CONTEXT-FREE

The simple model of phrase structure that we have developed here adopts a context-free view of language. For example, once we have expanded 'VP' to 'VBD NP' and then to 'sewed NP,' we can replace NP with whatever noun phrase we please. The context provided by the verb sewed is inaccessible when we decide how to expand NP. This inaccessibility of context is the key property of a context-free grammar. We could expand VP to a natural phrase like sewed clothes, but we can as easily choose a nonsensical expansion like sewed ***wood blocks.***

▼ How to include necessary dependencies is a central topic in probabilistic parsing, which we discuss in chapter 12.

3.2.2 Dependency: Arguments and adjuncts

DEPENDENCY

Another important organizing notion is the concept of ***dependents.*** In a sentence like:

(3.48) Sue watched the man at the next table.

Sue and ***the man*** are dependents of a watching event. We will say that they are the two arguments of the verb ***watch.*** The PP ***at the next table*** is a dependent of man. It modifies ***man.***

SEMANTIC ROLES
AGENT
PATIENT

Most commonly, noun phrases are arguments of verbs. The arguments of verbs can be described at various levels. One can classify the arguments via ***semantic roles.*** The ***agent*** of an action is the person or thing that is doing something, the ***patient*** is the person or thing that is having something done to it, and other roles like ***instrument*** and goal describe yet other classes of semantic relationships. Alternatively, one can describe the syntactic possibilities for arguments in terms of grammatical relations. All English verbs take a ***subject,*** which is the noun phrase that appears before the verb. Many verbs take an ***object*** noun phrase, which normally appears immediately after the verb. Pronouns are in the subject case when they are subjects of a verb, and in the object case when they are objects of a verb. In our earlier example, here repeated as sentence (3.491, ***children*** is the subject of ***eat*** (the children are the agents of the action of eating), and ***sweet candy*** is the object of eat (the sweet candy is the thing being acted upon, the patient of the action):

SUBJECT
OBJECT

(3.49) Children eat sweet candy.

Note that the morphological form of candy does not change. In English, pronouns are the only nouns that change their forms when used in the object case.

Some verbs take two object noun phrases after the verb, both in the object case:

(3.50) She gave him the book.

INDIRECT OBJECT
RECIPIENT
DIRECT OBJECT

In this sentence, ***him*** is the ***indirect object*** (describing the ***recipient,*** the one who indirectly gets something) and the book is the direct object (describing the patient). Other such verbs are verbs of sending and verbs of communication:

(3.51) a. She sent her mother the book.

b. She *emailed* him the letter.

Such verbs often allow an alternate expression of their arguments where the recipient appears in a prepositional phrase:

(3.52) She sent the book to her mother.

Languages with case markings normally distinguish these NPs and express patients in the accusative case and recipients in the dative case.

There are systematic associations between semantic roles and grammatical functions, for example agents are usually subjects, but there are also some dissociations. In ***Bill received a package from the mailman,*** it is the mailman who appears to be the agent. The relationships between semantic roles and grammatical functions are also changed by voice alternations (the one feature in table 3.3 which we did not discuss earlier).

ACTIVE VOICE
PASSIVE VOICE

Many language make a distinction between active voice and passive voice (or simply ***active*** and ***passive).*** Active corresponds to the default way of expressing the arguments of a verb: the agent is expressed as the subject, the patient as the object:

(3.53) Children eat sweet candy.

In the passive, the patient becomes the subject, and the agent is demoted to an oblique role. In English this means that the order of the two arguments is reversed, and the agent is expressed by means of a prepositional by-phrase. The passive is formed with the auxiliary ***be*** and the past participle:

(3.54) Candy is eaten by children.

In other languages, the passive alternation might just involve changes in case marking, and some morphology on the verb.

Subcategorization

As we have seen, different verbs differ in the number of entities (persons, animals, things) that they relate. One such difference is the contrast between *transitive* and *intransitive* verbs. Transitive verbs have a (direct) object, intransitive verbs don't:

TRANSITIVE
INTRANSITIVE

(3.55) a. She brought a bottle of whiskey.

b. She walked (along the river).

In sentence (3.55a), a bottle *ofwhiskey* is the object of brought. We cannot use the verb *bring* without an object: we cannot say She *brought.* The verb *walk* is an example of an intransitive verb. There is no object in sentence (3.55). There is, however, a prepositional phrase expressing the location of the activity.

Syntacticians try to classify the dependents of verbs. The first distinction they make is between arguments and adjuncts. The subject, object, and direct object are arguments. In general, *arguments* express entities that are centrally involved in the activity of the verb. Most arguments are expressed as NPs, but they may be expressed as PPs, VPs, or as clauses:

ARGUMENTS

(3.56) a. We deprived him *of food.*

b. John knows *that he is losing.*

Arguments are divided into the subject, and all non-subject arguments which are collectively referred to as *complements.*

COMPLEMENTS
ADJUNCTS

Adjuncts are phrases that have a less tight link to the verb. Adjuncts are always optional whereas many complements are obligatory (for example, the object of *bring* is obligatory). Adjuncts can also move around more easily than complements. Prototypical examples of adjuncts are phrases that tell us the time, place, or manner of the action or state that the verb describes as in the following examples:

(3.5 7) a. She saw a Woody Allen movie yesterday.

b. She saw a Woody Allen movie *in Paris.*

c. She saw the Woody Allen movie with *great interest.*

d. She saw a Woody Allen movie *with a couple of friends.*

SUBORDINATE CLAUSES

Subordinate clauses (sentences within a sentence) can also be either adjuncts or subcategorized arguments, and can express a variety of relationships to the verb. In the examples we saw earlier in (3.28), (a) involves an argument clause, while the rest are adjuncts.

Sometimes, it's difficult to distinguish adjuncts and complements. The prepositional phrase on *the table* is a complement in the first sentence (it is subcategorized for by *put* and cannot be omitted), an adjunct in the second (it is optional):

(3.58) She put the book on *the table.*

(3.59) He gave his presentation on *the stage.*

The traditional argument/adjunct distinction is really a reflection of the categorical basis of traditional linguistics. In many cases, such as the following, one seems to find an intermediate degree of selection:

(3.60) a. I straightened the nail with a *hammer.*

b. He will retire *in Florida.*

It is not clear whether the PPs in italics should be regarded as being centrally involved in the event described by the verb or not. Within a Statistical NLP approach, it probably makes sense to talk instead about the degree of association between a verb and a dependent.

SUBCATEGORIZATION

We refer to the classification of verbs according to the types of complements they permit as ***subcategorization.*** We say that a verb *subcategorizes for* a particular complement. For example, ***bring*** subcategorizes for an object. Here is a list of subcategorized arguments with example sentences.

- **Subject.** *The children* eat candy.
- **Object.** The children eat candy.
- **Prepositional phrase.** She put the book on *the table.*
- **Predicative adjective.** We made the man angry.
- **Bare infinitive.** She helped me *walk.*

- **Infinitive with *to*.** She likes *to walk.*
- **Participial phrase.** She stopped singing *that tune* eventually.
- **That-clause. She** thinks *that it will ruin tomorrow.* The *that* can usually be omitted: She thinks *it will ruin tomorrow.*
- **Question-form clauses.** She is wondering why *it is ruining in August.* She asked me *what book I was reading.*

While most of these complements are phrasal units that we have already seen, such as NPs and APs, the final entries are not, in that they are a unit bigger than an S. The clause why *it is ruining in August* consists of a whole sentence *it is ruining in August* plus an additional constituent out front. Such a "large clause" is referred to as an S' (pronounced "S Bar") constituent. Relative clauses and main clause questions are also analyzed as S' constituents.

SUBCATEGORIZATION FRAME

Often verbs have several possible patterns of arguments. A particular set of arguments that a verb can appear with is referred to as a *subcategorization frame.* Here are some subcategorization frames that are common in English.

- **Intransitive verb.** NP[subject]. The *womun walked.*
- **Transitive verb.** NP[subject], NP[object]. *John loves Mary.*
- **Ditransitive verb.** NP[subject], NP[direct object], NP[indirect object]. *Mary gave Peter flowers.*
- **Intransitive with PP.** NP[subject], PP. *I rent in Puddington.*
- **Transitive with PP. NP[subject],** NP[object], PP. She *put the book on the table.*
- **Sentential complement.** NP[subject], **clause.** *I know (that) she likes you.*
- **Transitive with sentential complement.** NP[subj], NP[obj], **clause.** *She told me that Gary is coming on Tuesday.*

SELECTIONAL RESTRICTIONS
SELECTIONAL PREFERENCES

Subcategorization frames capture *syntactic* regularities about complements. There are also *semantic* regularities which are called *selectional restrictions* or *selectional preferences.* For example, the verb *bark* prefers dogs as subjects. The verb *eat* prefers edible things as objects:

(3.61) *The* Chihuahua barked all night.

(3.62) I eat *vegetables* every day.

Sentences that violate selectional preferences sound odd:

(3.63) a. *The cut* barked all night.

b. I eat philosophy every day.

▼ Selectional preferences are further discussed in section 8.4.

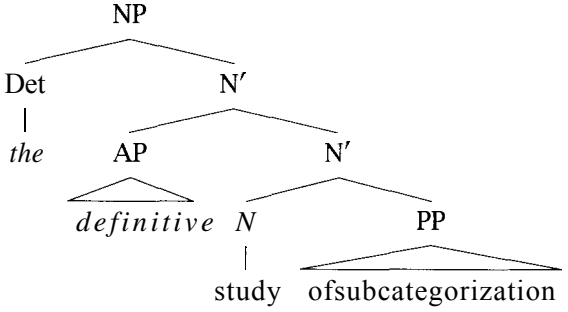
3.2.3 X' theory

Phrase structure rules as presented above do not predict any systematicity in the way that phrases in natural languages are made, nor any regularities for the appearance of different kinds of dependents in clauses. However, modern syntax has stressed that there are a lot of such regularities. An important idea is that a word will be the head of a phrase. The reason why we talk about noun phrases and prepositional phrases is because they are a constituent consisting of a noun or preposition respectively, and all their dependents. The noun or preposition heads the phrase.[3] Linguists have further argued that there is a broad systematicity in the way dependents arrange themselves around a head in a phrase. A head forms a small constituent with its complements. This constituent can be modified by adjuncts to form a bigger constituent, and finally this constituent can combine with a *specifier,* a subject or something like a determiner to form a maximal phrase. An example of the general picture is shown in (3.64):

HEAD

SPECIFIER

(3.64)



3. Recall, however, that verb phrases, as normally described, are slightly anomalous, since they include all the complements of the verb, but not the subject.

The intermediate constituents are referred to as N' nodes (pronounced "N bar nodes"). This is basically a two bar level theory (where we think of XP as X"), but is complicated by the fact that recursive *adjunction* of modifiers is allowed at the N' level to express that a noun can have any number of adjectival phrase modifiers. Sometimes people use theories with more or fewer bar levels.

ADJUNCTION

The final step of the argument is that while there may be differences in word order, this general pattern of constituency is repeated across phrase types. This idea is referred to as X' theory, where the X is taken to represent a variable across lexical categories.

3.2.4 Phrase structure ambiguity

GENERATION
PARSING

So far we have used rewrite rules to generate sentences. It is more common to use them in parsing, the process of reconstructing the derivation(s) or phrase structure tree(s) that give rise to a particular sequence of words. We call a phrase structure tree that is constructed from a sentence a parse. For example, the tree in (3.43) is a parse of sentence (3.41).

PARSE

In most cases, there are many different phrase structure trees that could all have given rise to a particular sequence of words. A parser based on a comprehensive grammar of English will usually find hundreds of parses for a sentence. This phenomenon is called phrase structure ambiguity or syntactic ambiguity. We saw an example of a syntactically ambiguous sentence in the introduction, example (1.10): *Our company is training workers.* One type of syntactic ambiguity that is particularly frequent is *attachment ambiguity.*

SYNTACTIC AMBIGUITY

ATTACHMENT AMBIGUITY

Attachment ambiguities occur with phrases that could have been generated by two different nodes. For example, according to the grammar in (3.39), there are two ways to generate the prepositional phrase *with a spoon* in sentence (3.65):

(3.65) The children ate the cake with a spoon.

It can be generated as a child of a verb phrase, as in the parse tree shown in figure 3.2 (a), or as a child of one of the noun phrases, as in the parse tree shown in figure 3.2 (b).

Different attachments have different meanings. The 'high' attachment to the verb phrase makes a statement about the instrument that the children used while eating the cake. The 'low' attachment to the noun phrase tells us which cake was eaten (the cake with a spoon, and not, say, the

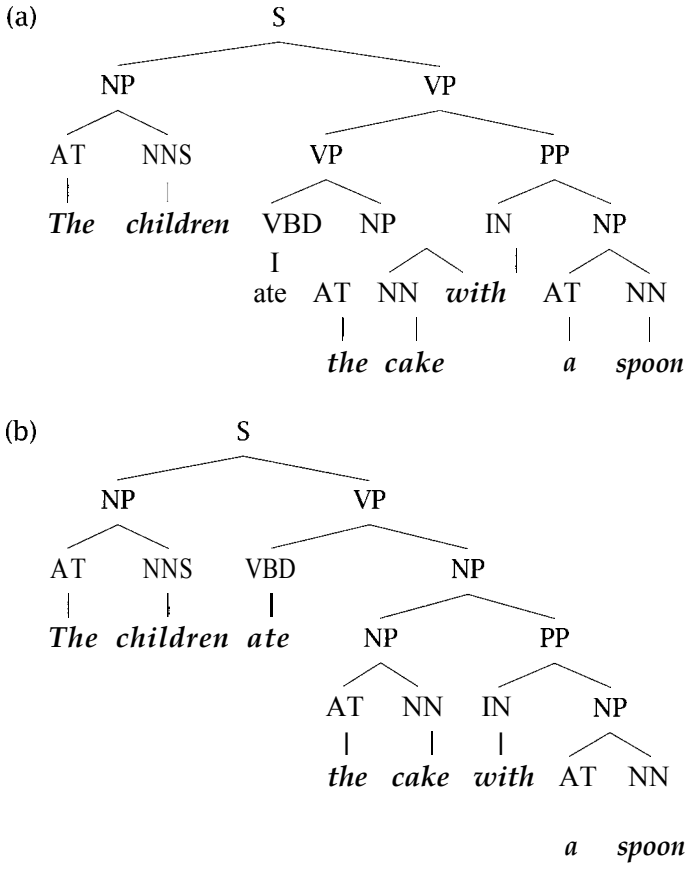


Figure 3.2 An example of a prepositional phrase attachment ambiguity.

cake with icing). So resolving attachment ambiguities can be important for finding the correct semantic interpretation.

GARDEN PATHS

A much-studied subclass of syntactic ambiguity is the phenomenon of *garden pathing*. A garden path sentence leads you along a path that suddenly turns out not to work. For example, there might turn out to be additional words in the sentence that do not seem to belong there:

(3.66) The horse raced past the barn fell.

Sentence (3.66) from (Bever 1970) is probably the most famous example of a garden path sentence. By the time most people get to the word *barn*, they have constructed a parse that roughly corresponds to the meaning

'The horse ran past the barn.' But then there is an additional word fell that cannot be incrementally added to this parse. We have to backtrack to ***raced*** and construct a completely different parse, corresponding to the meaning The horse fell ***after it had been raced past the burn.*** Garden pathing is the phenomenon of first being tricked into adopting a spurious parse and then having to backtrack to try to construct the right parse.

Garden-path sentences are rarely a problem in spoken language. Semantic preferences, the generosity of speakers in following communicative maxims, and intonational patterns all usually prevent us from garden pathing (MacDonald et al. 1994; Tanenhaus and Trueswell 1995). We can see this in sentence (3.66) where an intonational break between horse and ***raced*** would tip the hearer off that ***raced*** introduces a reduced relative clause, not the verb of the main clause. However, garden-pathing can be a real problem when reading complex sentences of written English.

We have seen examples of sentences with more than one parse due to syntactic ambiguity. Most sentences are of this type. But it is also possible that a sentence will have no parse at all. The reason could be that a rule was used in the generation of the sentence that is not covered by the grammar. The other possibility is that the sentence is *ungrammatical* or not syntactically well-formed. Here is an example of an ungrammatical sentence.

UNGRAMMATICAL

(3.67) *Slept children the.

It is important to distinguish ungrammaticality from semantic abnormality. Sentences like the following are odd, but they are jarring because their semantic interpretation is incoherent whereas (3.67) does not have an interpretation at all.

(3.68) a. Colorless green ideas sleep furiously.

b. The cat barked.

People often use a hash mark (#) to indicate semantic, pragmatic, or cultural oddness, as opposed to the marks we introduced earlier for syntactic illformedness.

3.3 Semantics and Pragmatics

Semantics is the study of the meaning of words, constructions, and utterances. We can divide semantics into two parts, the study of the meaning

LEXICAL SEMANTICS

of individual words (or *lexical semantics)* and the study of how meanings of individual words are combined into the meaning of sentences (or even larger units).

One way to approach lexical semantics is to study how word meanings are related to each other. We can organize words into a lexical hierarchy, as is the case, for example, in WordNet, which defines *hypernymy* and *hyponymy*. A hypernym or *hyperonym*[4] is a word with a more general sense, for example, animal is a hypernym of cat. A hyponym is a word with a more specialized meaning: *cut* is a hyponym of animal. (In general, if w^1 is a hypernym of w^2, then w^2 is a hyponym of w^1.) *Antonyms* are words with opposite meanings: *hot* and *cold* or long and *short.* The part-whole relationship is called *meronymy*. The word *tire* is a meronym of *car* and *leaf* is a meronym of tree. The whole corresponding to a part is called a *holonym.*

HYPERNYMY
HYPONYMY
HYPERONYM
ANTONYMS
MERONYMY
HOLONYM

Synonyms are words with the same meaning (or very similar meaning): *car* and *automobile* are synonyms. *Homonyms* are words that are written the same way, but are (historically or conceptually) really two different words with different meanings which seem unrelated. Examples are suit ('lawsuit' and 'set of garments') and bunk ('river bank' and 'financial institution'). If a word's meanings (or senses) are related, we call it a *polyseme.* The word brunch is polysemous because its senses ('natural subdivision of a plant' and 'a separate but dependent part of a central organization') are related. Lexical *ambiguity* can refer to both homonymy and polysemy. The subcase of homonymy where the two words are not only written the same way, but also have identical pronunciation, is called *homophony. So* the words buss for a species of fish and *bass* for a low-pitched sound are homonyms, but they are not homophones.

SYNONYMS
HOMONYMS
SENSES
POLY SEME
AMBIGUITY
HOMOPHONY

▼ Disambiguating word senses is the topic of chapter 7.

Once we have the meanings of individual words, we need to assemble them into the meaning of the whole sentence. That is a hard problem because natural language often does not obey the principle of *compositionality* by which the meaning of the whole can be strictly predicted from the meaning of the parts. The word *white* refers to very different colors in the following expressions:

COMPOSITIONALITY

(3.69) white paper, white hair, white skin, white wine

White hair is grey, a white skin really has a rosy color, and white wine

4. The latter is prescriptively correct. The former is more commonly used.

is actually yellow (but yellow wine doesn't sound very appealing). The groupings white *hair, white skin,* and *white wine* are examples of collocations. The meaning of the whole is the sum of the meanings of the part plus some additional semantic component that cannot be predicted from the parts.

COLLOCATIONS

▼ Collocations are the topic of chapter 5.

If the relationship between the meaning of the words and the meaning of the phrase is completely opaque, we call the phrase an *idiom.* For example, the idiom *to kick the bucket* describes a process, dying, that has nothing to do with kicking and buckets. We may be able to explain the historical origin of the idiom, but in today's language it is completely non-compositional. Another example is the noun-noun compound *carriage return* for the character that marks the end of a line. Most younger speakers are not aware of its original meaning: returning the carriage of a typewriter to its position on the left margin of the page when starting a new line.

IDIOM

There are many other important problems in assembling the meanings of larger units, which we will not discuss in detail here. One example is the problem of scope. Quantifiers and operators have a scope which extends over one or more phrases or clauses. In the following sentence, we can either interpret the quantifier everyone as having scope over the negative not (meaning that not one person went to the movie), or we can interpret the negation as having scope over the quantifier (meaning that at least one person didn't go to the movie):

SCOPE

3.70) Everyone didn't go to the movie.

In order to derive a correct representation of the meaning of the sentence, we need to determine which interpretation is correct in context.

The next larger unit to consider after words and sentences is a *discourse.* Studies of discourse seek to elucidate the covert relationships between sentences in a text. In a narrative discourse, one can seek to describe whether a following sentence is an example, an elaboration, a restatement, etc. In a conversation one wants to model the relationship between turns and the kinds of speech acts involved (questions, statements, requests, acknowledgments, etc.). A central problem in *discourse analysis* is the resolution of *anaphoric relations.*

DISCOURSE ANALYSIS

ANAPHORIC RELATIONS

(3.71) a. Mary helped *Peter* get out of the cab. *He* thanked her.

b. Mary helped *the other passenger* out of the cab. *The man* had asked her to help him because of his foot injury.

Anaphoric relations hold between noun phrases that refer to the same person or thing. The noun phrases Peter and He in sentence (3.71a) and *the other passenger* and *The man* in sentence (3.71b) refer to the same person. The resolution of anaphoric relations is important for *information extraction.* In information extraction, we are scanning a text for a specific type of event such as natural disasters, terrorist attacks or corporate acquisitions. The task is to identify the participants in the event and other information typical of such an event (for example the purchase price in a corporate merger). To do this task well, the correct identification of anaphoric relations is crucial in order to keep track of the participants.

INFORMATION EXTRACTION

(3.72) Hurricane Hugo destroyed 20,000 Florida homes. At an estimated cost of one billion dollars, the disaster has been the most costly in the state's history.

If we identify *Hurricane Hugo* and *the disaster* as referring to the same entity in mini-discourse (3.72), we will be able to give Hugo as an answer to the question: *Which hurricanes caused more than a billion dollars worth of damage?*

PRAGMATICS

Discourse analysis is part of *pragmatics*, the study of how knowledge about the world and language conventions interact with literal meaning. Anaphoric relations are a pragmatic phenomenon since they are constrained by world knowledge. For example, for resolving the relations in discourse (3.72), it is necessary to know that hurricanes are disasters. Most areas of pragmatics have not received much attention in Statistical NLP, both because it is hard to model the complexity of world knowledge with statistical means and due to the lack of training data. Two areas that are beginning to receive more attention are the resolution of anaphoric relations and the modeling of speech acts in dialogues.

3.4 Other Areas

Linguistics is traditionally subdivided into phonetics, phonology, morphology, syntax, semantics, and pragmatics. Phonetics is the study of the physical sounds of language, phenomena like consonants, vowels and intonation. The subject of phonology is the structure of the sound systems

in languages. Phonetics and phonology are important for speech recognition and speech synthesis, but since we do not cover speech, we will not cover them in this book. We will introduce the small number of phonetic and phonological concepts we need wherever we first refer to them.

In addition to areas of study that deal with different levels of language, there are also subfields of linguistics that look at particular aspects of language. *Sociolinguistics* studies the interactions of social organization and language. The change of languages over time is the subject of *historical linguistics*. Linguistic typology looks at how languages make different use of the inventory of linguistic devices and how they can be classified into groups based on the way they use these devices. Language acquisition investigates how children learn language. Psycholinguistics focuses on issues of real-time production and perception of language and on the way language is represented in the brain. Many of these areas hold rich possibilities for making use of quantitative methods. Mathematical linguistics is usually used to refer to approaches using non-quantitative mathematical methods.

SOCIOLINGUISTICS

HISTORICAL LINGUISTICS

3.5 Further Reading

In-depth overview articles of a large number of the subfields of linguistics can be found in (Newmeyer 1988). In many of these areas, the influence of Statistical NLP can now be felt, be it in the widespread use of corpora, or in the adoption of quantitative methods from Statistical NLP.

De Saussure 1962 is a landmark work in structuralist linguistics. An excellent in-depth overview of the field of linguistics for non-linguists is provided by the Cambridge Encyclopedia of Language (Crystal 1987). See also (Pinker 1994) for a recent popular book. Marchand (1969) presents an extremely thorough study of the possibilities for word derivation in English. Quirk et al. (1985) provide a comprehensive grammar of English. Finally, a good work of reference for looking up syntactic (and many morphological and semantic) terms is (Trask 1993).

Good introductions to speech recognition and speech synthesis are: (Waibel and Lee 1990; Rabiner and Juang 1993; Jelinek 1997).

3.6 Exercises

Exercise 3.1 [⋆]

What are the parts of speech of the words in the following paragraph?

(3.73) The lemon is an essential cooking ingredient. Its sharply fragrant juice and tangy rind is added to sweet and savory dishes in every cuisine. This enchanting book, written by cookbook author John Smith, offers a wonderful array of recipes celebrating this internationally popular, intensely flavored fruit.

Exercise 3.2 [⋆]

Think of five examples of noun-noun compounds.

Exercise 3.3 [⋆]

Identify subject, direct object and indirect object in the following sentence.

(3.74) He baked her an apple pie.

Exercise 3.4 [⋆]

What is the difference in meaning between the following two sentences?

(3.75) a. Mary defended her.

b. Mary defended herself.

Exercise 3.5 [⋆]

What is the standard word order in the English sentence (a) for declaratives, (b) for imperatives, (c) for interrogatives?

Exercise 3.6 [⋆]

What are the comparative and superlative forms for the following adjectives and adverbs?

(3.76) good, well, effective, big, curious, bad

Exercise 3.7 [⋆]

Give base form, third singular present tense form, past tense, past participle, and present participle for the following verbs.

(3.77) throw, do, laugh, change, carry, bring, dream

Exercise 3.8 [⋆]

Transform the following sentences into the passive voice.

(3.78) a. Mary carried the suitcase up the stairs.

b. Mary gave John the suitcase.

Exercise 3.9 [⋆]

What is the difference between a preposition and a particle? What grammatical function does ***in*** have in the following sentences?

(3.79) a. Mary lives in London.
b. When did Mary move in?
c. She puts in a lot of hours at work.
d. She put the document in the wrong folder.

Exercise 3.10 [⋆]

Give three examples each of transitive verbs and intransitive verbs.

Exercise 3.11 [⋆]

What is the difference between a complement and an adjunct? Are the italicized phrases in the following sentences complements or adjuncts? What type of complements or adjuncts?

(3.80) a. She goes to Church on ***Sundays.***
b. She went ***to London.***
c. Peter relies ***on Mary*** for help with his homework.
d. The book is lying ***on the table.***
e. She watched him ***with a telescope.***

Exercise 3.12 [⋆]

The italicized phrases in the following sentences are examples of attachment ambiguity. What are the two possible interpretations?

(3.81) Mary saw the man ***with the telescope.***

(3.82) The company experienced growth in classified advertising ***and preprinted inserts.***

Exercise 3.13 [⋆]

Are the following phrases compositional or non-compositional?

(3.83) to beat around the bush, to eat an orange, to kick butt, to twist somebody's arm, help desk, computer program, desktop publishing, book publishing, the publishing industry

Exercise 3.14 [⋆]

Are phrasal verbs compositional or non-compositional?

Exercise 3.15 [⋆]

In the following sentence, either a few ***actors*** or ***everybody*** can take wide scope over the sentence. What is the difference in meaning?

(3.84) A few actors are liked by everybody.