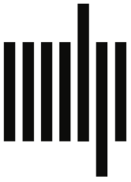
Foundations of Statistical Natural Language Processing

Christopher Manning and Hinrich Schuetze



The MIT Press

From The MIT Press

MITCogNet

Second printing with corrections 1999
Third printing 2000, fourth printing 2001
Fifth printing with corrections

Typeset in 10/13 Lucida Bright by the authors using LaTeX 2ε.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Information

Manning, Christopher D.
Foundations of statistical natural language processing / Christopher D. Manning, Hinrich Schütze.
p. cm.
Includes bibliographical references (p.) and index.
ISBN 0-262-13360-1
1. Computational linguistics—Statistical methods. I. Schütze, Hinrich. II. Title.
P98.5.S83M36 1999
410'.285—dc21 99-21137
CIP

10 9 8 7 6

3 *Linguistic Essentials*

THIS CHAPTER introduces basic linguistic concepts, which are necessary for making sense of discussions in the rest of the book. It may partly be a review of things you learned at school, but it will go into more depth for syntactic phenomena like attachment ambiguities and phrase structure that are important in NLP. Apart from syntax (sentence structure), we will cover some morphology (word formation) and semantics (meaning). The last section will give an overview of other areas of linguistics and pointers to further reading.

3.1 Parts of Speech and Morphology

SYNTACTIC CATEGORIES
GRAMMATICAL CATEGORIES
PARTS OF SPEECH
NOUN
VERB
ADJECTIVE
SUBSTITUTION TEST

Linguists group the words of a language into classes (sets) which show similar syntactic behavior, and often a typical semantic type. These word classes are otherwise called *syntactic* or *grammatical categories*, but more commonly still by the traditional name *parts of speech* (POS). Three important parts of speech are *noun*, *verb*, and *adjective*. *Nouns* typically refer to people, animals, concepts and things. The prototypical *verb* is used to express the action in a sentence. *Adjectives* describe properties of nouns. The most basic test for words belonging to the same class is the *substitution test*. Adjectives can be picked out as words that occur in the frame in (3.1):

(3.1) The $\left\{\begin{array}{l}\text{sad}\\ \text{intelligent}\\ \text{green}\\ \text{fat}\\ \ldots\end{array}\right\}$ one is in the corner.

In sentence (3.2), the noun *children* refers to a group of people (those of young age) and the noun *candy* to a particular type of food:

(3.2) Children eat sweet candy.

The verb *eat* describes what children do with candy. The adjective *sweet* tells us about a property of candy, namely that it is sweet. Many words have multiple parts of speech: *candy* can also be a verb (as in *Too much boiling will candy the molasses*), and, at least in British English, *sweet* can be a noun, meaning roughly the same as *candy*. Word classes are normally divided into two. The *open* or *lexical categories* are ones like nouns, verbs and adjectives which have a large number of members, and to which new words are commonly added. The *closed* or *functional categories* are categories such as prepositions and determiners (containing words like *of, on, the, a*) which have only a few members, and the members of which normally have a clear grammatical use. Normally, the various parts of speech for a word are listed in an online *dictionary*, otherwise known as a *lexicon*.

OPEN WORD CLASS
LEXICAL CATEGORIES
CLOSED WORD CLASS
FUNCTIONAL CATEGORIES
DICTIONARY
LEXICON

Traditional systems of parts of speech distinguish about 8 categories, but corpus linguists normally want to use more fine-grained classifications of word classes. There are well-established sets of abbreviations for naming these classes, usually referred to as POS *tags*. In this chapter, as we introduce syntactic categories, we will give the abbreviations used in the Brown corpus for the more important categories. For example, adjectives are tagged using the code JJ in the Brown corpus. Because of its pioneering role, the Brown corpus tags are particularly widely known.
▼ We briefly describe and compare several well-known tag sets in section 4.3.2.

TAGS

Word categories are systematically related by *morphological processes* such as the formation of the *plural* form (*dog-s*) from the *singular* form of the noun (*dog*). Morphology is important in NLP because language is productive: in any given text we will encounter words and word forms that we haven't seen before and that are not in our precompiled dictionary. Many of these new words are morphologically related to known words. So if we understand morphological processes, we can infer a lot about the syntactic and semantic properties of new words.

MORPHOLOGICAL PROCESSES
PLURAL
SINGULAR

It is important to be able to handle morphology in English, but it's absolutely essential when it comes to highly inflecting languages like Finnish. In English, a regular verb has only 4 distinct forms, and irregular verbs have at most 8 forms. One can accomplish a fair amount without mor-

phology, by just listing all word forms. In contrast, a Finnish verb has more than 10,000 forms! For a language like Finnish, it would be tedious and impractical to enumerate all verb forms as an enormous list.

INFLECTION
ROOT FORM
PREFIXES
SUFFIXES

The major types of morphological process are inflection, derivation, and compounding. *Inflections* are the systematic modifications of a *root form* by means of *prefixes* and *suffixes* to indicate grammatical distinctions like singular and plural. Inflection does not change word class or meaning significantly, but varies features such as tense, number, and plurality. All the inflectional forms of a word are often grouped as manifestations of a single *lexeme*.

LEXEME

DERIVATION

Derivation is less systematic. It usually results in a more radical change of syntactic category, and it often involves a change in meaning. An example is the derivation of the adverb *widely* from the adjective *wide* (by appending the suffix *-ly*). *Widely* in a phrase like *it is widely believed* means *among a large well-dispersed group of people*, a shift from the core meaning of *wide* (*extending over a vast area*). Adverb formation is also less systematic than plural inflection. Some adjectives like *old* or *difficult* don't have adverbs: **oldly* and **difficultly* are not words of English. Here are some other examples of derivations: the suffix *-en* transforms adjectives into verbs (*weak-en*, *soft-en*), the suffix *-able* transforms verbs into adjectives (*understand-able*, *accept-able*), and the suffix *-er* transforms verbs into nouns (*teach-er*, *lead-er*).

COMPOUNDING

Compounding refers to the merging of two or more words into a new word. English has many noun-noun compounds, nouns that are combinations of two other nouns. Examples are *tea kettle*, *disk drive*, or *college degree*. While these are (usually) written as separate words, they are pronounced as a single word, and denote a single semantic concept, which one would normally wish to list in the lexicon. There are also other compounds that involve parts of speech such as adjectives, verbs, and prepositions, such as *downmarket*, *(to) overtake*, and *mad cow disease*.

We will now introduce the major parts of speech of English.

3.1.1 Nouns and pronouns

Nouns typically refer to entities in the world like people, animals, and things. Examples are:

(3.3) dog, tree, person, hat, speech, idea, philosophy

Type of inflection	Instances
number	singular, plural
gender	feminine, masculine, neuter
case	nominative, genitive, dative, accusative

Table 3.1 Common inflections of nouns.

English, which is morphologically impoverished compared to many other languages, has only one inflection of the noun, the plural form. It is usually formed by appending the suffix *-s*. Here are some nouns with their singular and plural forms.

(3.4)

dog : dogs	tree : trees	person : persons
hat : hats	speech : speeches	woman : women
idea : ideas	philosophy : philosophies	child : children

The plural suffix has three pronunciations, /s/ as in *hats*, /z/, as in *boys*, and /əz/ as in *speeches*, the last case being represented by insertion of an *e* in the writing system. A few forms like *women* don't follow the regular pattern, and are termed *irregular*.

IRREGULAR

Number (singular and plural) is one common grammatical distinction that is marked on the noun. Two other types of inflection that are common for nouns across languages are gender and case as shown in table 3.1.

English does not have a system of gender inflections, but it does have different gender forms for the third person singular pronoun: *he* (masculine), *she* (feminine), and *it* (neuter). An example of gender inflection of nouns from Latin is the endings *-a* for feminine and *-us* for masculine. Examples: *fili-us* 'son, male child'; *fili-a* 'daughter, female child.' In some languages, grammatical gender is closely correlated with the sex of the person referred to as it is for these two Latin words (female → feminine, male → masculine, neither → neuter), but in other languages gender is a largely arbitrary grammatical category. An example linguists are fond of is the German word for girl, *Mädchen*, which is neuter.

In some languages, nouns appear in different forms when they have different functions (subject, object, etc.) in a sentence, and these forms are called *cases*. For example, the Latin for 'son' is *filius* when the subject, but *filium* when the object of a verb. Many languages have a rich array of case inflections, with cases for locatives, instrumentals, etc. English

CASE

has no real case inflections. The only case relationship that is systematically indicated is the genitive. The genitive describes the possessor. For example, the phrase *the woman's house* indicates that the woman owns the house. The genitive is usually written *'s*, but just as *'* after words that end in *s*, which includes most plural nouns such as in *the students' grievances*. Although *'s* initially looks like a case inflection, it is actually what is termed a *clitic*, also known as a *phrasal affix*, because it can appear not only attached to nouns, but after other words that modify a noun, as in *the person you met's house was broken into*.

CLITIC

PRONOUN

Pronouns are a separate small class of words that act like variables in that they refer to a person or thing that is somehow salient in the discourse context. For example, the pronoun *she* in sentence (3.5) refers to the most salient person (of feminine gender) in the context of use, which is Mary.

(3.5) After *Mary* arrived in the village, *she* looked for a bed-and-breakfast.

As well as distinguishing the number of their antecedent, they also mark person (1st = speaker, 2nd = hearer, or 3rd = other discourse entities). They are the only words in English which appear in different forms when they are used as the subject and the object of a sentence. We call these forms the *nominative* or *subject case* and *accusative* or *object case* personal pronouns, respectively. Pronouns also have special forms, *possessive pronouns*, for when they are a possessor, as in *my car*, which we can view as genitive case forms. Somewhat oddly, English pronouns have another possessive form, often called the 'second' possessive personal pronoun, used when the object of the preposition *of* describes the possessor: *a friend of* ***mine***. Finally, there are *reflexive pronouns*, which are used similarly to ordinary (personal) pronouns except that they always refer to a nearby antecedent in the same sentence, normally the subject of the sentence. For example, *herself* in sentence (3.6a) must refer to Mary whereas *her* in sentence (3.6b) cannot refer to Mary (that is, Mary saw a woman other than herself in the mirror).

NOMINATIVE
SUBJECT CASE
ACCUSATIVE
OBJECT CASE
POSSESSIVE PRONOUNS

REFLEXIVE PRONOUNS

(3.6) a. Mary saw herself in the mirror.

b. Mary saw her in the mirror.

Reflexive pronouns (and certain other expressions like *each other*) are often referred to as *anaphors*, and must refer to something very nearby in the text. Personal pronouns also refer to previously discussed people

ANAPHORS

	Nominative	Accusative	Possessive	2nd Possessive	Reflexive
Tag(s)	PPS (3SG) PPSS (1SG,2SG,PL)	PPO	PP$	PP$$	PPL (PPLS for PL)
1SG	I	me	my	mine	myself
2SG	you	you	your	yours	yourself
3SG MASC	he	him	his	his	himself
3SG FEM	she	her	her	hers	herself
3SG NEUT	it	it	its	its	itself
1PL	we	us	our	ours	ourselves
2PL	you	you	your	yours	yourselves
3PL	they	them	their	theirs	themselves

Table 3.2 Pronoun forms in English. Second person forms do not distinguish number, except in the reflexive, while third person singular forms distinguish gender.

and things, but at a slightly greater distance. All the forms for pronouns, and their Brown tags are summarized in table 3.2.

PROPER NAMES
ADVERBIAL NOUNS

Brown tags. NN is the Brown tag for singular nouns (*candy*, *woman*). The Brown tag set also distinguishes two special types of nouns, *proper nouns* (or *proper names*), and *adverbial nouns*. Proper nouns are names like *Mary*, *Smith*, or *United States* that refer to particular persons or things. Proper nouns are usually capitalized. The tag for proper nouns is NNP.[1] Adverbial nouns (tag NR) are nouns like *home*, *west* and *tomorrow* that can be used without modifiers to give information about the circumstances of the event described, for example the time or the location. They have a function similar to *adverbs* (see below). The tags mentioned so far have the following plural equivalents: NNS (plural nouns), NNPS (plural proper nouns), and NRS (plural adverbial nouns). Many also have possessive or genitive extensions: NN$ (possessive singular nouns), NNS$ (possessive plural nouns), NNP$ (possessive singular proper nouns), NNPS$ (possessive plural proper nouns), and NR$ (possessive adverbial nouns). The tags for pronouns are shown in table 3.2.

1. Actually, the Brown tag for proper nouns was NP, but we follow the Penn Treebank in substituting NNP, so that NP can maintain its conventional meaning within linguistics of a noun phrase (see below). We also follow the Penn Treebank in using a doubled N in the related tags mentioned subsequently.

3.1.2 Words that accompany nouns: Determiners and adjectives

DETERMINER
ARTICLE

Several other parts of speech most commonly appear accompanying nouns. *Determiners* describe the particular reference of a noun. A subtype of determiners is *articles*. The article *the* indicates that we're talking about someone or something that we already know about or can uniquely determine. We say *the tree* if we have already made reference to the tree or if the reference is clear from context such as when we are standing next to a tree and it is clear we are referring to it. The article *a* (or *an*) indicates that the person or thing we are talking about was not previously mentioned. If we say *a tree*, then we are indicating that we have not mentioned this tree before and its identity cannot be inferred from context. Other determiners include the *demonstratives*, such as *this* and *that*.

DEMONSTRATIVES
ADJECTIVE

Adjectives are used to describe properties of nouns. Here are some adjectives (in italics):

(3.7) a *red* rose, this *long* journey, many *intelligent* children, a very *trendy* magazine

ATTRIBUTIVE
ADNOMINAL
PREDICATIVE

Uses such as these modifying a noun are called *attributive* or *adnominal*. Adjectives also have a *predicative* use as a complement of *be*:

(3.8) The rose is *red*. The journey will be *long*.

AGREEMENT

Many languages mark distinctions of case, number, and gender on articles and adjectives as well as nouns, and we then say that the article or adjective *agrees* with the noun, that is, they have the same case, number, and gender. In English, the morphological modifications of adjectives are the derivational endings like *-ly* which we covered earlier, and the formation of *comparative* (*richer*, *trendier*), and *superlative* (*richest*, *trendiest*) forms. Only some, mainly shorter, adjectives form morphological comparatives and superlatives by suffixing *-er* and *-est*. For the rest, *periphrastic forms* are used (*more intelligent*, *most intelligent*). Periphrastic forms are formed with the auxiliary words, in this case *more* and *most*. The basic form of the adjective (*rich*, *trendy*, *intelligent*) is called the *positive* when contrasted with comparative and superlative. Comparative and superlative forms compare different degrees to which the property described by the adjective applies to nouns. The following example should be self-explanatory:

COMPARATIVE
SUPERLATIVE
PERIPHRASTIC FORMS
POSITIVE

(3.9) John is rich, Paul is richer, Mary is richest.

Brown tags. The Brown tag for adjectives (in the positive form) is JJ, for comparatives JJR, for superlatives JJT. There is a special tag, JJS, for the 'semantically' superlative adjectives *chief*, *main*, and *top*. Numbers are subclasses of adjectives. The cardinals, such as *one*, *two*, and *6,000,000*, have the tag CD. The ordinals, such as *first*, *second*, *tenth*, and *mid-twentieth* have the tag OD.

The Brown tag for articles is AT. Singular determiners, like *this*, *that*, have the tag DT; plural determiners (*these*, *those*) DTS; determiners that can be singular or plural (*some*, *any*) DTI, and 'double conjunction' determiners (*either*, *neither*) DTX.

QUANTIFIER

Quantifiers are words that express ideas like 'all,' 'many,' 'some.' The determiners *some* and *any* can function as quantifiers. Other parts of speech that correspond to quantifiers have the tags ABN (pre-quantifier: *all*, *many*) and PN (nominal pronoun: *one*, *something*, *anything*, *somebody*). The tag for *there* when used to express existence at the beginning of a sentence is EX.

INTERROGATIVE PRONOUNS
INTERROGATIVE DETERMINERS

A final group of words that occur with or instead of nouns are the *interrogative pronouns* and *determiners* which are used for questions and relative clauses. Their tags are WDT (*wh*-determiner: *what*, *which*), WP$ (possessive *wh*-pronoun: *whose*), WPO (objective *wh*-pronoun: *whom*, *which*, *that*), and WPS (nominative *wh*-pronoun: *who*, *which*, *that*).

3.1.3 Verbs

Verbs are used to describe actions (*She* ***threw*** *the stone*), activities (*She* ***walked*** *along the river*) and states (*I* ***have*** *$50*). A regular English verb has the following morphological forms:

BASE FORM

- the root or *base form*: *walk*
- the third singular present tense: *walks*
- the gerund and present participle: *walking*
- the past tense form and past/passive participle: *walked*

Most of these forms do duty in several functions. The base form is used for the *present tense*.

PRESENT TENSE

(3.10) I walk. You walk. We walk. You (guys) walk. They walk.

The third singular person has a different present tense form:

(3.11) She walks. He walks. It walks.

INFINITIVE The base form is also used for the *infinitive* with *to*:

(3.12) She likes *to walk*. She has *to walk*. *To walk* is fun.

and after modals and in the bare infinitive:

(3.13) She shouldn't *walk*. She helped me *walk*.

PROGRESSIVE The *-ing* form is used for the *progressive* (indicating that an action is in progress):

(3.14) She is walking. She was walking. She will be walking.

GERUND and as the *gerund*, a derived form where the verb gains some or all of the properties of nouns:

(3.15) This is the most vigorous *walking* I've done in a long time. *Walking* is fun.

The *-ed* form serves as past tense indicating an action that took place in the past:

(3.16) She walked.

PRESENT PERFECT It also functions as the past participle in the formation of *present perfect*:

(3.17) She has walked.

PAST PERFECT and *past perfect*:

(3.18) She had walked.

IRREGULAR A number of verbs are *irregular* and have different forms for past tense and past participle. Examples are *drive* and *take*:

(3.19) a. She *drove* the car. She has never *driven* a car.

b. She *took* off on Monday. She had already *taken* off on Monday.

Just as nouns are commonly marked for features like number and case, verbs are also commonly marked for certain features. Table 3.3 summarizes grammatical features that are commonly indicated on verbs across languages. These features can be indicated either morphologically

SYNTHETIC FORMS (also called *synthetically*), as in the case of the English endings *-s*, *-ing*,

Feature Category	Instances
subject number	singular, plural
subject person	first (*I walk*), second (*you walk*), third (*she walks*)
tense	present tense, past tense, future tense
aspect	progressive, perfect
mood/modality	possibility, subjunctive, irrealis
participles	present participle (*walking*), past participle (*walked*)
voice	active, passive, middle

Table 3.3 Features commonly marked on verbs.

AUXILIARIES
VERB GROUP
ANALYTIC FORMS

and *-ed*, or by means of *auxiliaries*, words that accompany verbs in a *verb group* (also called *analytically*). English uses the auxiliaries *have*, *be*, and *will* (and others) to express aspect, mood, and some tense information. The present and past perfect are formed with *have* as we saw in sentences (3.17) and (3.18). The progressive is formed with *be* (3.14). Forms that are built using auxiliaries, as opposed to direct inflection as in the case of the English past tense, are referred to as *periphrastic forms*.

PERIPHRASTIC FORMS

MODAL AUXILIARIES

In English, there is a class of verbs with special properties: the *modal auxiliaries* or *modals*. Modals lack some of the forms of ordinary verbs (no infinitive, no progressive form), and always come first in the verb group. They express modalities like possibility (*may*, *can*) or obligation (*should*) as illustrated in the following examples:

(3.20) a. With her abilities, she *can* do whatever she wants to.

b. He *may* or *may* not come to the meeting.

c. You *should* spend more time with your family.

In English, the formation of the future tense with the auxiliary *will* is in all ways parallel to that of other modalities:

(3.21) She *will* come. She *will* not come.

Brown tags. The Brown tag set uses VB for the base form (*take*), VBZ for the third person singular (*takes*), VBD for the past tense (*took*), VBG for gerund and present participle (*taking*), and VBN for the past participle (*taken*). The tag for modal auxiliaries (*can*, *may*, *must*, *could*, *might*, ...) is MD. Since *be*, *have*, and *do* are important in forming tenses and moods,

the Brown tag set has separate tags for all forms of these verbs. We omit them here, but they are listed in table 4.6.

3.1.4 Other parts of speech

Adverbs, prepositions, and particles

ADVERB

We have already encountered adverbs as an example of morphological derivation. *Adverbs* modify a verb in the same way that adjectives modify nouns. Adverbs specify place, time, manner or degree:

(3.22) a. She *often* travels to Las Vegas.

b. She *allegedly* committed perjury.

c. She started her career off very *impressively*.

Some adverbs, such as *often*, are not derived from adjectives and lack the suffix *-ly*.

Some adverbs can also modify adjectives ((3.23a) and (3.23b)) and other adverbs (3.23c).

(3.23) a. a *very* unlikely event

b. a *shockingly* frank exchange

c. She started her career off *very* impressively.

DEGREE ADVERBS

QUALIFIER

Certain adverbs like *very* are specialized to the role of modifying adjectives and adverbs and do not modify verbs. They are called *degree adverbs*. Their distribution is thus quite distinct from other adverbs, and they are sometimes regarded as a separate part of speech called *qualifiers*.

PREPOSITION

Prepositions are mainly small words that prototypically express spatial relationships:

(3.24) *in* the glass, *on* the table, *over* their heads, *about* an interesting idea, *concerning* your recent invention

PARTICLE

PHRASAL VERBS

Most prepositions do double duty as particles. *Particles* are a subclass of prepositions that can enter into strong bonds with verbs in the formation of so-called *phrasal verbs*. We can best think of a phrasal verb as a separate lexical entry with syntactic and semantic properties different from the verb it was formed from. Here are some examples:

(3.25) a. The plane *took off* at 8am.

b. Don't *give in* to him.

c. It is time to *take on* new responsibilities.

d. He was *put off* by so much rudeness.

Sometimes these constructions can occur with the preposition separated from the verb:

(3.26) a. I didn't want to *take* that responsibility *on* right now.

b. He *put* me *off*.

These phrasal verbs have particular meanings that are quite specialized, and unpredictable from the verb and particle that make them up.

Sometimes we need to know the meaning of a sentence to be able to distinguish particles and prepositions: *up* is a preposition in (3.27a) and a particle in (3.27b). Note the meaning shift from the literal meaning of running on an incline in (3.27a) to the figurative meaning of building up a large bill in (3.27b).

(3.27) a. She ran up a hill.

b. She ran up a bill.

Brown tags. The tags for adverbs are RB (ordinary adverb: *simply*, *late*, *well*, *little*), RBR (comparative adverb: *later*, *better*, *less*), RBT (superlative adverb: *latest*, *best*, *least*), $*$ (*not*), QL (qualifier: *very*, *too*, *extremely*), and QLP (post-qualifier: *enough*, *indeed*). Two tags stand for parts of speech that have both adverbial and interrogative functions: WQL (*wh*-qualifier: *how*) and WRB (*wh*-adverb: *how*, *when*, *where*).

The Brown tag for prepositions is IN, while particles have the tag RP.

Conjunctions and complementizers

COORDINATING CONJUNCTION
COORDINATION

The remaining important word categories are coordinating and subordinating conjunctions. *Coordinating conjunctions* 'conjoin' or *coordinate* two words or phrases of (usually) the same category:

- husband *and* wife [nouns]

- She bought *or* leased the car. [verbs]
- the green triangle *and* the blue square [noun phrases]
- She bought her car, *but* she also considered leasing it. [sentences]

CLAUSE
SUBORDINATING CONJUNCTION

One function of coordinating conjunctions is to link two sentences (or *clauses*) as shown in the last example. This can also be done by *subordinating conjunctions*. In the examples below, the subordinating conjunction is shown in italics.

(3.28) a. She said *that* he would be late. [proposition]

b. She complained *because* he was late. [reason]

c. I won't wait *if* he is late. [condition]

d. She thanked him *although* he was late. [concession]

e. She left *before* he arrived. [temporal]

COMPLEMENTIZERS

Cases of subordinating conjunctions like *that* in (3.28a) or use of *for* which introduce arguments of the verb are often alternatively regarded as *complementizers*. The difference between coordination and subordination is that, as the terms suggest, coordination joins two sentences as equals whereas subordination attaches a secondary sentence to a primary sentence. The secondary sentence often expresses a proposition, a reason, a condition, a concession or a temporally related event.

Brown tags. The tag for conjunctions is CC. The tag for subordinating conjunctions is CS.

3.2 Phrase Structure

WORD ORDER
PHRASES
SYNTAX

Words do not occur in just any old order. Languages have constraints on *word order*. But it is also the case that the words in a sentence are not just strung together as a sequence of parts of speech, like beads on a necklace. Instead, words are organized into *phrases*, groupings of words that are clumped as a unit. *Syntax* is the study of the regularities and constraints of word order and phrase structure.

CONSTITUENT

One fundamental idea is that certain groupings of words behave as *constituents*. Constituents can be detected by their being able to occur

in various positions, and showing uniform syntactic possibilities for expansion. The examples in (3.29) and (3.30) show evidence from positioning and phrasal expansion for a constituent that groups nouns and their modifiers:

(3.29) a. I put *the bagels* in the freezer.

b. *The bagels*, I put in the freezer.

c. I put in the fridge *the bagels* (that John had given me)

(3.30) $$\left\{\begin{array}{c}\text{She}\\ \text{the woman}\\ \text{the tall woman}\\ \text{the very tall woman}\\ \text{the tall woman with sad eyes}\\ \ldots\end{array}\right\} \text{ saw } \left\{\begin{array}{c}\text{him}\\ \text{the man}\\ \text{the short man}\\ \text{the very short man}\\ \text{the short man with red hair}\\ \ldots\end{array}\right\}.$$

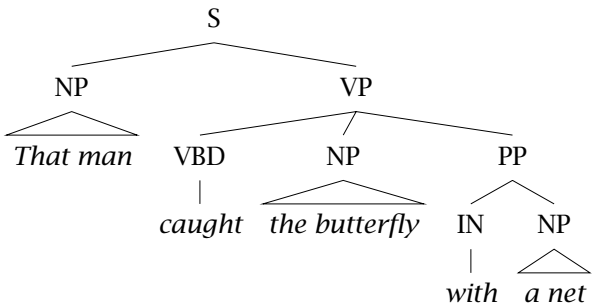
PARADIGMATIC RELATIONSHIP

This is the notion of a *paradigmatic relationship* in Saussurean linguistics. All elements that can be replaced for each other in a certain syntactic position (like the noun phrase constituent above) are members of one *paradigm*. In contrast, two words bear a *syntagmatic relationship* if they can form a phrase (or *syntagma*) like *sewed clothes* or *sewed a dress*. An important class of syntagmatically related words are *collocations* (chapter 5).

PARADIGM
SYNTAGMATIC RELATIONSHIP
SYNTAGMA
COLLOCATIONS

In this section we will briefly mention some of the major phrase types, and then introduce techniques linguists use to model phrase structure. The upshot will be to suggest that English sentences typically have an overall phrase structure that looks as follows:

(3.31)



A whole sentence is given the category S. A sentence normally rewrites as a subject noun phrase and a verb phrase.

NOUN PHRASE

Noun phrases. A noun is usually embedded in a *noun phrase* (NP), a syntactic unit of the sentence in which information about the noun is gathered. The noun is the *head* of the noun phrase, the central constituent that determines the syntactic character of the phrase. Noun phrases are usually the *arguments* of verbs, the participants in the action, activity or state described by the verb. Noun phrases normally consist of an optional determiner, zero or more adjective phrases, a noun head, and then perhaps some post-modifiers, such as prepositional phrases or clausal modifiers, with the constituents appearing in that order. Clausal modifiers of nouns are referred to as *relative clauses*. Here is a large noun phrase that indicates many of these possibilities:

HEAD

RELATIVE CLAUSES

(3.32) The homeless old man in the park that I tried to help yesterday

PREPOSITIONAL PHRASES

Prepositional phrases. *Prepositional phrases* (PPs) are headed by a preposition and contain a noun phrase complement. They can appear within all the other major phrase types. They are particularly common in noun phrases and verb phrases where they usually express spatial and temporal locations and other attributes.

VERB PHRASE

Verb phrases. Analogous to the way nouns head noun phrases, the verb is the head of the *verb phrase* (VP). In general, the verb phrase organizes all elements of the sentence that depend syntactically on the verb (except that in most syntactic theories the verb phrase does not contain the subject noun phrase). Some examples of verb phrases appear in (3.33):

(3.33) a. *Getting to school on time* was a struggle.

b. He *was trying to keep his temper*.

c. That woman *quickly showed me the way to hide*.

ADJECTIVE PHRASES

Adjective phrases. Complex *adjective phrases* (APs) are less common, but encompass examples like the phrases shown in bold in these sentences: *She is* ***very sure of herself***; *He seemed a man who was* ***quite certain to succeed***.

3.2.1 Phrase structure grammars

A syntactic analysis of a sentence tells us how to determine the meaning of the sentence from the meaning of the words. For example, it will tell us who does what to whom in the event described in a sentence. Compare:

(3.34) Mary gave Peter a book.

(3.35) Peter gave Mary a book.

Sentences (3.34) and (3.35) use the same words, but have different meanings. In the first sentence, the book is transferred from Mary to Peter, in the second from Peter to Mary. It is the word order that allows us to infer who did what to whom.

Some languages like Latin or Russian permit many different ways of ordering the words in a sentence without a change in meaning, and instead use case markings to indicate who did what to whom. This type of language is called a *free word order* language, meaning that word order isn't used to indicate who the doer is – word order is then usually used mainly to indicate discourse structure. Other languages such as English are more restrictive in the extent to which words can move around in a sentence. In English, the basic word order is Subject – Verb – Object:

FREE WORD ORDER

(3.36) *The children* (subject) *should* (auxiliary verb) eat *spinach* (object).

In general, this order is modified only to express particular 'mood' categories. In *interrogatives* (or questions), the subject and first auxiliary verb are *inverted*:

INTERROGATIVES
INVERTED

(3.37) *Should* (auxiliary verb) *the children* (subject) eat *spinach* (object)?

If the statement would involve no auxiliary, a form of *do* appears in the initial position (*Did he cry?*). In *imperatives* (commands or requests), there is no subject (it is inferred to be the person who is addressed):

IMPERATIVES

(3.38) Eat spinach!

Basic sentences are called *declaratives* when contrasted with interrogatives and imperatives.

DECLARATIVES

The regularities of word order are often captured by means of *rewrite rules*. A rewrite rule has the form 'category → category*' and states that the symbol on the left side can be rewritten as the sequence of symbols on the right side. To produce a sentence of the language, we start with the *start symbol* 'S' (for sentence). Here is a simple set of rewrite rules:

REWRITE RULES
START SYMBOL

(3.39)
$$
\begin{array}{lcl@{\qquad}lcl}
\text{S} & \rightarrow & \text{NP VP} & \text{AT} & \rightarrow & \textit{the} \\
\text{NP} & \rightarrow & \left\{\begin{array}{l}\text{AT NNS}\\ \text{AT NN}\\ \text{NP PP}\end{array}\right\} & \text{NNS} & \rightarrow & \left\{\begin{array}{l}\textit{children}\\ \textit{students}\\ \textit{mountains}\end{array}\right\} \\
\text{VP} & \rightarrow & \left\{\begin{array}{l}\text{VP PP}\\ \text{VBD}\\ \text{VBD NP}\end{array}\right\} & \text{VBD} & \rightarrow & \left\{\begin{array}{l}\textit{slept}\\ \textit{ate}\\ \textit{saw}\end{array}\right\} \\
\text{P} & \rightarrow & \text{IN NP} & \text{IN} & \rightarrow & \left\{\begin{array}{l}\textit{in}\\ \textit{of}\end{array}\right\} \\
 & & & \text{NN} & \rightarrow & \textit{cake}
\end{array}
$$

The rules on the righthand side rewrite one of the syntactic categories (or part of speech symbols) introduced in the previous sections into a word of the corresponding category. This part of the grammar is often separated off as the *lexicon*. The nature of these rules is that a certain syntactic category can be rewritten as one or more other syntactic categories or words. The possibilities for rewriting depend solely on the category, and not on any surrounding context, so such phrase structure grammars are commonly referred to as *context-free grammars*.

LEXICON

CONTEXT-FREE GRAMMARS

With these rules, we can derive sentences. Derivations (3.40) and (3.41) are two simple examples.

(3.40)
S
→ NP VP
→ AT NNS VBD
→ *The children slept*

(3.41)
S
→ NP VP
→ AT NNS VBD NP
→ AT NNS VBD AT NN
→ *The children ate the cake*

The more intuitive way to represent phrase structure is as a tree. We refer to the leaf nodes of the tree as *terminal nodes* and to internal nodes as *nonterminal nodes*. In such a tree each nonterminal node and its immediate daughters, otherwise known as a *local tree* corresponds to the application of a rewrite rule. The order of daughters generates the word order of the sentence, and the tree has a single root node, which is the start symbol of the grammar. Trees (3.42) and (3.43) correspond to deriva-

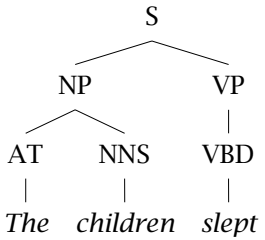
TERMINAL NODES

NONTERMINAL NODES

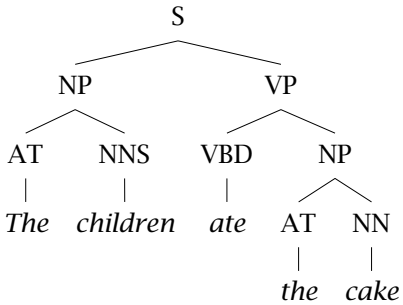
LOCAL TREE

tions (3.40) and (3.41). Each node in the tree shows something that we are hypothesising to be a *constituent*.

(3.42)



(3.43)



BRACKETING A third and final way to show constituency is via a *(labeled) bracketing*. Sets of brackets delimit constituents and may be labeled to show the category of the nonterminal nodes. The labeled bracketing for (3.43) is (3.44):

(3.44) [S [NP [AT *The*] [NNS *children*]] [VP[VBD *ate*] [NP [AT *the*] [NN *cake*]]]]

RECURSIVITY A property of most formalizations of natural language syntax in terms of rewrite rules is *recursivity:* the fact that there are constellations in which rewrite rules can be applied an arbitrary number of times. In our example grammar, a PP contains an NP which can in turn contain another PP. Thus we can get recursive expansions as in the example in figure 3.1. Here, the sequence of prepositional phrases is generated by multiple application of the rewrite rule cycle "NP → NP PP; PP → IN NP." The derivation applies the cycle twice, but we could apply it three, four, or a hundred times.

Recursivity makes it possible for a single nonterminal symbol like VP or NP to be expanded to a large number of words. (For example, in figure 3.1 the symbol VP is expanded to nine words: *ate the cake of the children in*

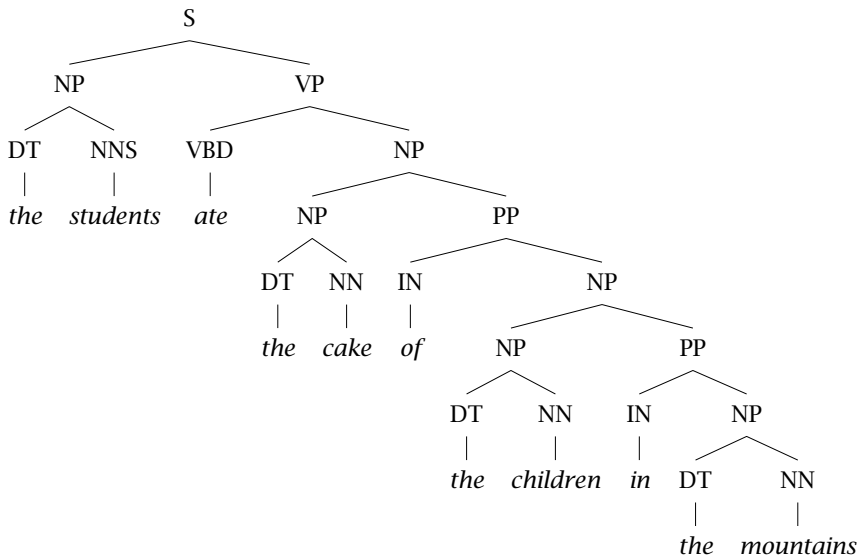


Figure 3.1 An example of recursive phrase structure expansion.

the mountains.) One consequence is that two words that were generated by a common rewrite rule and are syntactically linked can become separated by intervening words as the derivation of a sentence proceeds.

NON-LOCAL DEPENDENCIES

These types of phenomena are called *non-local dependencies* because two words can be syntactically dependent even though they occur far apart in a sentence.

SUBJECT-VERB AGREEMENT

One example of a dependency that can be non-local is *subject-verb agreement*, the fact that the subject and verb of a sentence agree in number and person. We have *She walks*, *He walks*, *It walks* versus *I walk*, *You walk*, *We walk*, *They walk*. That is, the verb has the ending *-s* indicating third person singular if and only if the subject is in the third person singular. Subject and verb agree even if other words and phrases intervene as in the following example.

(3.45) The **women** who found the wallet **were** given a reward.

If we looked only at immediate neighbors it would seem that we would have to say *the wallet was*. Only a complete syntactic analysis of the sentence reveals that *The women* is the subject and the form of *to be* has to be in the plural.

LONG-DISTANCE DEPENDENCIES
WH-EXTRACTION

Another important case of non-local dependencies is the class known as *long-distance dependencies*, such as *wh-extraction*.[2] The name is based on the theory that phrases such as *which book* in (3.46b) are moved (or extracted) from an underlying position (after the verb as in (3.46a)) to their "surface" position (the beginning of the sentence as in (3.46b)).

(3.46) a. Should Peter buy *a book*?

b. *Which book* should Peter buy?

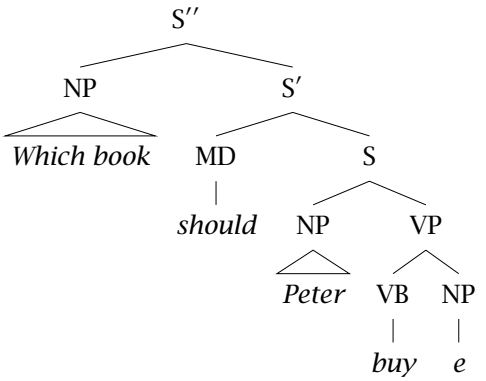
Without making any commitment to such a movement theory, it is clear that we have to recognize a long distance dependency between *buy* and *which book*. Otherwise we would not be able to tell that *book* is an argument of *buy*.

▼ Non-local phenomena are a challenge for some Statistical NLP approaches like *n*-grams that model local dependencies. An *n*-gram model would predict that the word after *wallet* in (3.45) is *was*, not *were*. These issues are further discussed at the beginning of chapter 11.

EMPTY NODES

A final feature of many versions of phrase structure grammar is empty nodes. *Empty nodes* occur when a nonterminal may be rewritten as nothing. For example, noting that one can also say *Eat the cake!* without a subject NP, one might suggest adding a rule NP → ∅. An NP nonterminal is then allowed to be rewritten as nothing. This is often represented by putting a ∅ or an *e* under the node in the tree. Using this notation, the tree in (3.46b) could be given the structure in (3.47):

(3.47)



2. In the speech literature, the term 'long-distance dependencies' regularly refers to anything beyond the range of a trigram model. We have termed such effects 'non-local dependencies,' and have reserved the term 'long-distance dependencies' for its usual linguistic meaning of a dependency that appears to be able to cross any number of nodes in a phrase structure tree.

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*.

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 (3.49), *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):

SEMANTIC ROLES
AGENT
PATIENT

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 languages 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 of whiskey* 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 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:

ADJUNCTS

(3.57) 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 rain tomorrow*. The *that* can usually be omitted: She thinks *it will rain tomorrow*.
- **Question-form clauses.** She is wondering *why it is raining 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 raining in August* consists of a whole sentence *it is raining 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 woman 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 Paddington*.
- **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 strong selectional preferences sound odd:

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

b. I eat *philosophy* every day.

▼ Selectional preferences are further discussed in section 8.4.

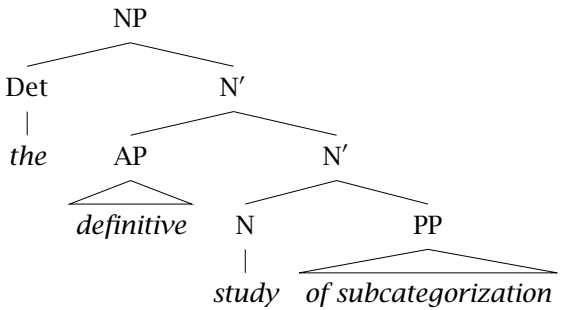
3.2.3 X′ theory

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

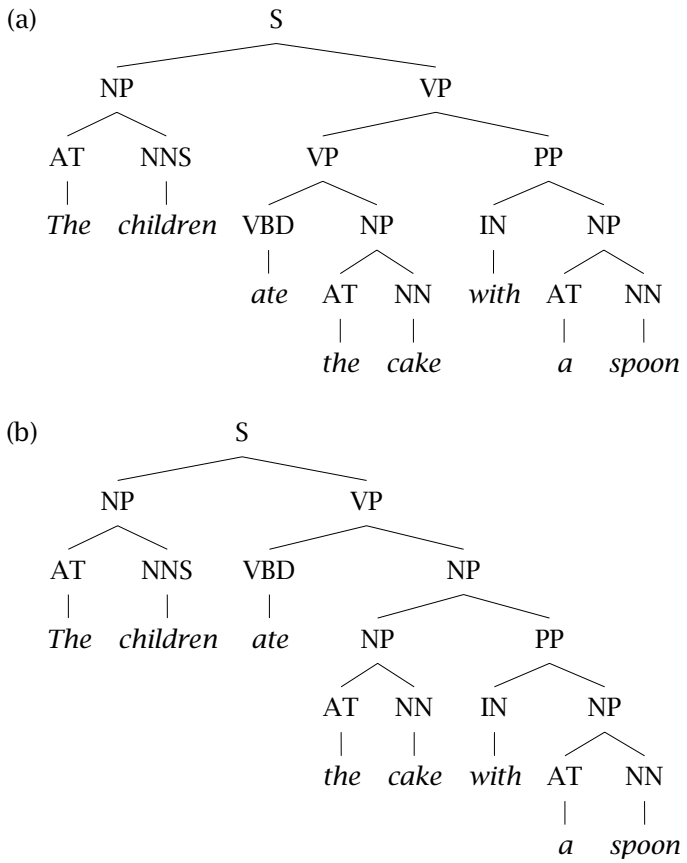


Figure 3.2 An example of a prepositional phrase attachment ambiguity.

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 barn.* 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,

HYPERNYMY
HYPONYMY
HYPERONYM

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: *cat* 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

Antonyms are words with opposite meanings: *hot* and *cold* or *long* and *short*.

MERONYMY

The part-whole relationship is called *meronymy*. The word *tire* is a meronym of *car* and *leaf* is a meronym of *tree*.

HOLONYM

The whole corresponding to a part is called a *holonym*.

SYNONYMS

Synonyms are words with the same meaning (or very similar meaning): *car* and *automobile* are synonyms.

HOMONYMS

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 *bank* ('river bank' and 'financial institution').

SENSES
POLYSEME

If a word's meanings (or *senses*) are related, we call it a *polyseme*. The word *branch* is polysemous because its senses ('natural subdivision of a plant' and 'a separate but dependent part of a central organization') are related.

AMBIGUITY

Lexical *ambiguity* can refer to both homonymy and polysemy.

HOMOPHONY

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 *bass* for a species of fish and *bass* for a low-pitched sound are homonyms, but they are not homophones.

▼ 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

COMPOSITIONALITY

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:

(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.

SOCIOLINGUISTICS
HISTORICAL LINGUISTICS

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.

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.