in Bloomington modifying the main clause. This gives a meaning akin to a because clause, partly because surprised can be intransitive. Thus, we can mark this as a typical subordinate clause, as in figure 6.7.

One debatable issue is whether for should be marked as a lexical violation. In this context, and especially without an intervening comma, it does seem odd, even with a because meaning. Thus, it seems like marking a LEX violation is appropriate, as indicated by the asterisk.

As a minor point, note that surprised is a PRED, because it is an adjective and not a verb—e.g., very surprised is grammatical here (see VC vs. (X)PRED under section 3.2.5).

### 6.8 Example 8: complement clause

We find a strange case of a complement clause in (209). There are two issues with somebody smiled to me. The first, which is easy to diagnose, is that to is an odd-sounding preposition. Thus, in figure 6.8, we show it as a lexical violation (marked with an asterisk).

(209) When I walked on the road, I always found somebody smiled to me.

The second issue is more complicated, namely: what is the relationship between found and somebody smiled to me? To figure this out, we should consider what found selects for vs. what it finds. In (210), we have at least a few possible complements: a noun phrase with a relative clause (NP$_{rel}$), a noun phrase with a gerund (NP$_{ger}$), and a sentence started by that (S$_{that}$). We can even find a non-that sentence (S), depending upon the exact intonation and pragmatics.

(210) a. NP$_{rel}$: I found somebody who smiled at me.
    b. NP$_{ger}$: I found somebody smiling at me.
c. $S_{that}$: I found that somebody smiled at me.

d. $S$: ?I found somebody would smile at me.

Indeed, this shows us that a finite sentence can follow $found$ in this context, though a bit oddly without the $that$. To indicate the oddness (which seems ungrammatical), we mark a CPZR being subcategorized for by $smiled$, but not being realized.

![Figure 6.8: An odd-sounding but quasi-okay sentence](image)

### 6.9 Example 9: Lemma, POS ambiguity

In sentence (211), we see a few problems with the forms of words, namely $study$ and $lean$. We will consider each in turn, with the tree in figure 6.9.

(211)  

First, for $study$, we have a noun/verb ambiguity, and the fact that it occurs after $like$ makes the decision challenging. If the sentence were simply $I like study$, we would have a very difficult decision in determining its noun vs. verb status (which we won’t go into here). Note, though, that coordination is another piece of evidence. In this sentence, because $study$ coordinates with $hang out$ and $lean$, we can see that it is coordinating with verbs and thus is a verb itself.

The question, then, is how to handle the relation between the verb $like$—which usually takes an infinitive clause, as noted in the subcategorization for VC—and the verb $study$. By convention, (see section 5.1.1, especially surrounding figure 5.4), we handle this as an underspecified relation.

Turning now to $lean$, it seems clear from context that $learn$ would be a better fit semantically. However, this is not a simple typographical mistake and is not similar enough phonologically to warrant treating the lemma as $learn$. Thus, we lemma-annotate as $lean$; falling out from this lemma, we have a subcategorization which selects for an indirect object (cf. $we lean on your culture$). And then, regardless of the lemma annotation, there is an object (OBJ) relation between $lean$ and $culture$. 
CHAPTER 6. EXTENDED EXAMPLES OF DIFFICULT CASES

One annotation is not marked: , even though lean is semantically anomalous, it is not a lexical violation. Lexical violations are only for wrong words which lead to ungrammaticality, not semantic oddness.

In passing, note how culture selects for a DET which is not present.

6.10 Example 10: non-finite subordinate clause, problematic misspellings

There are a number of complications in (212). First, as discussed in section 2.2.2, prison, the lemmas of no and prison are unaltered forms, i.e., no and prison, respectively. For prison, this doesn’t affect the tree at all, but for no, this means that we do not have a true label between no and what it modifies, as shown in figure 6.10.

(212) After to start, I want to tell that this exercise is very important in the life, no only as a prison.

The multi-word expression not only is discussed in section 4.12 where not is the head and only is a JCT dependent. We stick to the structure of not only as much as we can, with the modifier only treated as a regular JCT dependent of no. However, as mentioned above, the relation of the phrase as a whole is marked as unspecified, i.e., an unspecified relation between as and no.

Secondly, notice the placement of the comma between life and no. As discussed in section 3.3.5 this attaches to important, the same word that the no only as a prison clause attaches to. This phrase attachment shows us that the comma scopes over important and thus does not introduce any non-projectivity.

The final issue to deal with is unusual non-finite subordinate clause, After to start. While there is an issue with finiteness, note that we can construct a normal subordinate clause, since to start is a verbal form, meaning that After can and should be treated as a complementizer (CPZR) of to, which is itself the head of the clause (see below for subcategorization). See section 4.5 for more on analyzing the arguments of words like after which vary between prepositions and subordinating conjunctions.

What, then, is the issue with After to start? Firstly, non-finite clauses do not typically select for a complementizer, and thus to does not include CPZR on its subcategorization list. Secondly, this position after a complementizer is generally a position for a finite verb (though, in that case a subject would need to be present) or an -ing non-finite clause, not an infinitive. This is an issue of distribution and is marked on the distributional POS layer, as shown in (213). We do not know which verb form is appropriate here (VV), and the marker to really serves no purpose.
(213) After to start ...
    ICS TO VV0 ... ⇐ Morph.
    ICS _ VV ... ⇐ Dist.
Figure 6.9: Coordination as evidence, and a difficult lemma
After to start, I want to tell that this exercise is very important in the life, no only as a prison.
Appendix A

Practical matters

A.1 Brat annotation tool

We will be using the Brat rapid annotation tool (http://brat.nlplab.org/) to do annotation. An example annotated sentence is shown in figure A.1.

From the bottom to the top, we have the following layers:

1. Words, including ROOT. You will be starting with words segmented into sentences. If the sentence segmentation looks incorrect, notify one of us immediately, so we can fix the underlying files before annotation is added.

2. Dependency POS tag

3. Morphological POS tag ... Note that the following annotations are captured within the morphological (blue-bordered) tag, so please keep them in the morphological layer:

   (a) lemma information

   (b) subcategorization information
APPENDIX A. PRACTICAL MATTERS

(c) the presence of a lexical violation

Annotating all these is discussed below, but you can see little asterisks indicating which words contain subcategorization or lexical violation issues. The blue border around the morphological tags indicates that lemma information is present.

4. Morphosyntactic dependencies, shown in arcs.

A.1.1 Getting started

We will send you a link to the annotation files. When you go to that site, you will see an information page and then have the opportunity to select an annotation file. After doing this, in the upper right corner, select the option to login.

A.1.2 Basics of annotation

The Brat manual is at: [http://brat.nlplab.org/manual.html](http://brat.nlplab.org/manual.html) and this may help you at times to annotate. Note that to create dependency arcs, you will drag the morphological tag of the head to the morphological tag of the dependent. Upon doing this, you will then select a label. This process is illustrated below in section [A.1.4](#A.1.4).

Note that the POS annotations are referred to as Entities in Brat, while dependencies are Relations. (You do not need to worry about Events.) Subcategorization is handled via Attributes, described below.

A.1.3 Annotating a word

For the morphological tag, there are a number of annotations to include, illustrated in figure [A.2](#A.2).

**Lemmata** Lemma information is included in the Notes field, as this is the only place where Brat allows open-ended annotations. Be sure to double-check lemma annotation, in cases where the word is inflected and, more importantly, for reasonable misspellings.

**Subcategorization** Subcategorization is handled via 4 separate annotation slots. In the case of enjoyed in the figure, for instance, you can see that the first slot is SUBJ, the second OBJ, and the third and fourth are left empty. We think we will never need more than 4 subcategorization slots, but if you find this necessary, let us know.

**Lexical violations** Lexical violations are simply annotated by checking the LEX box. Otherwise, leave it unchecked.
POS hierarchies  For both the morphological and distributional tags, you will have to wade through a hierarchical tagset. The good parts are that: a) you start with the lead category, and b) you can easily underspecify the tag as much or as little as necessary. An example of annotating *enjoy* is shown in figure A.3.

Re-segmenting words  When re-segmenting words, you will have to remove the POS annotations which are present and then highlight the appropriate token(s) for a new annotation. Note that you will be adding two POS labels and that this can be tricky in Brat. The problem is that if you create two POS labels with exactly the same values, Brat seems to think that you only want to create a single label.

One way to work around this, after creating one label, is to create a second annotation (i.e., re-highlighting the same token(s)) with an initially different label. Then, you
A.1.4 Annotating dependencies

The process for creating dependency relations is fairly straightforward. Taking our example sentence, the starting point for annotating dependencies will be as in figure A.4. In this case, the asterisks indicate that In, enjoyed and city all have subcategorization already marked. No arrows are marked yet.

Figure A.5 shows the situation right after really has been annotated as the JCT of enjoyed. The next step may then be to annotate city as the OBJ of enjoyed. This is done by dragging an arrow from the morphological POS box for enjoyed (the head) to the
A.2. CONLL FILE FORMAT

When you have dragged the arrow, an annotation box appears, as in figure A.6. Here, you mark the dependency relation, OBJ in this case. The resulting dependency tree is then shown in figure A.7.

A.2 CoNLL file format

You will not need to directly work with CoNLL files, but we often want to use this format for working with the annotated files. Thus, we have scripts available to convert the Brat annotation files into conll.

The annotation is encoded in CoNLL format (Buchholz and Marsi, 2006), as seen in figure A.8. Columns 1–8 indicate the following information, in order: ID (1), token (2), lemma (3), morphosyntactic POS tag (4), distributional POS tag (5), subcategorization frame (6), head ID (7), and morphosyntactic dependency relation (8). Column 9 is left blank (fitting the CoNLL standards), and column 10 encodes secondary dependencies as a semi-colon separated list of head:relation pairs (though, none are shown here).
Figure A.6: Selecting the object relation

Figure A.7: Sentence after annotating the object
1 In in II II <POBJ> 6 TRANS _ _
2 conclusion conclusion NN1 NN1 _ 1 OBJ _ _
3 , , Y Y _ 1 PUNCT _ _
4 I I PPIS1 PPIS1 _ 6 SUBJ _ _
5 really really RR RR _ 6 JCT _ _
6 enjoyed enjoy VVD VVD <SUBJ,OBJ> 0 ROOT _ _
7 this this DD1 DD1 _ 8 DET _ _
8 city city NN1 NN1 <DET> 6 OBJ _ _
9 . . Y Y _ 6 PUNCT _ _

Figure A.8: CoNLL format
Bibliography


Index

adjunct, 17, 20, 25, 29, 33, 34, 40, 41, 48, 69
argument, 17–21, 24, 29, 40–42, 48, 66, 67, 69

CHILDES, 19, 20, 22–24, 33, 37
COORD, 24–29, 46, 62
NJCT, 17, 19, 20, 33, 36, 37

lemma, 5–9, 14, 15, 30, 40, 65, 70–72

subcategorization, 6, 12, 13, 15, 16, 21, 24, 25, 27, 30, 32, 37, 40, 42–46, 48, 61, 66, 67, 70–72
SUSANNE, 10, 11, 38, 40

underspecification, 5, 61