LORI LEVIN AND DAVID R. MORTENSEN

HUMAN LANGUAGE FOR ARTIFICIAL INTELLIGENCE

Contents

1 Introduction

This is a course about human language as it relates to artificial intelligence applications. The goal of the course is to introduce students to the several levels of linguistic structure that impinge upon human language technologies and related AI endeavors. This course is necessarily introductory—it is a survey that can only scratch the surface of what is known about each topic—but it is also intended to be demanding. Unlike most introductions to linguistics intended for humanities undergraduates, or as undergraduate general education courses, this course will be taught with expectations similar to those of an introductory course in a natural science (biology or physics, for example). Students will be expected to show no shyness towards scientific, formal, or quantitative reasoning. They will also be expected to have considerable familiarity with computers, both in practice and in theory. For example, students will find aspects of the course challenging if they have no experience with the Unix/Linux shell or the notion of an algorithm.

1.1 An Introductory Exercise

The original Turing Test asks whether a conversational agent, communicating via text, can trick a human interlocutor into believing that it is human. Imagine a more difficult version of this test:

- Rather than a conversational agent, the computational interlocutor is a humanoid robot.
- The robot communicates through speech (or, alternatively, through signed language).
- The robot passes the test if it can fool humans that interact with it into believing that it is human.

This is a difficult scenario. Of course, there are many conditions that would have to be satisfied in order for the robot to pass the test. It must look like a human, move like a human, display "body language" like that of a human, and so on. However, our concern is specifically with what it must know about human language—one of the most difficult aspects of this test. What must a robot know about language in order to fool humans into believing that it is also human?

The robot would have to be able to engage language at various levels (sounds, grammar, meaning, coherence of dialogues). It would have to be able to manipulate units such as words and sentences with various degrees of internal complexity. We do not mean to imply that each of these levels must be programmed individually; this may or may not be the case. However, even if our robot is a perfect end-to-end system, it must necessarily encode—somewhere in its architecture or memory—each of these levels. In deep neural networks for speech or image classification, it has been found that lower layers in a network typically encode more concrete, low-level units. Higher layers encode more abstract units. We have every reason to believe that an end-to-end language system for a robot would follow roughly the same pattern. The more concrete levels or representation, which we will discuss first, would belong to the lower layers and the more levels of representation are more abstract (or are farther removed from fundamental units) would be encoded in higher layers. For an end-to-end system that uses some approach other than deep learning, a similar condition is likely to hold: somewhere, somehow, this knowledge is present in the system.

But what knowledge?

The robot would have to be able to interact via speech (or, alternatively, sign). This involves being able produce speech but also to understand speech. At a very basic level, the robot must be able to segment acoustic signals and assign these segments to abstract categories that smooth over their idiosyncratic and contextual variation. It must also be able to take a string of such categories and render them as an acoustic signal. These problems belong to the sphere of PHONETICS.

However, when it comes to sound, these low-level segments are not the most abstract units required to produce and understand speech. There are higher-level representations (phonemic representations) that factor out non-contrastive information and ensure that—in different contexts—the same meaningful unit (morpheme) always has the same representation. Finally, there are constraints on how phonemes can combine, which the robot must know about in order to correctly manipulate these representations. This field is called PHONOLOGY.

Words, too, have internal structure. To a first approximation, they are made by concatenating morphemes. However, the knowledge necessary to predict which combinations of morphemes are legitimate words in a language and which are illicit is quite involved. This field is called MORPHOLOGY.

In language, however, words do not combine without constraint. Our robot must know about SYNTAX, the constraints on how words combine with one another to form higher-level structures called PHRASES and SENTENCES. This structure is a way of encoding SEMANTICS, including notions like possession, causation, and "who did what to whom?"

Structure does not end at the sentence, though. Our robot must know how to string together sentences into meaningful discourses and carry on

sensible conversations. Just knowing what sentences are part of the language in question will not be sufficient to do this becausex DISCOURSE has a structure all its own.

Finally, language does not exist in a vacuum. Humans are very sensitive to social cues and social identity; our robot must be the same. The field in the language sciences that studies social meaning and the role of language in society is called SOCIOLINGUISTICS.

1.2 What is linguistics?

Linguistics is the study of human language as a naturally occurring phenomenon. It is not about "correct" language or "talented" language. It is about language as it is used by normal people every day, including those who use signed languages. Linguistics also pertains to language disorders, trying to understand what the disorder tells about language in the human brain and in human physiology.

In this class, you will learn to understand the terminology and notation that linguists use to describe languages and talk about languages. You will learn to analyze, describe, and talk about languages that you do not speak.

Like the life sciences such as biology, the human language sciences cover behavior as well as structure. Both at behavioral level and structural levels, the objective study of human language involves becoming aware of what is "folk science" (how untrained people tend to think about a science) and even "school linguistics" (the terminology that language teachers use) and replacing them with rigorously defined concepts and methods.

1.2.1 How many languages are there and how do we know?

Linguistics is about all human languages that exist now and have ever existed, so let's get an idea of the scope of the subject matter. There are about 7000 human languages currently in existence. There are also many extinct languages for which recordings or written material remains. But what is a language as opposed to a dialect or a slang? You may have heard the humorous statement that a language is a dialect with an army. There is some truth to the statement. Many languages are mutually intelligible and seem only to be called different languages because they are spoken in different countries or use a different writing system. Examples are Serbian and Croatian or Hindi and Urdu. Azerbajani also might not be a language if it didn't have its own country.

However, linguists have other ways of distinguishing languages, aside from observing national and ethnic boundaries. Ethnologue and Glottolog (insert links) are comprehensive lists of the world's languages with a unique ISO code for each language. In order to compile such comprehensive reference works, field linguists visit every remote inhabited place, or find people from those places who have moved to cities. The linguists ask people to translate word

lists including words that exist in most or all languages like words for body parts, stones, dirt, and parts of plants and animals. If the word lists of two villages are more than thirty percent different, the linguists label the languages as distinct from each other.

The process of counting languages is not precise or stable. Globalization, war, and shifting economic trends are leading to a rapid extinction of languages, and it is estimated that the world could lose most of its distinct languages by the end of the twenty-first century. It is difficult to measure the exact day when a language dies. It could be measured by the day the last native speaker dies, but there may be many partial native speakers, a community of people who learn the language in adulthood as a second language, and the community may be in the process of revitalizing the language by teaching it to children. At the same time, new languages continue to arise. The BBC started reporting news in Nigerian Pidgin in 2017. Nigerian Pidgin is spoken by tens of millions of people in at least four countries, and certainly did not arise suddenly in 2017. It is not possible to name a day or year when children began speaking Nigerian Pidgin as one of their first languages or the day when informal adaptations English became a conventionalized system of communication.

1.2.2 A case study in counting languages: African American Vernacular English

As a case study in the difficulty of counting how many human languages there are, we will consider African American Vernacular English (AAVE). AAVE is spoken in the United States. AAVE speakers and Standard American English (SAE) speakers can usually understand each other. However, there are grammatical differences. For example, an SAE speaker may say *She has been doing her homework* whereas an AAVE speaker might say *She been doing her homework*, both might convey that she has been doing her homework for a long time. There are also differences in pronounciation. For example, AAVE speakers might not pronounce a final *r* in a word like *car*, and when a word ends in two consonants, they might drop the last consonant so that *past* and *passed* sound like *pass*. To complicate matters, like all bilingual people, AAVE speakers switch between their languages, but SAE speakers may perceive AAVE speakers as being inconsistent when they mix AAVE and SAE. For decades, SAE speakers considered AAVE to be a slang or sloppy way of speaking English. In the 1970's however, AAVE was recognized as a system within English, meaning that the differences between AAVE and SAE are not random, but are systematic rules of grammar and pronunciation. Schools still struggle with the question of how to honor AAVE as the home language of many students in a setting where most education is in SAE.

1.3 Is linguistics a science?

Most linguistics departments in the United States were formed between 1960 and 1980 as spin-offs of humanities departments such as anthropology and philosophy departments. Beginning in the 1980's linguistics in the US was also considered to be a cognitive science, along with psychology, philosophy, and artificial intelligence. Now, many linguistics departments are being re-classified as natural sciences or data sciences. In many places, linguistics is re-inventing itself to be *the human language sciences* in order to be inclusive of linguists working in industry and linguists doing applied work in speech and language disorders. So, is linguistics a science?

We will discuss three ways in which linguistics is like a science, summarizing from a YouTube video by Martin Hilpert () . The first way in which linguistics is like a science is that it uses the scientific method. Scientists start with a theoretical framework, within which they formulate a hypothesis. They design an experiment to test the hypothesis, collect results, and make observations. Depending on the outcome of the observations, scientists may revise the theoretical framework or change the hypothesis. Experiments such as this are common in psycholinguistics, which is concerned with how human language is processed in the brain and how the human brain produces the structures of human language. An example of a theoretical construct in psycholinguistics is the *mental lexicon*. A lexicon is a list of words. In the mental lexicon theory, the words are not stored in alphabetic order as they are in a dictionary, but in networks of related words such as *doctor* and *hospital*. A specific hypothesis in mental lexicon theory pertains to *reaction time* in a *word-decision* task. In a word-decision task, experimental subjects are asked to decide whether something is a word or not in their language. For example, if English-speaking experimental subjects are shown the word *doctor*, they should respond that it is a word, but if shown *roctod*, they should say that it is not a word. The experimental subjects respond by pressing buttons, and an experimental apparatus records the number of milliseconds it takes them to press the button (reaction time). The hypothesis is that if the mental lexicon stores related words near each other in the brain, or related words *activate* each other, then reaction time to decide whether *hospital* is an English word should be shorter when the word *hospital* is seen shortly after the word *doctor* has been seen, and the reaction time to decide whether *hospital* is an English word should be longer when no related words have been seen recently. This turns out to be true. Words with related meanings activate each other, resulting in shorter reaction times in word-decision tasks.

Another way that linguistics is like a science, according to Hilpert, is that linguists are on a quest for the laws of nature. One example of a "law" of language is Zipf's law. Zipf's Law: the frequency of a word is proportional to 1/the-word's-frequency rank. For example, the second most common word in English is "of", so the frequency rank of "of" is two. Therefore "of" is half

as common as the most common word, "the". The third most common word, "and" is one third as common as the most common word, "the". Zipf's law holds in all languages. Figure **??** from () illustrates Zipf's Law in English.

The third way in which linguistics is like a science is that some linguists use big data such as the ten billion word News on the Web corpus, the British National Corpus, or the American National Corpus.

To sum up, linguistics is like a science in that it uses the scientific method, seeks laws of nature, and uses big data.

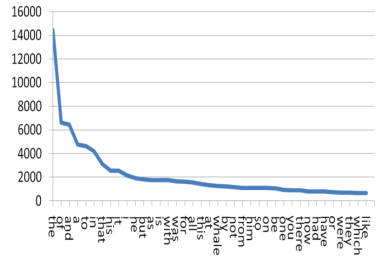


Figure 1.1: Zipf's Law in English

1.4 *Symbolic Systems: how linguistics is different from some other sciences*

If you have studied biology, chemistry, or physics, you might be used to laws that are formulated in continuous numerical systems. In contrast, in linguistics, laws are often formulated in terms of symbolic systems. Symbolic systems include some things you may have heard of like formal logic, programming languages, finite state machines, and context-free grammars. Since you will be using symbolic systems extensively in this course, we will present an example here as a preview of what is to come.

1.4.1 *The "law" of English reflexive pronouns represented as a symbolic system*

In order to examine the law of English reflexive pronouns, we will briefly introduce notions such as *grammaticality* and *constituent structure*, which will be treated in more detail later in the course. We will start by listing some English pronouns. The columns indicate the grammatical roles of the pronouns in sentences like *I will go* (subject), *It bothers me* (object), *This is my book* (possessor), *This book is mine* (possessive predicate), *I saw myself* (reflexive pronoun).

I	me	my	mine	myself
we	us	our	ours	ourselves
you	you	your	yours	yourself
you	you	your	yours	yourselves
it	it	its	its	itself
she	her	her	hers	herself
he	him	his	his	himself
they	them	their	theirs	themselves

We will be concerned here with the pronouns in the last column, which are called *reflexive pronouns*, for reasons that will become apparent.

2 Articulation and Acoustics in Speech

2.1 Introduction

In mid-2018, a craze swept the Internet—an audio recording that pitted parent against child, brother against sister, and friend against friend. Like an earlier debate that involved the color of a dress (gold, blue, or white?) this debate involved a matter of perception. Hearing the recording, some heard *Laurel* while others heard *Yanny*. On the page, these two words seem quite dissimilar. As with the ambiguously-colored dress, there are good perceptual reasons of this confusion. For the auditory case, these are grounded in the acoustic properties of speech, or acoustic phonetics.

Language encodes meaning. This coding system can be analyzed at many levels, but the most fundamental of these coincides with the modality through which language is transmitted. The most common modality is speech, though signed languages are also widely used (particularly in deaf communities) and written language is ubiquitous too. But speech is so common, and so important to so many people, that we will devote a whole unit to phonetics, the study of the physical aspects of speech[1].

A basic understanding of speech requires knowing something about how speech sounds are produced. This is called ARTICULATORY PHONETICS[2]. This will serve as a foundation for understanding the International Phonetic Alphabet, a standard way of representing speech sounds. More central, for language technologies and artificial intelligence, is ACOUSTIC PHONETICS[3]. AUTOMATIC SPEECH RECOGNITION[4] and SPEECH SYNTHESIS[5] both concern the relationship between speech acoustics and written representations of language. As such, acoustic phonetics is highly relevant to these tasks. Finally, no discussion of phonetics would be complete without an exploration of PROSODY[6], including INTONATION[7], and how it can encode non-lexical meaning[8].

Figure 2.1: The Dress

[1] Sign phonetics is a field of study as well (dealing with the production and perception of the visual behaviors that make up sign languages).

[2] Articulatory phonetics is the study of the physiological mechanisms by which humans produce speech sounds.

[3] Acoustic phonetics is the study of the acoustic signals produced by speech. It is closely allied to auditory phonetics, which explores how speech sounds are sensed and perceived by human listeners.

[4] The task in which the sound waves from speech are converted to meaningful linguistic representations by a computer.

[5] The production of synthetic speech.

[6] Aspects of speech that are "spread out" over multiple segments (consonants and vowels) like tone, stress, and intonation

[7] Intonation is the use of pitch to express information other than distinctions between words.

[8] Meaning that does not come from words.

2.2 Articulatory phonetics

Language users are largely unaware of how systematic the mechanisms by which they speak really are. They are unaware that there are a small number of relatively simple parameters that can characterize all of the speech

sounds of the world's languages. This is because speech production is highly routinized[9] and, when speaking our first language at least, we execute "programs" for whole words without any thought for the individual parts—the consonants, vowels, and prosodies that combine to realize those words.

[9] That is, they are automatic routines.

In fact, most speech sounds can be characterized on just four dimensions: what the source of airflow is, where the VOCAL TRACT is most constricted, how these constrictions are made (including their extent), and whether to vocal folds are vibrating contemporaneously with the constriction[10]. Most speech sounds (and all speech sounds in English) share a common source of airflow—a common AIRSTREAM MECHANISM—characterized as PULMONIC EGRESSIVE. This means that air flows out from the lungs. Other airstream mechanisms include GLOTTALIC EGRESSIVE (ejectives), GLOTTALIC INGRESSIVE (implosives), and VELARIC INGRESSIVE (clicks). Because pulmonic egressive sounds are so much more common than the others, and because the other parameters are the same regardless of airstream mechanism, we will concentrate on the last three parameters: PLACE, MANNER, and VOICING.

[10] Really, voicing as a parameter refers to the timing of the onset of voicing relative to the timing of the constriction, but this approximation is good enough to start.

The signs in sign languages can also be represented in terms of a small finite number of parameters, but these are—of course—different from those used in spoken languages. For example HAND SHAPE is one phonetic dimension of sign. Compared to airstream mechanism, place, manner, and voicing, there are a relatively large number of possible settings for this parameter (especially if you look across all the sign languages in the world). Because research on sign languages has a shorter history than research in spoken languages, less is known about sign phonetics and characterizing the range of possibilities is still very much an open research area.

2.2.1 *Place, manner, and voicing*

In order to understand what phoneticians mean by place, it is necessary to learn something about the anatomy of the vocal tract. The vocal tract is a tube that extends from the lips to the glottis (or vocal folds), but it is a rather articulated tube. An illustration of the vocal tract, and its landmarks, is given in Figure 2.2. This is an example of what is called a MIDSAGITTAL SECTION. It is the result of taking a human head and splitting it in half. Phoneticians do this all the time, for fun.

The articulatory landmarks correspond to PLACES OF ARTICULATION. Memorizing these, or being familiar with them, will help you interpret phonetic descriptions. They are also important to phonology (as are manner of articulation and voicing). Following is a list of the most important places of articulation for consonants:

- **Labial** Constriction made with the lips
  - **Bilabial** Constriction made by bringing the two lips together, like the ⟨b⟩ in *ban*

Figure 2.2: Anatomy of the vocal tract

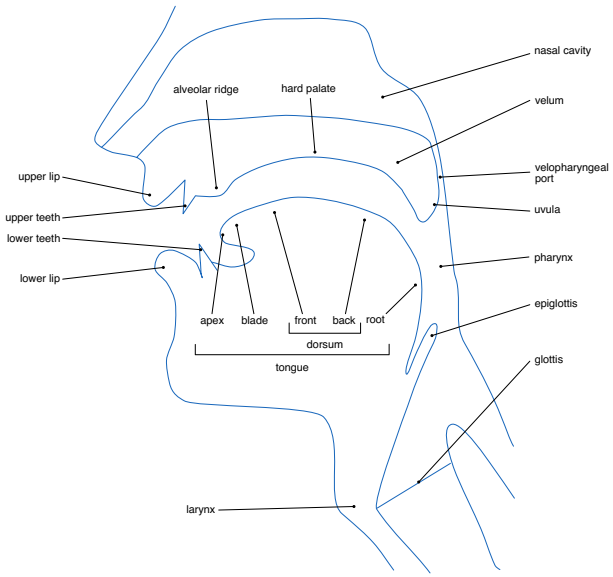
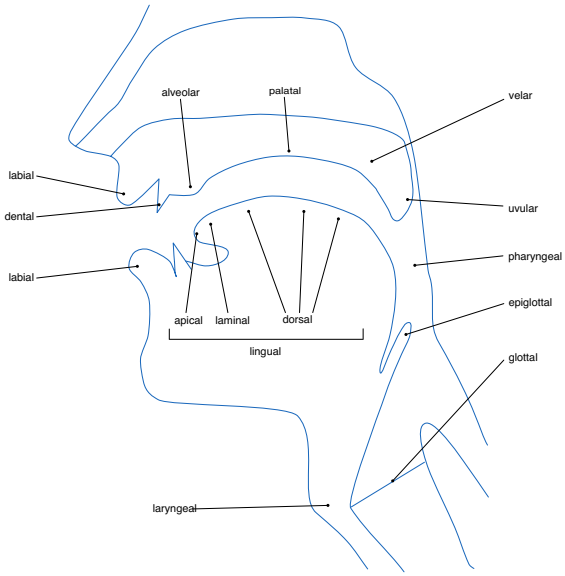


Figure 2.3: Adjectives for places of articulation anchored to the vocal tract anatomy



  - **Labiodental** Constriction by bringing the bottom lip together with the upper teeth, like the ⟨f⟩ in *fine*

- **Coronal** Constrictions made with the tip or blade of the tongue
  - **Interdental** Constriction made by bringing the front of the tongue into the proximity of the upper teeth, like the ⟨th⟩ in *think*
  - **Dental** Constriction made by bringing the blade (or less often, the apex) of the tongue together with the upper teeth
  - **Alveolar** Constriction made by bringing the apex (or less often, the blade) of the tongue together with the alveolar ridge, like the ⟨s⟩ in American English *suspect*
  - **Post-alveolar** Constriction made by bringing the apex or blade of the tongue together with region directly behind the alveolar ridge, like the ⟨j⟩ in *judge*
  - **Retroflex** Constriction made by raising the tongue so that the tip is vertical or curling the tongue back so that the bottom of the apex is touching the hard palate, like ⟨t⟩, ⟨d⟩, ⟨s⟩, ⟨l⟩ and ⟨r⟩ in Indian English
- **Dorsal** Constrictions made with the body of the tongue and the palate (the roof of the mouth)
  - **Palatal** Constriction made by bringing the body of the tongue together with the hard palate
  - **Velar** Constriction made by bringing the body of the tongue together with the soft palate (velum), like the ⟨g⟩ in *glamorous*
  - **Uvular** Constriction ma1de by bringing the body of the tongue together with the uvula
- **Pharyngeal** Constriction made by bringing the base of the tongue toward the back wall of the pharynx
- **Laryngeal/glottal** Constriction made with the vocal folds, like the ⟨h⟩ in *hellish*

It may appear from this list that place of articulation for consonants is a one-dimensional affair. In fact, there can be secondary articulations. For example, the most common pronunciation of American English ⟨r⟩ actually has two places of articulation: it is produced with rounded lips (bilabial) and is produced with the front of the tongue "bunched" and advanced towards the alveolar ridge.

Even more explicitly multidimensional is vowel place. Vowels are usually described in three dimensions: height, backness, and roundness. Height refers to the position of the body of the tongue. The vowel in *meet* is produced with the tongue body relatively height. Compare this with (General American English) *mat*, where the vowel is low (produced with the tongue

Place	Examples
BILABIAL	**p**in, **b**in, **m**in
LABIODENTAL	**f**in, **v**im
ALVEOLAR	**t**in, **d**in, **n**ine, **s**in, **z**ip
PALATO-ALVEOLAR	**ch**in, **j**inn, **sh**in, a**z**ure
VELAR	**k**in, **g**ain
GLOTTAL	**h**it

Table 2.1: American English examples of place of articulation

body relatively distant from the roof of the mouth. Now compare the vowels is *meet* and *moot*[11]: the differences between the two vowels are not of the same quality. First, when producing *moot*, the corners of the mouth are drawn together and the lips protrude somewhat. When producing *meet*, the corners of the mouth are spread[12]. This is not the only difference between the *meet* and *moot* vowels, however.

Pay attention to the position of your tongue—in *meet*, the tongue body is advanced, with the closest constriction being between the tongue and the hard palate. In *moot*, by contrast, the closest constriction is between the dorsum of the tongue and the velum.

In most languages, there are degrees of height and degrees of backness, but most languages only have one degree of lip rounding. These dimensions define a three-dimension space called the VOWEL SPACE. Because of the limitations of human anatomy, the front-back dimension is compressed progressively one moves lower in the vowel space. This yields the shape commonly known as the VOWEL QUADRILATERAL, as shown in Figure 2.4. This illustration only covers the height and backness dimensions. With the rounding dimension added, the space looks like Figure 2.5.

Place of articulation is one major parameter of phonetic description. Another is manner of articulation. We have already invoked the largest difference in manner—that between consonants and vowels. Vowels are produced with a relatively open vocal tract (with minimal constriction). Consonants are produced with some constraint on the flow of air through the vocal tract. As you might expect, this isn't a categorical distinction and there are intermediate sounds—SEMIVOWELS or GLIDES. English has two glides, the initials sounds in *yes* and *Wes*, but other languages have more. They are usually considered to be consonants, but consonants with vowel characteristics. While it is the case that there are manner distinctions among vowels, these are marginal and are usually conflated with height. By contrast, there are many different manners of consonants. They differ in how tightly the vocal tract is constricted, what the nature of the constriction is, and whether the velopharyngeal port is open. The primary manners of articulation are listed below:

- **plosives** or **oral stops** Characterized by the complete obstruction of the vocal tract and the closure of the velopharyngeal port; like the ⟨p⟩ in *porpoise*
- **nasal stops** or NASALS Characterized by the complete obstruction of the vocal tract but with the velopharyngeal port open; like the ⟨m⟩ in *muggle*
- **trills** Produced with a "loose" closure so that the passage of air produces an oscillation
- **flap** or **tap** essentially a momentary plosive produced when an ACTIVE ARTICULATOR strikes a PASSIVE ARTICULATOR; like the

[11] Do so without thinking of ents

[12] This is why photographers in the English-speaking world often request that their subjects say "Cheese"—producing the ⟨ee⟩ vowel forces the mouth into something like a smile.

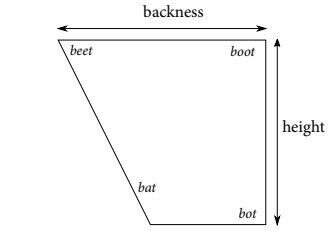


Figure 2.4: The vowel quadrilateral

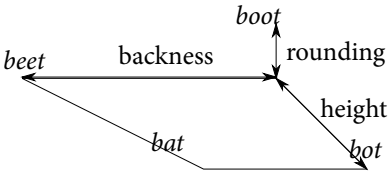


Figure 2.5: Vowel place in three dimensions

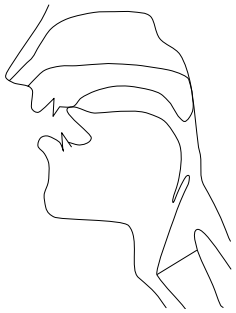


Figure 2.6: Midsagittal section of a person producing [t], a plosive

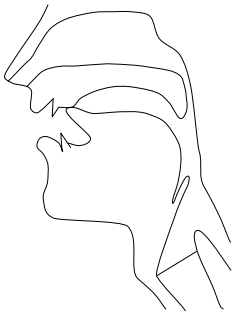


Figure 2.7: Midsagittal section of a person producing [n], a nasal stop

- **fricatives** Characterized by a tight constriction that produced turbulence when air is blown through it; like the ⟨s⟩ in *slither*
- **lateral fricative** A special kind of fricative in which the opening is on one or both sides of the tongue; common in exotic languages like Hmong and Welsh
- **approximant** Characterized by a loose constriction; includes glides like the ⟨w⟩ in *wand* and other sounds like the ⟨r⟩ in *raven*
- **lateral approximant** A special type of approximant in which there is an opening on one or both sides of the tongue; like the ⟨l⟩ in *leprechaun*

Every sound in the world's languages has one of these manners (and one of the above places). As mentioned, some use a different airstream mechanism.

However, there is one other parameter in which sounds can vary and that is voicing. Voicing refers to phonation—the oscillation of the vocal folds (colloquially called the vocal cords). The vocal folds are pictured in Figure 2.8.

To a first approximation, *voicing* refers to whether or not phonation is occurring during a constriction. For example, vowels are typically voiced because, during the whole duration when the vowel articulation is being made, the vocal folds are vibrating. Traditionally, beginning students are told to treat consonant voicing in the same way: the vocal folds are vibrating during the production of ⟨b⟩ as in *bill*, but not during the production of ⟨p⟩ as in *pill*. This certainly works for fricatives and approximants. If you want to test it, try producing the ⟨s⟩ sound and then the ⟨z⟩ sound with your fingers on your larynx. During ⟨z⟩, you will feel a buzzing that is absent during ⟨s⟩. This buzzing is phonation. However, we are going to tell you the whole truth: voicing is actually a continuum and is based on the relative timing of a constriction and the onset of phonation. This is called VOT or VOICE-ONSET TIME. In English, the ⟨p⟩ in *pit* has a longer VOT than ⟨p⟩ in *spit*, which has a slightly longer VOT than the ⟨b⟩ in *bit*. In other languages, ⟨b⟩ sounds (and other voiced plosives) may have a negative VOT.

This means that when we talk about voiced versus voiceless consonants (especially plosives), we are actually abstracting over the real situation. Most languages distinguish at most two sets of sounds via voicing, but some distinguish more. Even when languages only make two-way distinctions, the categories are not necessarily comparable. If we were to build a cross-lingual speech recognizer that was trained on data with ⟨p⟩ sounds and ⟨b⟩ sounds in one language, we cannot expect it to necessarily map ⟨p⟩ and ⟨b⟩ onto the ⟨p⟩ and ⟨b⟩ sounds of another language, since VOT can vary considerably within these categories.

	FRONT	CENTRAL	BACK
HIGH	b**ee**t		b**oo**t
	b**i**t		b**oo**k
MID	b**ai**t	b**u**t	b**oa**t
	b**e**t	b**u**tt	b**ough**t
LOW	b**a**t		b**o**t

Table 2.2: American English examples of vowel place of articulation, assuming a dialect in which *bought* and *bot* are pronounced differently. Rounded vowels are in cyan. *But* is the unstressed version as in "I went but you stayed."

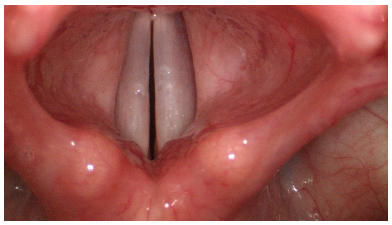


Figure 2.8: Normal female vocal folds (glottis)

2.2.2 *The International Phonetic Alphabet*

Over years of research, phoneticians have developed this systematic approach to phonetic description using the parameters air stream mechanism,

place, manner, and voicing. They have also developed a standard notation for transcribing speech sounds based upon these parameters. This is called the International Phonetic Alphabet (IPA). This alphabet has a unique symbolic representation for every speech sound in every language of the world.

The majority of the pulmonic egressive consonants of the IPA are show in Table 2.3. Additionally, there are symbols for a few other pulmonic consonants and the non-pulmonic consonants, for which see the official IPA chart.

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lat. Fric.				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lat. Approx.				l		ɭ	ʎ	ʟ			

Table 2.3: The pulmonic egressive consonants of the IPA. In each column, the symbol on the left is for the voiceless phone and the symbol on the right is for the voiced phone.

The IPA vowels are show in Figure 2.9. Finally, there are a large number of diacritics and modifiers that alter the meaning of symbols in defined ways. By combining the base symbols with the diacritics and modifiers, it is possible to represent all of the world's speech sounds.

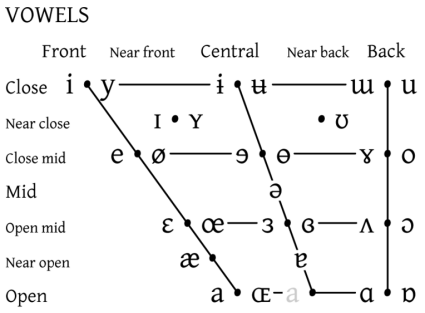


Figure 2.9: The vowels of the IPA

For a complete IPA chart and other information about the IPA, see the website of the International Phonetic Association at `https://www.internationalphoneticassociation.org/`. You are also encouraged to experiment with the interactive IPA chart at `http://www.ipachart.com/`. This chart allows you to hear the pronunciations of all the basic IPA symbols and can be very helpful in learning to apply the IPA. You should be aware, however, that the recordings on this site are not always accurate

PHONE	EXAMPLE	PHONE	EXAMPLE
[pʰ]	**p**it	[b]	**b**it
[p]	s**p**it		
[tʰ]	**t**ick	[d]	**d**ot
[t]	s**t**ick		
[kʰ]	**c**ot	[g]	**g**ot
[k]	S**c**ot		

Table 2.4: Examples of plosives as they are produced by some English speakers in North America.

PHONE	EXAMPLE	EXAMPLE
[m]	**m**it	si**m**
[n]	**n**it	si**n**
[ŋ]	—	si**ng**

Table 2.5: Examples of nasal stops.

realizations of the symbols.

2.2.3 *Coarticulation*

Sounds are not produced in isolation. In fact, speech sounds overlap with one another considerably. This is why it is somewhat naïve to speak of dividing a speech recording into non-overlapping SEGMENTS (consonants and vowels). It is not inaccurate to think of speech as a series of consonants floating by on a stream of vowels. Speech segments often do not have sharp borders or well-defined beginnings or endings. To illustrate this, carefully observe your articulators as you say *caw*, *coo*, and *key*. They all start with a ⟨k⟩-sound, and the initial consonants probably sound very similar to you, but you probably noticed that the position of the closure differs with the following vowel (and matches the place of articulation of the vowel). This is because the articulation of the consonant and the following vowel overlap. This phenomenon, called COARTICULATION, is pervasive in speech.

2.3 *Acoustic phonetics*

Articulatory phonetics is useful for describing speech sounds. You cannot understand the IPA without articulatory phonetics. However, the real action is in acoustic (and auditory) phonetics. This is the study of the acoustic signals produced by speech. To understand speech acoustics, it is necessary to review some basic acoustic physics, much of which will already be familiar to some readers. Then we can talk about the elements of acoustics that are particularly important to speech and relate speech acoustics to actual speech.

2.3.1 *Complex waves, Fourier analysis, and spectra*

The simplest kind of wave is a sine wave, which can be described by the sin function.

The vocal tract does not produce sine waves. Or rather, the vocal tract produces lots of sine waves superimposed on top of one another. Speech signals, like most naturally-occurring sounds, are encoded in COMPLEX WAVES. These waves are encoded, computationally, as WAVEFORMS. An example waveform can be found in 2.10. It is a mathematical fact that any wave, no matter how complex, can be analyzed into a series of sine waves with various wavelengths and amplitudes. The means by which this is done is FOURIER ANALYSIS. There is an efficient algorithm for doing Fourier analysis called the FAST FOURIER TRANSFORM (FFT). Imagine arranging all such sine waves from a complex wave from longest to shortest in wavelength, then taking a time window from the waves and averaging, in some way, the amplitude of the waves over the window. You would get a series of numbers that you could then plot. The resulting plot is a SPECTRUM (see Figure 2.11). A three dimensional plot (or numeric representation) that gives spectra over time is

PHONE	EXAMPLE	PHONE	EXAMPLE
[f]	**f**an	[v]	**v**an
[θ]	**th**igh	[ð]	**th**y
[s]	**s**ink	[z]	**z**inc
[ʃ]	a**ss**ure	[ʒ]	a**z**ure
[h]	**h**at		

Table 2.6: Examples of plosives.

PHONE	EXAMPLE	PHONE	EXAMPLE
[l]	**l**ip	[ɬ]	pi**ll**
[w]	**w**ack	[j]	**y**ack
[ɹ]	**r**ack		

Table 2.7: Examples of approximants.

PH	EX	PH	EX	PH	EX
[i]	b**ee**t			[u]	b**oo**t
[ɪ]	b**i**t			[ʊ]	b**oo**k
[ej]	b**ai**t	[ə]	b**u**t	[ow]	b**oa**t
[ɛ]	b**e**t	[ʌ]	b**u**tt	[ɔ]	b**ough**t
[æ]	b**a**t			[ɑ]	b**o**t

Table 2.8: Examples of vowels.

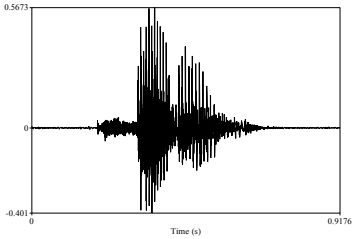


Figure 2.10: An example waveform of *Potter*

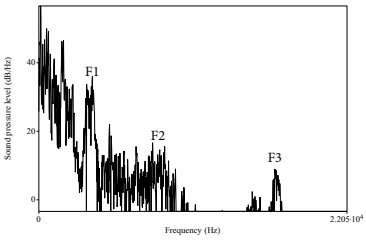
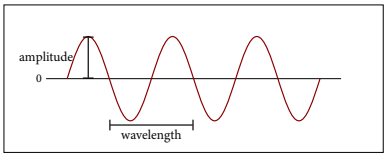


Figure 2.11: An example of a spectrum of speech



a SPECTROGRAM. Typically, in a spectrogram, higher amplitude/energy is indicated by a darker shade of gray while lower amplitude is indicated by a lighter shade of gray. An example spectrogram is shown in Figure 2.12.

An experienced phonetician can read speech from a spectrogram and identify both speech sounds themselves and the effects of coarticulation. But even if you cannot read spectrograms, understanding terms in which they are analyzed can allow you to understand other speech concepts which are important to the construction of speech technologies.

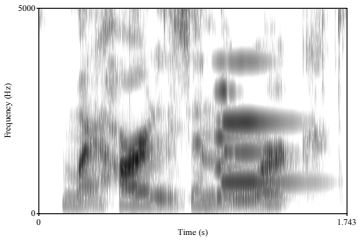


Figure 2.12: A spectrogram of "My heart is in the work."

2.3.2 Harmonics, resonances, and formants

Let us return to Fourier analysis. Each of the component waves, into which we analyzed complex waves, are multiples of a single frequency, which we will call the FUNDAMENTAL FREQUENCY or F0. In many cases, this is the frequency at which the vocal folds are vibrating. Each of these multiples, including the fundamental itself, are called harmonics. We call the fundamental H1 and each of the higher harmonics H2, H3, and so on. Harmonics are not all of the same amplitude. F0/H1 tends to have a lot of energy; successive harmonics tend to trail off in amplitude. However, in speech there are often bands of relatively high-energy harmonics (from the viewpoint of a spectrogram). These are called formants. We label them as F1, F2, F3, and so on. The most important formants for the speech sciences are F1, F2, and F3 (though F4 and F5 are occasionally important). An example spectrogram with the formants labeled is given in Figure 2.13.

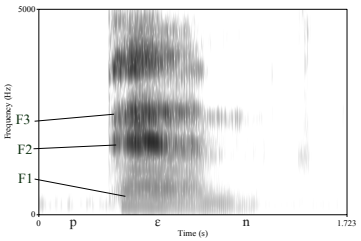


Figure 2.13: Spectrogram of the word *pin* with the first three formants labeled.

Why do these bands of higher energy exist? The vocal tract, like any physical system, has one or more resonant frequencies—frequencies at which the response amplitude is a relative maximum. You might compare the vocal tract to a pendulum, like a child's swing. No matter the frequency at which you push the child, the frequency at which the swing oscillates will remain basically constant. Pushing the swing at a different frequency will elicit less of a response for the same amount of pushing. A (potentially branched) tube like the vocal tract has resonant frequencies in much the same way, but it has an infinite number of them. If a harmonic is relatively near a resonant frequency, it will be relatively higher in amplitude, other things being equal, than one that is far away. Formants are clusters of harmonics that are near a resonant frequency.

The vocal tract is essentially a tube that is closed at one end. The resonances of such a tube correspond to the kinds of standing waves it could support. An illustration of the first three such wave patterns is given in Figure 2.14. The waves show the relative velocity of air particles. For all of the waves, the particles have a very low velocity near the closed end of the tube and a relatively high velocity at the open end of the tube. The areas of lowest particle velocity (the points where the two lines intersect) are called velocity nodes; those areas with the highest particle velocity (where the lines are

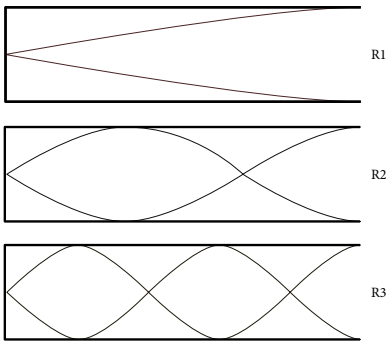


Figure 2.14: The first three resonances of a tube closed on one end

furthest apart) are called velocity antinodes.

The physics behind this are well-established and widely known. First, it is important to know that there is a deterministic relationship between frequency, wavelength, and velocity (of sound waves). This is stated in Equation 2.1:

$$f = \frac{c}{\lambda} \tag{2.1}$$

where f is the frequency of a wave, c is it's velocity (the speed of sound, 343 meters/second) and λ is the wavelength of the wave. From this formula, we can derive an approximation for the resonant frequencies of a tube closed at one end:

$$f = \frac{nc}{4L} \tag{2.2}$$

where n is an odd integer (1, 3, 5, ...) and L is the length of the tube in meters. In reality, however, the diameter of the tube also influences the resonant frequencies. A more accurate formula is:

$$f = \frac{nc}{4(L + 0.4d)} \tag{2.3}$$

where d is the diameter of the tube (in meters).

You may be wondering how we get from the rather static illustration in Figure 2.14 to the dynamics of actual speech. In fact, the principles are quite simple. A constriction near a velocity will raise (in frequency) the corresponding resonance. An expansion near a velocity node will have the opposite effect. Likewise. a constriction near a velocity antinode will depress the corresponding resonance and an expansion will raise it.

With these tools in hand, you equipped to understand a great deal about how speech acoustics relates to articulation. The conceptual connection between these concepts is provided by the source-filter model of speech production.

2.3.3 Perturbation theory and the source-filter model of speech production

At its foundation, the source-filter model is simple: there is a source of sound and an acoustic filter that dampens some frequencies within the signal produced by that source. Prototypically, the sound source is the vibrating glottis (the vocal folds). However, it can be other sources of sound like a uvular trill, the hiss of a palatal fricative, or the release of a dental click. The filter is always the vocal tract. It may be a simple tube; it may also be a branched tube (when the velopharyngel port is open), in which case it has markedly different acoustic properties.

The source produces a complex wave with many components. The filter has certain resonances. Components close to these harmonics pass through the filter unaffected. The other components are damped. The resonances are predictable from the shape of the tube, with constrictions and expansions

near nodes and antinodes having a marked effect upon the frequencies of resonances.

The alignment between the first three resonances and the anatomy of the vocal tract is given in Figure 2.15. This diagram can serve as a kind of

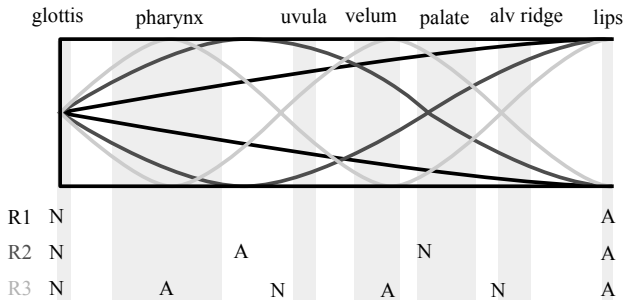


Figure 2.15: Velocity nodes and antinodes for the first three formants shown relative to articulatory landmarks

"decoder ring" except that it is not a torus. If a vowel has a low F1 and a high F2 and F3, this can be explained by a palatal constriction. This is consistent with the vowel [i] (the vowel in *beet*). This vowel is high and front—it has a relatively close palatal constriction—meaning that it has a constriction near an antinode of R1, a node of R2, and (to a lesser extent) a node of R3. Likewise, if we wish to predict the influence on a neighboring mid-central vowel of a uvular stop, we can observe that the constriction will be somewhat near a node for R1, near an antinode for R2, and near a node for R3. We thus expect F1 to be elevated, F2 to be depressed, and F3 to be elevated.

This model is heuristically useful. It is possible to produce a more exact mathematical model of vocal tract acoustics using an idealization of the vocal tract. A more detailed description of how this can be done is found in Johnson (2011).

2.3.4 The acoustics of vowels

All vowels have formants. Just as vowels have two primary articulatory dimensions, they have two primary acoustic dimensions, in terms of formants. There is a mapping between these two spaces. Vowel height is inversely correlated with the frequency of F1—high vowels have a low F1 and low vowels have a high F1. Vowel backness is inversely correlated with F2—front vowels have a high F2 and back vowels have a low F2. Lip rounding lowers all formants, an effect that is particularly pronounced with F2 and F3 (and the difference between F2 and F3). Some languages have sequences of vowels, or of a vowel and a glide, that behave as a single sound. These are called diphthongs. Examples include [ɑɪ] (the ⟨y⟩ in *why*) or [ɔʊ] (the ⟨o⟩ in *woe*). They are characterized by dynamic formants. For example, in [ɑɪ], F1 drops progressively. F2 gradually rises till it is very close to F3.

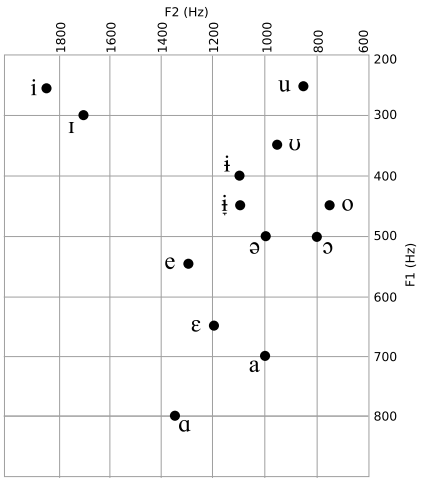


Figure 2.16: The vowels of Welsh plotted according to average F1 and F2

2.3.5 *The acoustics of consonants*

The acoustics of consonants is more complicated. Different manners of articulation are associated with very different acoustic signals. Here are a few of the major classes:

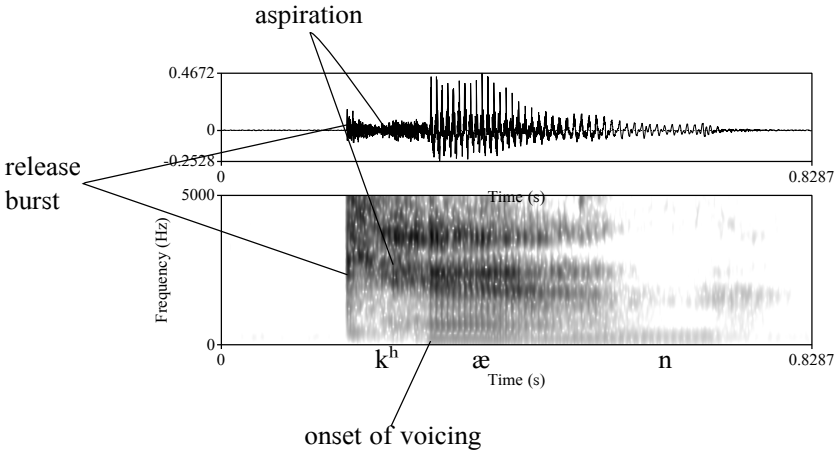


Figure 2.17: Spectrogram and waveform of *can* with acoustic landmarks labeled

Plosives The common element of plosives is the STOP RELEASE. This is a burst of energy produced when closure is opened. Before this closure, there may be a period of silence (for voiceless plosives) or a largely silent period with a detectable F0 (for "strongly voiced" plosives). When a stop is aspirated, as in Figure 2.17, the release is followed by a brief interval of aperiodic noise. This can be detected on the waveform as the absence of any repeating pattern. On a spectrogram, aperiodic noise looks vaguely fuzzy and lacks the structure seen when turbulence is absent. The cues for place of articulation of plosives are encoded two principle places: in the spectral properties of the release and

in the formant transitions between a plosive and neighboring vowels.

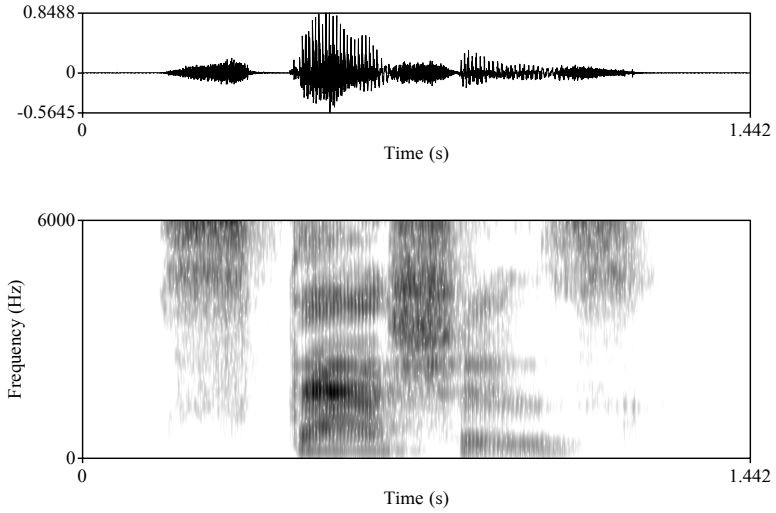


Figure 2.18: Spectrogram and waveform of *stashes*

Fricatives Fricatives consist just of noise, so they look like longer versions of the aspirated portions of aspirated plosives. The place cues for fricatives are based largely on the energy distribution of the noise. Note that, in Figure 2.18, [s] and [z] have most of their energy at or above 6000 Hz, while in [ʃ], the center of gravity of the energy is quite a bit lower. The spectral properties of fricatives depend on the state of the vocal tract (as resonator) but also on the nature of the constriction the produces the turbulence.

Nasals and laterals

2.3.6 *The acoustics of tone and intonation*

Tone is the use of pitch (the perceptual correlate of the fundamental frequency or F_0) to make LEXICAL DISTINCTIONS[13]. As such, it is a phonological notion—that is, it is about the way sounds function in the grammar of a language, not about the sounds themselves—rather than a phonetic one. Phonetically, it can be grouped with intonation, which is the use of pitch for other purposes. Both of these ways of using pitch are SUPRASEGMENTAL, that is, they do not characterize a single segment (a single consonant or vowel) but a larger span like a syllable, a word, a phrase, or an utterance.

[13] Lexical distinctions are distinctions between words

Pitch may be high or low, on a certain scale, as well as level, falling, rising, or some combination of these. Many tone languages of East and Southeast Asia have pitch contours from tone that are distributed over syllables and may have a range of contours. Take the example below from Hmong, a language of China and Southeast Asia: Intonational contours offer the same options, but

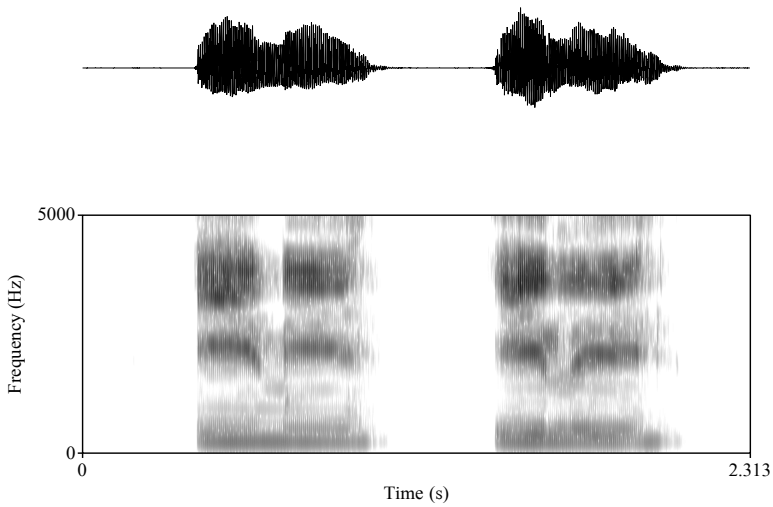


Figure 2.19: Spectrogram and waveform of /ini/ and /ili/

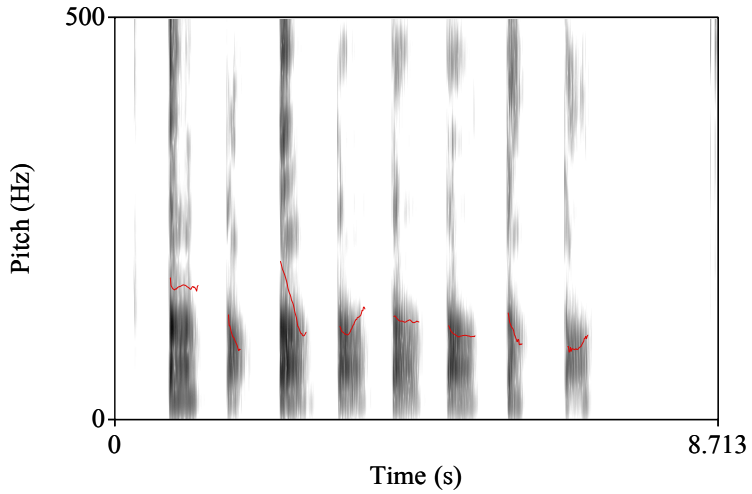


Figure 2.20: The pitch contours (shown in red) of the Hmong Daw words *tob* 'deep', *tom* 'to bite', *toj* 'hill', *tov* 'to mix', *to* 'to be punctured', *tos* 'to wait', *tog* 'to sink', and *tod* 'over there'

they are distributed over a potentially wider domain. They may also interact with phonological constructs like stress. For example, pitch is elevated around the primary stresses of words in most varieties of English. Thus, the distinction between *'object* 'thing' and *ob'ject* 'take issue with' is partially indicated by a difference in vowel quality and partially indicated by a difference in pitch contour. Amplitude (and thus loudness) may also play a role.

2.4 Phonetics and speech technologies

Phonetics is very relevant to speech technologies, both speech synthesis and speech recognition. As such, you might assume that speech researchers and engineers must know a lot about phonetics. In the past, this was definitely true. However, with the advent of neural models—particularly for automatic speech recognition—domain knowledge has been less necessary (or at least, less valued) and attention has been focused on very general architectural considerations that obscure the differences between human speech and other kinds of digital signals.

2.4.1 Phonetics and speech synthesis

2.4.2 Phonetics and speech recognition

The task in ASR[14] is, given a speech signal, to decode it into text. This can be understood in terms of the noisy channel model of Shannon (XYZ). This model, which can be derived from Bayes's Rule, has two parts: a channel model (in ASR this is the ACOUSTIC MODEL and a LANGUAGE MODEL. The speech signal comes to the ASR system as a waveform. When we are processing the speech signal as speech scientists, a logical next step is to convert the waveform into a spectrogram (or a time series of spectra). Historically, machine learning models were not powerful enough to learn patterns from raw waveforms or spectra. Insteam, systems extracted MFCCs—Mel frequency cepstral coefficients. This is a way of representing a spectral slice in a relatively small sequence of numbers. The task is then to learn an acoustic model that can translate these MFCCs into sequences of words or phones and a language model that evaluates how probable such a sequence is. Contemporary neural ASR systems are so powerful that explicit feature engineering is no longer necessary, at least for languages and domains where there are extensive data resources.

[14] Automatic speech recognition

Are ASR researchers and engineers freed from the need to learn phonetics, then? Actually, no. While end-to-end systems[15] are state of the art for many commercially important tasks, they are less well-adapted to scenarios where there is limited data. For example, it is sometimes helpful (particularly where writing systems are not transparent) to convert text to a sequence of phones[16]. This requires a detailed knowledge of the sounds of a language and how they relate to the writing system. This is a matter of articulatory and descriptive

[15] In this case, an END-TO-END SYSTEM is one that maps directly from input to output with no explicit intermediate steps.

[16] This is called GRAPHEME-TO-PHONEME TRANSDUCTION or G2P. In reality, it should often be characterized as "grapheme-to-phone transduction"

phonetics. Acoustics phonetics also has a role to play in the error analysis of ASR systems, since it can tell an investigator which phones are likely to be confused with one another. For example, the vowels /y/ and /u/ are more likely to be confused by an ASR system than the vowels /y/ and /ɑ/. This kind of knowledge can help an engineer or scientist know why a model is failing in a particular case and help them, for example, present training data to a system in such a way as to mitigate the failure.

2.5 Exercise: plotting a vowel space

The exercise for this unit will be creating a vowel plot, in F2 by F1, for your native language. You will turn in the following things:

1. A list of example words for each of the vowels in your language (the same words that you will record)
2. A spectrogram of one example of each of the vowels
3. A plot of each vowel token recorded (at least 10 per type) with F1 on the y-axis and F2 on the x-axis.
4. Paragraphs answering each of the following questions:
   (a) How compact are the "clouds" of exemplars for any given type? How much do they overlap? How do you explain this observation?
   (b) Are there any outliers? To what do you attribute them?
   (c) Human first-language learners do not know, in advance, how many vowel types their language will have. Given data like that you have extracted, how would you construct a program that learned the optimal number and identity of vowels in a vowel system?

2.5.1 Determining what vowels your language has

The second most challenging part of this exercise is figuring out how many vowels your language has an what they are (in articulatory terms). Naïve answers to this question often turn out to be wrong. For example, a speaker of English might suppose that there are five English vowels, *a*, *e*, *i*, *o*, and *u* based on the fact that there are five vowel letters in the English alphabet. In fact, American English has ten or eleven vowels and a single vowel letter may be used to write more than one vowel sound. Other writing systems are similarly deceptive and cannot be used uncritically as a guide.

The best place to find out about the vowels of your language is in the Illustrations of the IPA published in the Journal of the International Phonetic Association (JIPA). Many of these have been collected together in the *Handbook of the International Phonetic Association*. However, there is not an illustration of the IPA for every language. The second-best place to look is in a reference

grammar of the language. Barring that, it can sometimes be useful to consult online resources like Wikipedia and Omniglot.

2.5.2 Finding an optimal set of words exemplifying the vowels

Once you know what the vowels are, it is necessary to develop a set of words in which they can be elicited. In the ideal case, these will be a minimal set—a set of words that differ only in the vowel under consideration. An example of a (near)-minimal set for American English is *beet*, *bit*, *bait*, *bet*, *bat*, *boot*, *book*, *boat*, *bought*, *bot*, and *but*. It is unlikely that you will be able to come up with such a list off the top of your head. The structure of your language may also make such a list unlikely. Do not worry if you cannot identify a complete minimal set, but do try to keep the vowels in as similar a context as possible (if one is word-final, all should be word-final. If one is stressed, all should be stressed, etc.)

2.5.3 Making a recording

You can make your recording using Praat, Audacity, or any other software that can record audio as a WAV file. Be sure to make your recording in a quiet place where you will not be felt strange for repeating words to a computer. It is helpful if you wear a microphone (this will boost the signal-to-noise ratio) but it is not absolutely necessary.

To record in Praat, go to the Praat Objects window and select Record Mono Sound... from the New menu in the top-left corner of the window. You can keep the default Sampling Frequency (44100 Hz). Click record when you're ready and wait until a gray background appears in the meter. Then start recording. When you are done, click Stop. Then click Save to list & Close.

Record ten tokens of each vowel type.

2.5.4 Annotating the recording

You now need to tell Praat, or another program you are using to extract the formant frequencies, where the vowels are located in the recording. To do this, you need to create a Praat TextGrid. Select the Sound object for which you want to make an annotation and select "To TextGrid..." from the "Annotate" menu. For All tier names enter "vowels". By "Which of these are point tiers?" leave the text input blank. Click OK. This will create the a TextGrid with one tier, called "vowels". Shift-click to select both the Sound object and the matching TextGrid object. Then select "View & Edit," which will open the TextGrid editor. You can see this editor in Figure 2.21.

Click on the spectrogram or waveform where you wish to place the initial boundary for each vowel. You will see a vertical line with a circle at the top of the tier.

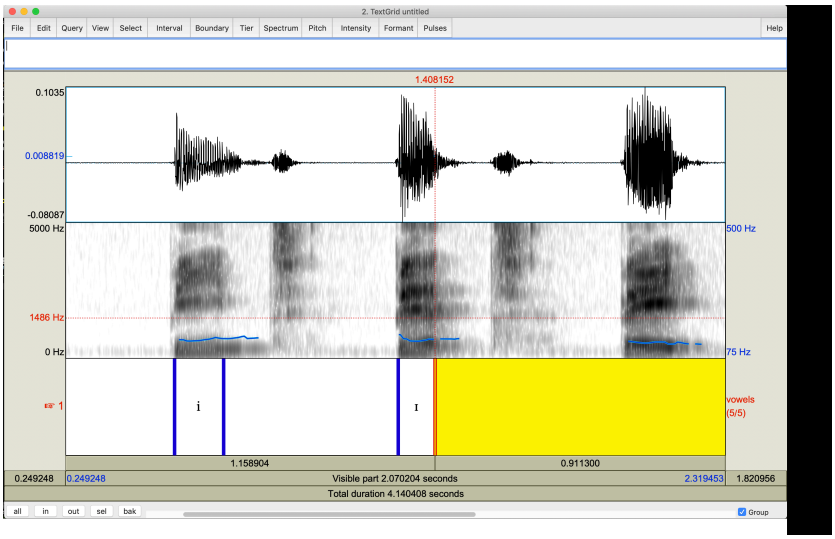


Figure 2.21: Annotating a sound file by marking up vowels in a TextGrid

Click in this circle and a new boundary will be created on that tier. If you start typing, the annotation will apply to the span that begins at that boundary. It appears both in the tier (centered between the initial boundary and the end of the window or the next boundary) and in a text entry box near the top of the window.

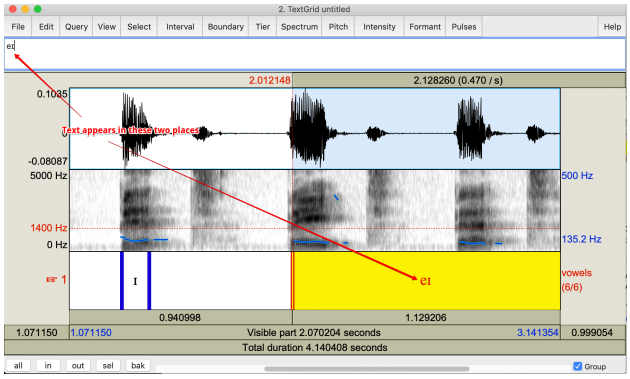


Figure 2.22: Adding a TextGrid boundary

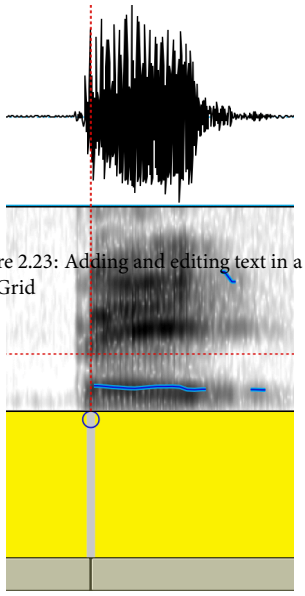


Figure 2.23: Adding and editing text in a TextGrid

When you have finished annotating your sound file (and probably before that, since Praat is not incredibly stable) save the sound file and TextGrid using the "Praat Objects" window. Save the WAV files and the TextGrids in one folder. When you are done, you can run the script that will extract the formant measurements and provide a preliminary plot of the data.

2.5.5 *Producing the plot*

To extract the data that you need to produce the plot, you may download a Praat script for extracting formant values. You should save this script in

a folder that is a parent to the folder where the sound files are saved. For example, you might save the whole project in the folder `vowel_plot/`, placing the script here, and save the sound files in `vowel_plot/sounds/`.

The script is not hard to use, but some details require special explanation. To open the script, use the "Read from file..." option from the "Open" menu on the "Praat Objects" window. This will open a script window with the text of the script. You can then run the script by clicking the "Run" menu on the top of the script window and selecting "Run." You will see a dialog box like the one in Figure 2.24. The first text input provides the **relative** path to the

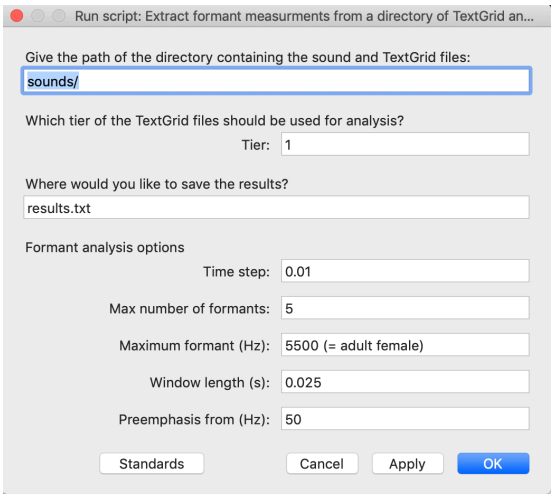


Figure 2.24: Dialog box for Praat formant value extraction script

sound files and TextGrids. If this is something other that `sounds/`, you may want to edit the script to change it. Since our TextGrid(s) only have one tier, we can leave "Tier" set to "1". The formant measurements will be written to a tab-delimited textfile which is called, by default, `results.txt`.

All of that should have been fairly obvious. The tricky part is next: speakers of a language differ in their physical properties—the length and thickness of their vocal folds and the length of their vocal tracts—and these influence the acoustics of speech. Humans normalize for this easily, but Praat does not. It uses algorithms for extracting formants, that are not robust to this kind of variation. Thus, in order to get optimal results, it is necessary to adjust some parameters.

The most important of these is "Maximum formant (Hz)," which constrains the highest formant that the algorithm is allowed to find. For adult females, the default of 5500 Hz is reasonable and usually works (unless you have an unusually high-pitched voice). If you are relatively tall, or male, or both (and thus have a "deeper" voice), try lowering this parameter to 5000 Hz (or even lower). How do you know if you should adjust this value? If when you plot your vowel space there are many outliers, especially if vowels of the same type are located in to different parts of the plot, this suggests that the formant extraction algorithm is inappropriately calibrated. You may also

experiment with "Max number of formants," but this is usually unnecessary.

If your vocal tract is especially long or short, you might consider editing the script to change the dimensions of the plot. This may make debugging the parameters of the formant extraction algorithm easier. However, you are not encouraged to turn in the plot from Praat as your final product, since it does not provide a clear visualization of the distribution of vowel tokens when the number of tokens is relatively large. Instead, use the data from `results.txt` to make a different plot, using the best tools and ideas at your disposal, that presents your vowel space in as informative a way as possible.

2.6 *Conclusion: Back to Laurel and Yanny*

You should now have the knowledge you need to untangle the puzzle with which we started the unit: why, when listening to the same low-quality recording of speech do some listeners hear Laurel and some hear Yanny. The answer has to do with formants and how consonant and vowel articulations influence them.

The "Laurel-Yanny" recording has two unusual characteristics: F2 is largely indistinct and there are strange artifacts in the higher frequencies of the recording. It is as if the original recording (of *Laurel*) had copies, at successively higher frequencies, superimposed upon it. We will concentrate on the first issue, which is illustrated in Figure 2.25. In a higher-quality

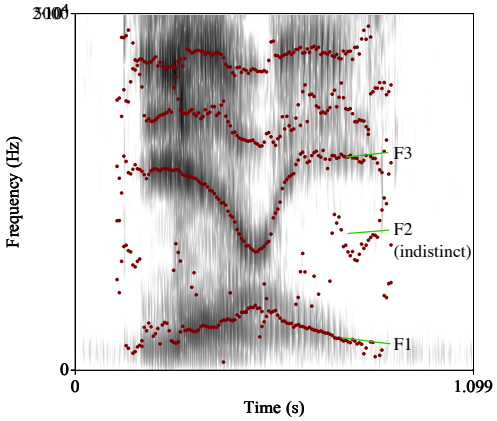


Figure 2.25: Spectrogram of the "Laurel-Yanny" recording (originally of *Laurel*). F2 is indistinct leading some listeners to perceive the sound as if F3 was F2.

recording of *Laurel*, there would be a formant audible between F1 and F3 at the beginning of the word. Under the right conditions, a listener will fill this in perceptually. If they do not, however, they will perceive F3 as F2 and F4 as F3. This means that the first sound will be perceived as [j] (with a low F1, a high F2, and an F3 very close to F2). This initial perceptual "decision" predisposes the listener to interpret the rest of the signal in a particular way. For example, the first vowel, with its low F1 and reasonably high (apparent) F2 is perceived as [æ] instead of [ɑ]. The drop in F3 that is characteristic of [ɹ]

is reinterpreted an effect of nasal [n] on the surrounding front vowels. Finally, The part of the word that would be a syllabic [l] is perceived instead as the vowel [i] because of the elevated F2 and the low F1.

Human speech is complicated, but it need not be mysterious. Just as apparent perceptual mysteries like "Laurel-Yanny" dissolve when examined in light of acoustic phonetics, so many of the mysteries of everyday speech can be resolved through the findings of speech science.

3 Structure in the Sounds of Human Language

Phonetics makes a lot of things that are the same seem different. The ⟨t⟩ sounds in *top* and *stop* are, as they function in English, the same sound. However, acoustically they are quite unalike. Likewise, we are adding the same plural suffix to *dog*, *cat*, and *horse* to form *dogs*, *cats*, and *horses*, but this suffix has a different phonetic realization in each of these words based on the final sound in the base to which it attaches. The job of phonology is to find the underlying unity in this sea of variation and reduce it to computationally tractable rules and constraints.

3.1 Phonology as normalization

Phonology sits between phonetics and morphology.

It maps between an abstract morphological world, where every unit of meaning has one realization, and the messy, concrete world of phonetics where the same meaningful unit may have multiple realizations, conditioned upon the environment in which it occurs. As such, phonology can be seen as normalization.

Such normalization is possible because the contextually conditioned changes that produce the observed variation are not random—they are rule governed and can be described by systems of rules and constraints. A great many of these are what we would call "phonological rules" and would attribute to the computational capacities of human cognition. It is true that some phonetic variation arises from purely mechanical factors in articulation (you never step in the same river twice and you never produces the same sound twice) or from perceptual factors (variation may be perceived where, acoustically, there is none, or may be factored out where it is present). Such facts, while they are not phonology *per se*, are often incorporated into phonological systems through historical processes (PHONOLOGIZATION[1]) and are thus important to understanding phonological patterns. In fact, they may be the *chief* source of phonological patterns, resulting in the apparent "phonetic naturalness" of phonological rules.

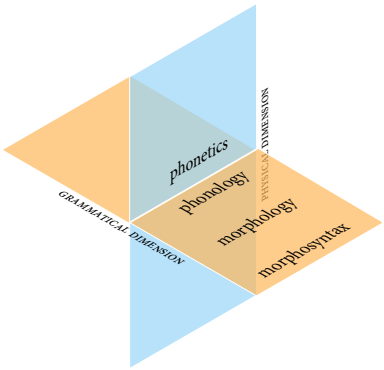


Figure 3.1: Phonology sits at the interface between physics (phonetics) and grammar (morphology and syntax)

orthographic	phonetic
im-possible	[ɪm-pʰɑsəbl̩]
in-tolerable	[ɪn-tʰɑlərəbl̩]
in-conceivable	[ɪŋ-kʰənsivəbl̩]
il-legal	[ɪl-ligl̩]
ir-regular	[ɪɹ-ɹɛgjulɹ̩]

Table 3.1: Two representation of words beginning with /ɪN/ 'negative.'

[1] Phonologization is the process through which automatic, phonetic variation is incorporated into the phonology of a language.

3.1.1 Allophony

It is traditional to begin discussions of phonology with an explanation of allophony—of phones, allophones, and phonemes. This has a number of pitfalls in that it often leads to confusion of concepts like PHONEMIC REPRESENTATION and UNDERLYING REPRESENTATION. Nevertheless, because the notion of the phoneme and the concept of allophony are important to language technologies, we will give them pride of place.

An idealized representation of a physical sound is called a PHONE. Phones and phonetic representations are typically written between square brackets, so we might talk of the phone [tʰ] or the word [tʰɪk] 'tick'. As discussed in Section XYZ above, phonetic representations can be transcribed with varying degrees of detail, which can be confusing. The transcription [tʰɪk̚] is a BROAD TRANSCRIPTION that includes only information that is relevant to the matter under discussion. With such a system of transcription, the word 'stick' would be transcribed as [stɪk̚]—the "t" sound is pronounced differently. This might be surprising at first: in some sense, the [tʰ] in 'tick' and the [t] in 'stick' are the same sound. Certainly, they sound the same to native speakers of English. This is because they are variants ("allophones") of one PHONEME[2].

ALLOPHONES are phonetic realizations of a phoneme that occur predictably in a specific set of environments. For example, [tʰ] is a phone that occurs only at the beginning of words and at the beginning of stressed syllables. [t̚], as in [kʰɪt̚], occurs at the end of words. [t], to a first approximation, occurs in all other environments. These three sounds are allophones of the same phoneme. By convention, we call a phoneme after the phone with the least restricted distribution, so we would call the English phoneme with the allophones [tʰ], [t̚], and [t] /t/ (between slashes[3]).

Allophonic rules tend not to be completely idiosyncratic. For example, in English, all of the voiceless plosives phonemes have aspirated allophones that occur word initially (and in the onsets of stressed syllables), unreleased allophones that occur word-finally, and unaspirated, released allophones that occur elsewhere.

Allophonic rules are language specific. In many other languages, [k] and [kʰ] are allophones of different phonemes. One example is Mandarin Chinese, where 古 /ku˧˩˧/ means 'old' but 苦 / kʰu˧˩˧/ means 'bitter'. Likewise, in Hmong, /pi˧˩/ means 'vulva' but /pʰi˧˩/ means 'to fit; to suit´. In these languages, unaspirated and aspirated plosive phones belong to separate phonemes.

The task of phonemic analysis should return two products: a new representation of the phonetic input in which all phones have been given a normalized representation as phonemes and a set of rules by which this normalized representation can be converted back into a phonetic representation. There are various ways this could be done computationally, but we will start with a

[2] Phonemes are abstract representations of a sound (consonant, vowel, tone, etc.) If one phoneme in it is substituted for another, the meaning of a word is changed. Thus, phonemes are primarily conceived of in terms of contrast—they are the minimal contrasting units of sound.

[3] In order to confuse novices, phonologists also use slashes to enclose what are called "underlying representations," a type of representation still more abstract than phonemes. We will adopt that same, unfortunate, notation in order to perpetuate the injustices of the past.

aspirated	unreleased	unaspirated
p̲in	nip̲	spin
t̲ick	kit̲	stick
k̲in	nic̲k̲	skin

Table 3.2: Examples of words with different allophones of English voiceless plosives.

pencil-and-paper version.

Consider the following data from Korean: To perform a phonemic anal-

kal	'that'll go'	ilkop	'seven'	iɾɯni	'name'
kɯnɯl	'shade'	ipalsa	'barber'	kiɾi	'road'
mul	'water'	onɯlpːam	'tonight'	kɯɾəm	'then'
pal	'leg'	pulpʰjən	'discomfort'	kəɾiɾo	'to the street'
pʰal	'arm'	silkwa	'fruit'	saɾam	'person'
səul	'Seoul'	tɯlt͡ʃaŋ	'window'	uɾi	'we'
tatɯl	'all of them'	əlmana	'how much'	yəɾɯm	'summer'

Table 3.3: Data from Korean

ysis, the first step is to find the distribution of each of the phones. To do this, we place a "window" around each of the phones and record the phones occurring before and after it (the phone). We will use the hash mark to indicated word boundaries (before the first phone and after the last phone). To illustrate this technique, we sometimes use a notation called the "t-chart". T-charts for Korean [l] and [ɾ], based on the data in Table 3.3 would look list this:

l	
a	#
ɯ	#
u	#
a	#
a	#
u	#
ɯ	#
i	k
a	s
ɯ	pː
u	pʰ
i	k
ɯ	t͡ʃ
ə	m

ɾ	
i	ɯ
ɯ	ə
ə	i
a	a
u	i
ə	ɯ

Table 3.4: Illustrations of t-charts

Since it is always predictable, in Korean, whether a given context can have an [l] or a [ɾ], would could produce a transcription of Korean in which these two sounds were only represented by /l/[4] If we applied this procedure to every pair of phones in a corpus of Korean, we would arrive at a phonemic representation—a representation from which all predictable phonetic detail has been factored (normalized) out.

It is important to note that phonemic solutions are not unique. This was shown by the linguist Y.R. Chao in the 1930s[5].

[4] Slashes are used here to represent PHONEMIC representations as opposed to PHONETIC representations, which are enclosed in square brackets.

[5] Chao, Y.-R. (1934). The non-uniqueness of phonemic solutions of phonetic systems. *Bulletin of the Institute of History and Philology 4*, 363–397

3.1.2 Allomorphy

Phonemic analysis was one of the earliest phonological activities (after phonological reconstructions, which we will not discuss at length). Phonemic anal-

ysis reduces a set of redundant phones to a set of contrasting phonemes and a set of rules by which the phonetic representation can be derived from the phonemic representation. However, there are other kinds of regularities in the sound systems of languages and these are captured by MORPHOPHONEMICS or MORPHOPHONOLOGY.

In order to understand morphophonology, it is necessary to know a little bit about MORPHEMES[6]. A morpheme is a pairing of form (a representation in terms of segments[7] or GRAPHEMES[8]) and meaning. Like phonemes, morphemes are defined in terms of contrast (each morpheme in a language differs from every other morpheme in that language in either form or meaning (content), or both.

[6] A morpheme is a minimal form-meaning pairing (that is, a minimal sign or construction).

[7] A segment, to a first approximation, means a consonant or vowel, whether phonetic or phonemic.

[8] Graphemes are the minimal contrasting units of orthography (or writing. In langauges written with an alphabet, letters are graphemes.)

A morpheme may have different forms. These are called ALLOMORPHS and this phenomenon is called ALLOMORPHY. Allomorphy may depend on (may be CONDITIONED by) different factors; here we will consider only allomorphy that is conditioned by speech sounds (or phonologically conditioned allomorphy). A simple example can be drawn from English.

There are many ways to produce the plural of nouns in English, but the most common is to suffix an ⟨-s⟩ or ⟨-es⟩ to the noun. This *-s/-es* is a single morpheme, but it is pronounced three different ways: Each of these different

Singular	Phonemic	Plural	Phonemic
dog	/dɑg/	dogs	/dɑg-z/
cat	/kæt/	cats	/kæt-s/
horse	/hɔɹs/	horses	/hɔɹs-əz/

Table 3.5: Three realizations of the English nominal plural

pronunciations of the plural is an ALLOMORPH of the plural morpheme. They are predictable based on their phonological context. Consider various words that take the /-s/ allomorph: *ants*, *elephants*, *bandicoots*, *storks*, *cops*, *fifths*, etc. All of these STEMS end in a voiceless sound that is not a SIBILANT[9]. Contrast the words that take the /-əs/ allomorph: *foxes*, *asses*, *dishes*, *watches*. These occur only after sibiliants. That leaves /z/, which occurs after other sounds. We can write a rule to describe this distribution. It would be different from an allophonic rule in that it changes one phoneme into another (since /z/ and /s/ are different phonemes) and it inserts (or perhaps deletes) a phoneme (/ə/). Therefore, our rule cannot be an allophonic rule.

[9] A sibilant is a loud fricative or affricate sound that is produced by blowing a stream of air against the back of the teeth. Examples include the ⟨s⟩ /s/, ⟨sh⟩ /ʃ/, and ⟨ch⟩ /t͡ʃ/ sounds.

We might be tempted to formulate a rule set that accounts for allomorphy in the English plural. However, this would be a mistake as the processes illustrated by this case are more general. Consider the third person singluar suffix ⟨-s/-es⟩ that occurs in most English verbs. It displays exactly the same behavior as the plural suffix: It is clear that there is a general phonological pattern to be accounted for. Morphophonology is about developing rules and representations such that each morpheme has exactly one form and all variant forms can be predicted by the GRAMMAR.

To illustrate this further, let us consider an example from Maori, the in-

Infinitive	Phonemic	3SG	Phonemic
take	/tejk/	takes	/tejks/
give	/gɪv/	gives	/gɪvz/
watch	/wɑt͡ʃ/	watches	/wɑt͡ʃəz/

Table 3.6: Three allomorphs of the English third person singular suffix

digenous language of New Zealand: If we parse off the suffixes, we notice

Verb	Passive	Gerund	Gloss
hopu	hopukia	hopukaŋa	'to catch'
aru	arumia	arumaŋa	'to follow'
tohu	tohuŋia	tohuŋaŋa	'to point out'
maatu	maaturia	maaturaŋa	'to know'

Table 3.7: Maori alternations

xb

something interesting: The root morphemes have two different forms: one

Verb	Passive	Gerund	Gloss
hopu	hopuk-ia	hopuk-aŋa	'to catch'
aru	arum-ia	arum-aŋa	'to follow'
tohu	tohuŋ-ia	tohuŋ-aŋa	'to point out'
maatu	maatur-ia	maatur-aŋa	'to know'

Table 3.8: Maori alternations with morphs parsed

that appears when the root occurs without suffixes (in isolation) and one that occurs when there is a following suffix. It is possible to predict the isolation form from the other form, but the reverse is not true. In other words, you could model this ALTERNATION[10] as deletion of a word-final consonant, but not as insertion of a consonant before a suffix (since there would be no way of predicting which consonant should be inserted).

[10] An alternation is a rule-governed "exchange" between two segments, or a segment and the empty string, across instances of the same morpheme in different contexts.

3.2 Phonology as language modeling

Phonology describes the relationship between surface forms and some more abstract representation. It also describes patterns within the surface form. The system of constraints that govern how a surface form can be structured—what sounds can occur next to what other sounds, for example—is called PHONOTACTICS. To name a famous English example due to Morris Halle, *blick* is a possible, but unattested, English word; however, *bnick* is not. At the beginning of an English word, [bl] is a possible sequence but [bn] is not (even though other languages allow initial [bn] sequences. In essence, in phonotactics one is trying to predict—given a preceding string of segments—the possible or probable following segments. It is an open question in phonology whether phonotactic patterns are the result of the same types of rules or constraints that are used to model the relationship between abstract and surface representations (the "channel model") or if they have a cognitive life of their own.

3.3 Phonology as symbol decomposition

As we will see later, another way of looking at phonology is in terms of NATURAL CLASSES—sets of segments that behave the same (whether in allophonic and morphophonological alternations or phonotactic patterns) and share phonetic properties in common. One way that this is captured, in phonology, is by decomposing individual segments into PHONOLOGICAL FEATURES. For example, there is held to be a feature [syllabic] that has the value [+syllabic] in segments that are syllabic nuclei (like vowels) and [−syllabic] in segments that are not (like most consonants). Using phonological features allows us to write rules and constraints for alternations and phonotactics that are relatively simple and insightful. However, first we will look at a version of phonology that does not have features.

3.4 Phonology as a computational system

From its inception, phonological theory has been primarily a computational discipline. It has been about deterministically predicting an output from one or more inputs. A good phonological analysis functions like a good algorithm. If the inputs are well-defined, it produces well-defined, and correct, outputs.

3.4.1 Underlying representations

We have previously alluded to the idea that phonology is (among other things) about factoring out all phonologically conditioned allomorphy to yield "basic" forms for each morpheme. These basic forms are called UNDERLYING FORMS and are the counterparts of SURFACE FORMS—the various allomorphs of a morpheme that are actually observed.

For example, in Table 3.7, we observed that the basic form of the verbs must be those that end in a consonant (in at least these four cases), as they do in the passive and the gerund. Thus, the underlying forms of 'to catch', 'to follow', 'to point out', and 'to know' would be /hopuk/, /arum/, /tohuŋ/, and /maatur/, respectively.

3.4.2 Phonological rewrite rules

Phonological alternations have been formalized, since the 1960s, as sequences of CONTEXT DEPENDENT REWRITE RULES in a Post production system[11]. These rules function largely as REGULAR EXPRESSION replacements. They replace one set of substrings with another string or set of substring. For example, a rule may replace a [z] with an [s] when it occurs after voiceless consonants word finally. Such a rule, for English, might be written as:

[11] Halle, M. (1962). Phonology in generative grammar. *Word 18*(1–3), 54–72; and Chomsky, N. and M. Halle (1968). *The sound pattern of English*. Studies in language. New York: Harper & Row

(1) $z \rightarrow s \,/ \left\{ \begin{matrix} p \\ t \\ k \\ f \\ \theta \\ s \\ \int \\ \widehat{t\int} \end{matrix} \right\} \_ \#$

This rule says: replace z with s when it occurs in the context between a member of the set {p, t, k, f, θ, s, ʃ} and a word boundary (represented by #). Likewise, a rule that inserts a ə between two neighboring sibilants could be represented as:

(2) $0 \rightarrow ə \,/ \left\{ \begin{matrix} s \\ z \\ \int \\ ʒ \\ \widehat{t\int} \\ \widehat{dʒ} \end{matrix} \right\} \_ \left\{ \begin{matrix} s \\ z \\ \int \\ ʒ \\ \widehat{t\int} \\ \widehat{dʒ} \end{matrix} \right\}$

This rule says: replace ϵ (the empty string) with ə when sibilants are in the left and the right context.

3.4.3 Rule ordering

The order in which rules apply is significant[12]. Consider the rules that induce allomorphy into the form of the English plural. These rules are given in (3):

[12] In finite state terms, the order in which FSTs are composed is significant

(3) a. **Voicing assimilation**

$$z \rightarrow s \,/ \left\{ \begin{matrix} p \\ t \\ k \\ f \\ \theta \\ s \\ \int \\ \widehat{t\int} \end{matrix} \right\} \_ \#$$

b. **Epenthesis**

$$0 \rightarrow ə \,/ \left\{ \begin{matrix} s \\ z \\ \int \\ ʒ \\ \widehat{t\int} \\ \widehat{dʒ} \end{matrix} \right\} \_ \left\{ \begin{matrix} s \\ z \\ \int \\ ʒ \\ \widehat{t\int} \\ \widehat{dʒ} \end{matrix} \right\}$$

These rules must be ordered so that (3b) applies before (3a). Otherwise, *watches* /wɑ$\widehat{t\int}$+z/ would yield [wɑ$\widehat{t\int}$əs] rather than [wɑ$\widehat{t\int}$əz] (the /z/ would

assimilate to the voiceless /t͡ʃ/ before the schwa [ə] was inserted. This type of counterfactual reasoning is essential to determining the correct ordering of rules, as you will see in your homework assignment for this unit.

There are four kinds of interactions between two rules that show them to be CRUCIALLY ORDERED (that their order matters):

feeding Rule A creates an environment where Rule B can apply

bleeding Rule A destroys and environment where Rule A would otherwise apply

counter-feeding Rule B would feed Rule A if their relative orders were reversed

counter-bleeding Rule B would bleed Rule A if their relative orders were reversed.

The ordering of the rules in the English *-s* example above is BLEEDING. Epenthesis breaks up clusters of two sibilants where assimilation would otherwise occur. This is different from counter-feeding. Consider the words *latter* and *ladder* in American English. These two words are not pronounced the same, by many speakers, but the consonants are the same because the orthographic *tt* and *dd* are realized as a single sound, the flap [ɾ]. What is different, then? The ⟨a⟩ vowel is short [æ] in *latter* but long [æː] in *ladder*. In other words, there are two phonological rules:

flapping /ipat/ and /d/ become /ɾ/ between a stressed and unstressed vowel

lengthening vowels are lengthened before voiced consonants

In dialects where *latter* and *ladder* are distinct, lengthening must apply before flapping. Otherwise, the difference between the two words would be obliterated (the medial consonants in both words would become voiced /ɾ/) before lengthening could apply.

3.5 Rewrite rules as finite-state transducers

3.5.1 Introduction to FSTs

A finite state transducer (FST) is a type of finite state machine (an abstract machine with a finite number of states and a finite number of transitions between these states). It differs from a finite state automaton, or FSA. An FSA has only one memory tape; an FST has two tapes: an input tape and an output tape. It can be thought of as a "translator" between input strings and output strings (rather than just an ACCEPTOR).

Formally, an FST is a 6-tuple $(Q, \Sigma, \Gamma, I, F, \delta)$ such that:

- Q is a finite set of states;

- Σ is a finite set of symbols, the *input alphabet*;
- Γ is a finite set of symbols, the *output alphabet*;
- I is a subset of Q, the set of *initial states*;
- F is a subset of Q, the set of *final states* or *acceptor states*;
- $\delta \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times (\Gamma \cup \{\epsilon\}) \times Q$ (where ϵ is the empty string), the *transition relation*.

In practice, this means that every transition is decorated with two labels: an input label drawn from the input alphabet and an output label drawn from the output alphabet. FSTs are often represented graphically. An illustration is given in Figure 3.2. This FST takes as input a string that consists of any

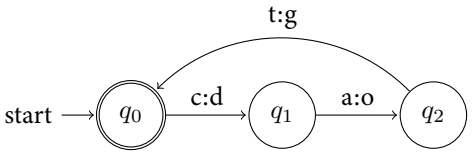


Figure 3.2: Simple FST

number of concatenated repetitions of "cat" and outputs a string with the same number of repetitions of "dog." It works in the following fashion: We start at the initial state, q_0. The we also start at the beginning of the input string. We can traverse a transition (an arc) when the symbol on the input side of the transition label is the next symbol on the input tape. When we traverse the transition, we write the symbol on the output label to the output tape. But a path through an FST is only valid if one is at a final state when the input tape is entirely consumed. That, the FST in Figure 3.2 would not produce any output for the inputs "c", "ca", or "catc".

It is easy to get confused about FSTs if you understand them procedurally. In fact, an FST is a compact description of a RELATION. Such a relation can be one-to-many. Thus, when it appears that you could take more than one path in an FST, is most correct to assume that *all* of those paths were taken. Perhaps the best analogy is to depth-first search. You are scanning through the FST looking for all of the paths that satisfy the necessary conditions. There is an output mapping for each such path.

3.5.2 Operations on FSTs

FSTs can be operated on in various useful ways. First of all, they can be inverted. For example, the inversion of Figure 3.2 would transduce "dogdog" as "catcat". This operation simple involves swapping the labels on all transitions. This is useful because it is possible to write an FST that transduces in one direction ("down") and easily produce an FST that transduces in the opposite direction ("up").

Secondly, two FSTs can be composed. COMPOSITION create a new FST that produces the same mappings as passing an input to the first FST, collecting the output from the first FST and passing it to the second FST. When an FST is applied to a string, it is actually through composition. The string is converted into a LINEAR CHAIN AUTOMATON, a chain of states where each symbol in the string has a corresponding transition, and the FST to be applied is composed with this automaton. The paths through the resulting lattice indicate the outputs of the FST. However, composition has many uses. It is not uncommon to break a task up into composable units, define them individually, then combine them into a larger FST through composition.

FSTs can also be concatenated. Consider the FST in Figure 3.2 and its inverted form. If these two were concatenated together, the resulting FST would transduce "catcatcatdog" as "dogdogdogcat".

Finally, it is possible to take the union of two FSTs, so that any mapping that is generated by either FST is generated by the new transducer. If you took the same two transducers as in the concatenation example above and instead took the union, the new transducer would map "catcat" to "dogdog" and "dogdogdog" to "catcatcat".

FSTs are closed[13] under inversion, composition, concatenation, and union. They are not closed under INTERSECTION. A transducer produced via intersection only generates the mappings that are generated by both intersected transducers. This sometimes results in relations that cannot be generated with the finite-state formalism. Some FST toolkits nevertheless allow you to intersect transducers, so it is useful to know about this operation.

[13] To be closed, in the context, means that—given an FST as an input—the operation will always produce an FST as its output.

3.5.3 *Johnson's discovery*

In his dissertation, C. Douglas Johnson reported an interesting discovery: most phonological rules were equivalent to finite state transducers [14]. The exceptions were rules that were self-feeding or self-bleeding. Self feeding rules had been proposed to account for some rules that seemed to apply iteratively (vowel harmony and tone spreading rules, in particular). However, by the 1970s, it because evident that these phenomena were better analyzed in another way. This means that basically all widely agreed-upon rule-based phonological analyses can be restated in finite state terms.

[14] Johnson, C. D. (1972). *Formal Aspects of Phonological Description*. The Hague: Mouton

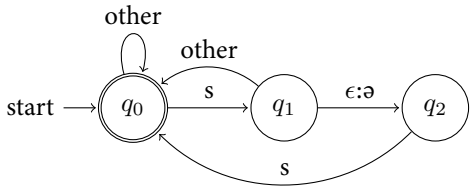


Figure 3.3: FST that epenthesizes /ə/ between two /s/s

3.5.4 Implementing phonological rewrite rules in XFST/Foma

Foma is an interpreter for grammatical notations that provides a high-level interface to FSTs. It is a clone of an earlier piece of software called XFST, which was produced by Xerox and is subject to onerous licensing requirements. Foma is open source software and it presents a more pleasant user interface.

XFST/Foma is a complete FST library that is suitable for writing tokenizers, phonological models, and morphological analyzers, among other things. Here we will be concerned mostly with using Foma to write phonological rules. The rules that we will write have the following form:

```
A -> B || L _ R
```

In this forumla `A`, `B`, `L`, and `R` are regular expressions of arbitrary complexity. It means, "rewrite substrings in `A` as substrings in `B` in the context between substrings in `L` and substrings in `R`. For example, a rule stating that /z/ is devoiced to /s/ between a voiceless consonant and the end of a word might look like this (for English):

```
z -> s || [p | t | k | f | θ | s | ʃ] _ .#.
```

In this rule, `[p | t | k | f | θ | s | ʃ]` is the union of regular expressions for each of the voiceless sounds. The square brackets make the enclosed regular expression into a group. The symbol `.#.` matches at word boundaries (at the beginning and the end of a string). It is equivalent to `#` in standard phonological rule notation.

We cannot enter a rule like this directly into Foma. Instead, we need to use one of two statements. The first is `read regex`. This statement compiles the following rule/regular expression and adds the resulting FST to the stack[15]. The second is `define`. It takes a name and a rule/regular expression and compiles the regular expression, assigning it to the name (as a kind of constant). These are illustrated in the following example:

[15] Foma (like XFST) is stack-based. Interactions with the interpreter, whether directly or via scripts, pushes FSTs on the stack, pops FSTs from the stack, reorders the fact, etc.

```
define Sibilant s | z | ʃ | ʒ | t ͡ ʃ | d ͡ ʒ;
define Voiceless p | t | k | f | θ | s | ʃ;
define Epenthesis [..] -> ə || Sibilant _ Sibilant;
define Devoicing z -> s || Voiceless _ .#.;
read regex Epenthesis .o. Devoicing;
```

This script defines `Sibilant` as a regular expression matching any of the sibilant consonants of English. It then defines `Voiceless` as a regular expression matching any of the voiceless (consonants) of English. Then it uses these two constants to define FSTs that implement rewrite rules. The fourth line states that the empty string is rewritten as ə between two sibilants. It uses the special notation `[]` for the empty string (ϵ); there is another notation for the empty string, `0`, that cannot be used here because it would trigger a potentially infinite number of insertions.

We have two choices when interacting with the Foma interpreter: we can either issue commands through an interactive session (initiated by typing `foma` at the command line) or we may pass a script to Foma (`foma -l myscript.xfst` to run a script and keep the interactive session open or `foma -f myscript.xfst` to run a script, then exit). If you are running the script, then exiting, you likely want to add the `save stack` command (no semicolon) to the end of your script. This will save the stack as a set of FSTs in binary format that can then be used with the `flookup` utility.

The `flookup` utility takes a binary stack as an argument and reads a list of words (one per line) from STDIN. It outputs the result of applying the topmost FST to each word in an upward direction (you can apply the FST in a downward direction instead by using the `-i` switch).

3.5.5 Example: Catalan

Consider the following data from Catalan, a Romance language of Spain (or Catalonia, depending on whom you ask):

MASC SG	FEM SG		MASC SG	FEM SG	
əkelʲ	əkelʲə	'that'	mal	malə	'bad'
siβil	siβilə	'civil'	əskerp	əskerpə	'shy'
ʃop	ʃopə	'drenched'	sɛk	sɛkə	'dry'
əspɛs	əspɛsə	'thick'	gros	grosə	'large'
baʃ	baʃə	'short'	koʃ	koʃə	'lame'
tot	totə	'all'	brut	brutə	'dirty'
pɔk	pɔkə	'little'	prəsis	prəsizə	'precise'
frənses	frənsezə	'French'	gris	grizə	'grey'
kəzat	kəzaðə	'married'	bwit	bwiðə	'empty'
rɔt͡ʃ	rɔʒə	'red'	bot͡ʃ	boʒə	'crazy'
orp	orβə	'blind'	lʲark	lʲarɣə	'long'
sek	seɣə	'blind'	fəʃuk	fəʃuɣə	'heavy'
grok	groɣə	'yellow'	puruk	puruɣə	'fearful'
kandit	kandiðə	'candid'	frɛt	frɛðə	'cold'
səɣu	səɣurə	'sure'	du	durə	'hard'
səɣəðo	səɣəðorə	'reaper'	kla	klarə	'clear'
nu	nuə	'nude'	kru	kruə	'raw'
flɔɲd͡ʒu	flɔɲd͡ʒə	'soft'	dropu	dropə	'lazy'
əgzaktə	əgzaktə	'exact'	əlβi	əlβinə	'albino'
sa	sanə	'healthy'	pla	planə	'level'
bo	bonə	'good'	sərɛ	sərɛnə	'calm'
suβlim	suβlimə	'sublime'	al	altə	'tall'
fɔr	fɔrtə	'strong'	kur	kurtə	'short'
sor	sorðə	'deaf'	bɛr	bɛrðə	'green'
san	santə	'saint'	kəlɛn	kəlɛntə	'hot'
prufun	prufundə	'deep'	fəkun	fəkundə	'fertile'
dəsen	dəsentə	'decent'	dulen	dulentə	'bad'
əstuðian	əstuðiantə	'student'	blaŋ	blaŋkə	'white'

In the first few lines, there are no morphophonemic alternations. The root

in the masculine forms is the same as a the root in the feminine forms. This means that the surface representation is the same as the underlying representation. This, for 'short', the UR of the masculine is /baʃ/ and the UR of the feminine is /baʃə/ (or /baʃ+ə/ if we use + to represent morpheme boundaries).

When we reach 'French' and 'grey', however, the situation becomes more complicated: compare [frənses] 'French.MASC.SG' with [frənsezə] 'French.FEM.SG'. Here, [s] alternates with [z]. It makes sense that these sounds should be related to one another—they differ only in voicing. The question is: which is underlying and which is derived by rule. To answer this question, let us refer back to 'large' ([gros] and [grosə]). Here, there is no alternation, so we can confidently say that the URs of these two word forms have /s/. Since we do have an alternation in 'French' and 'grey', the UR of the root in this word must end in a /z/. There must be a rule, then, that transforms word-final /z/ to [s]. As we will see, this same rule is operative elsewhere.

Many people find this counterintuitive. After all, they ask, shouldn't the root form (the masculine is just a bare root) be the same as the underlying form? On an empirical basis, the answer is quite simply "no." The optimal normalization of a morpheme does not follow from its position in a paradigm (whether it is masculine or feminine, nominative or accusative, infinite or finite form). The underlying representation is purely a function of phonology.

Stepping back, let's look at the distribution of some sounds in Catalan. In particular, let's compare the voiced plosives [b d g] with the voiced fricatives [β ð ɣ]. Are they in contrastive or complementary distribution? A quick survey will show that the plosives occur at the beginnings of words and after consonants while the fricatives occur after vowels. This means that they are in complementary distribution, that their distribution can be accounted for by a rule, and that they likely share a common underlying representation. A similar relation seems to hold between the voiced affricate [d͡ʒ] and the voiced fricative [ʒ].

Consider, now, the word for 'married'. In this word, [t] alternates with [ð]. These sounds differ in two ways:

- [t] is a plosive while [ð] is a fricative
- [t] is voiceless while [ð] is voiced

We have already proposed, but have not formalized, a rule for devoicing consonants at the ends of words (devoicing). This gets us halfway to our goal. We have also suggested that [ð] has the same underlying representation as [d]. That means we must have a rule, for example, that changes voiced stops to fricatives after vowels (spirantization). In other words, the best UR for 'married.MASC.SG' is neither /kəzat/ nor /kəzað/ but /kəzad/. In what order to they apply? If spirantization applied first, we would expect 'married.MASC.SG' to be [kəzaθ] instead of the observed [kəzat], but we get the right result if we adopt the bleeding order: devoicing before spirantization.

How do we formalize all of this in terms of Foma? Let's start with the natural classes:

```
define Vowel a | e | i | o | u;
```

This allows us to write the following rules[16]:

[16] Note that the comma between rewrites means that they occur simultaneously.

```
define Devoicing b -> p,
                 d -> t,
                 g -> k,
                 z -> s,
                 d ‿ ʒ -> t ‿ ʃ || _ .#. ;
define Spirantization b -> β,
                      d -> ð,
                      g -> ɣ,
                      d ‿ ʒ -> ʒ || Vowel _ ;
read regex Devoicing .o. Spirantization ;
```

3.6 Constraint-based approaches to phonology

So far, we have looked at phonology derivationally—underlying representations are rewritten by a succession of rules to derive surface representations. Alternatively, we can obtain the range of underlying representations for a given surface representation by running the cascade "backwards" or by applying the corresponding FSTs "up" instead of "down". This is the most common approach to phonology in actual computational implementations. It is simple and computationally tractable. However, there is another approach that can provide considerable insight: constraint-based (rather than rule-based) approaches to phonology.

3.6.1 Feature Theory

So far, we have treated segments as atomic symbols. However, we have seen cases where certain symbols pattern together based upon shared phonetic properties. For example, in Catalan all of the voiced plosives (/b d g/ devoiced word-finally (becomeing [p t k] and spirantized (became fricatives) when preceded by a vowel and followed by a vowel or approximate. /b d g/ all have things in common (they are plosives and they are voiced) and [p t k] all likewise have things in common (they are plosives and they are voiceless). Both of these groupings form what are called NATURAL CLASSES—sets of segments that share both phonetic properties and phonological behaviors.

Feature theory proposes that segments can be decomposed into vectors of binary features, each dimension corresponding to membership in a natural class. These vectors are typically held to have 20 or so dimensions.

	i	y	ɨ	u	e	ø	ʌ	o	æ	œ	ɑ	ɒ
high	+	+	+	+	−	−	−	−	−	−	−	−
low	−	−	−	−	−	−	−	−	+	+	+	+
back	−	−	+	+	−	−	+	+	−	−	+	+
round	−	+	−	+	−	+	−	+	−	+	−	+

Table 3.9: Four place features for vowels

	p	t̪	t	t͡ʃ	ʈ	c	k	q	ʕ	ʔ
anterior	+	+	+	−	−	−	−	−	−	−
coronal	−	+	+	+	+	−	−	−	−	−
distributed		+	−	+	−					
high	−	−	−	−	−	+	+	−	−	−
back	−	−	−	−	−	−	+	+	+	−
low	−	−	−	−	−	−	−	−	+	−

Table 3.10: Five place of articulation features for consonants

3.6.2 Phonology as typology

One goal of phonological theory, dating back to the 19th century, has been to characterize the range of variation and degree of similarities among the sound structures of languages. Phonologists have been interested in questions like the following:

- Do all languages have vowels?
- What is the minimum number of vowel phonemes a language can have?
- What is the maximum number of vowel phonemes a language can have?
- Do languages that have more vowel phonemes tend to have fewer consonant phonemes?
- Why are some vowel systems (/a i u/ and /a e i o u/) much more common than the others?

These questions concern phonological inventories. They have also been interested in questions about phonotactics and alternations:

- What kinds of syllables are most common (V, CV, CVC, CCV, etc., where C = consonant and V = vowel)
- In what contexts does vowel deletion occur?
- Do alternations fix bad syllables or other phonotactically suboptimal structures?

This general line of research is called TYPOLOGY. Doing typology with ordered rules proved to be rather unwieldy. In order to capture cross-linguistic generalizations and typological patterns, phonologists started encoding more and more of the sound structure of languages in static constraints—statements of what was allowed (or what was favored).

3.6.3 *Introducing Optimality Theory*

The most fully-explored constraint-based approach to phonology is called Optimality Theory. It is a theory of ranked constraints. It proposes that the grammar—the component that relates underlying representations to surface representations—consists of a strictly ordered hierarchy of universal but violable phonological constraints. For a given input (the underlying representation) the phonology generates a set of output candidates[17]. The actual output is the candidate that violates the least-high ranked constraint.

[17] We will discuss the component that generates these candidates, GEN, in a moment

Take a simple example from earlier in this chapter: Maori /hopuk/ 'catch' → [hopu]. Compare this with English /bik/ 'beak' → [bik]. We can view these mappings as the result of competition between two universal constraints. One says that there should not be consonants at the ends of syllables; we will call this constraint NOCODA. The other says that you should not delete segments; we will call this constraint MAX. Here is what we see for Maori[18].

[18] Note that this is greatly simplified, both in terms of the number of constraints and the number and diversity of candidates

/hopuk/	NOCODA	MAX
☞ a. hopu		*
b. hopuk	*	

The candidates are in the rows and the constraints are in the columns. Asterisks represent violations of constraints. Was a constraint violated more than once by a candidate, there would be more than one asterisk. Candidate (a), [hopu] wins even though it violates DEP because its competitor ([hopuk]) violates the higher ranked NOCODA. The ranking of these constraints is different in English:

/bik/	MAX	NOCODA
a. bi	*	
☞ b. bik		*

In Optimality Theory terms, English keeps codas like the [k] in *beak* not because it "likes" codas, but because it hates deleting segments more. In other words, it ranks the universal constraint MAX above the universal constraint NOCODA. However, English phonology may "repair" violations of NOCODA in other ways that do not involve deletion (such as in decisions about how to break a word up into syllables).

Where do the candidates come from? In some formulations of the theory, the candidate space is essentially the set of all strings (for a particular alphabet). This is not computationally tractable, and is also not suitable for pen-and-paper calculations. Another way of looking at this problem is to view the candidate set in terms of repairs—a constraint like NOCODA is violated by the input; what are ways of changing the input to satisfy this constraint? For each of these repairs, include a candidate.

3.6.4 Correspondence Theory

The most widely accepted version of Optimality Theory is called Correspondence theory. This theory is based upon a correspondence relation between the segments (or other elements) in the input and those in the output candidates. We can indicate segments that are in correspondence with subscripts. Drawing on the Maori example above, /$h_1o_2p_3u_4k_5$/ → [$h_1o_2p_3u_4$] (that is, every segment in the input corresponds to an identical segment in the output except for k_5, which has no correspondent in the output. That is what a MAX violation looks like in Correspondence Theory: it is a segment in the input with no corresponding segment in the output candidate. MAX is just one of a class of constraints that refer to these correspondence relations. These are called FAITHFULNESS constraints. To a first approximation, faithfulness constraints enforce similarity between inputs and output candidates. Here are some of the most important families of faithfulness constraints:

- INDENT Corresponding segments are identical in input and output. Do not change /d/ to [t], etc.
- MAX Segments in the input have corresponding segments in the output. Do no delete.
- DEP Segments in the output have corresponding segments in the input. Do not insert.

Each of these families of constraints consists of many more specific constraints that can be freely ranked relative to one another and other constraints.

Faithfulness constraints compete with one another and with another class of constraints, MARKEDNESS constraints. These constraints penalize dispreferred structures in output candidates. NOCODA is an example of such a constraint. Coda consonants—consonants at the ends of syllables—are universally disfavored and many languages have alternations that get rid of them, either through deleting them or by enforcing a particular pattern of syllabification. There are many other proposed markedness constraints. One example is *COMPLEX. This is a constraint against having more than one consonant in the ONSET or coda of a syllable. Another is ONSET, which states that syllables start with consonants. The AGREE family of constraints enforces similarity between segments in an output. For example, on such constraint might require neighboring consonants to share the same place of articulation or one vowel to share the same height, rounding, or backness as the preceding vowel.

For those who have a background in language technologies, correspondence theory may have a more than passing resemblance to the noisy channel model. Imagine a channel that factors all of the predictable information out of a surface representation, yielding an underlying representation. Decoding

means going from the UR to the surface representation or, in other words, mapping from input to output. Markedness tells us what a probable surface form is like. In other words, it is a source model, rather like a language model. Our channel model is faithfulness—it tells us how probable a given output-input mapping is. Of course, Optimality Theory does not speak probability, so the analogy breaks down at some level. A related grammatical formalism, Harmonic Grammar, uses the same type of constraints but is more explicitly probabilistic and presents an even clearer analog to the noisy model.

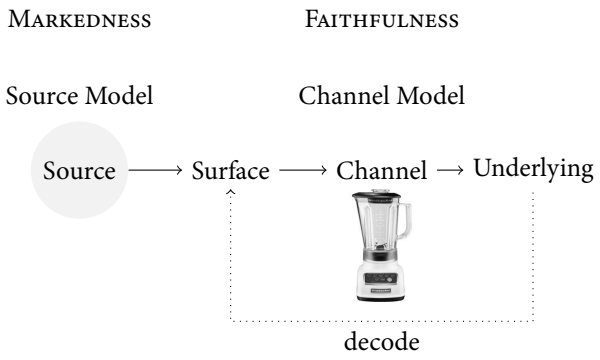


Figure 3.4: Correspondence Theory and the noisy channel model

3.7 A digression: orthography

Engineers sometimes encounter language as sound, when they are working with speech. In these cases, models that incorporate phonetic and phonological knowledge are appropriate. However, many other engineers of language technologies encounter only text, and this tends to be written in ORTHOGRAPHY, that is to say, a conventional writing system used by the speakers of a language. Orthography often reflects, and sometimes does not reflect, the phonology of a language. In order to understand the issues involved in working with orthographic representations, it will be useful to survey the types of writing systems that exist (and are in use) as well as the relationship between writing systems and phonological representations.

3.7.1 Types of writing systems

The most familiar type of writing system for many of those who will be reading this book is the ALPHABETIC. In their ideal form, alphabets have one symbol (letter) for each phoneme in the language, whether it is a vowel or consonant. In other words, there is a one-to-one mapping between logical sound and written symbol. Actual alphabets diverge from this ideal is a number of ways. For example, in the English word *laugh*, there are five letters but only three phonemes (/l/, /æ/, and /f/). English, in fact, is notorious for this type of divergence and shows a highly irregular correspondence between writing and sound. Even when the correspondences are somewhat

predictable, English makes extensive use of DIGRAPHS, sequences of two letters that stand for one sound. For example, the *sh* in SHIBBOLETH stands for a single sound (/ʃ/). Many alphabets employ digraphs and trigraphs to some degree, but in the typical alphabetic orthography, there is a predictable mapping between written form and pronunciation. An alphabet represents all sounds and places consonants and vowels on an equal footing.

Some older types of orthographies treat vowels differently. In ABJADS, some vowels (especially short vowels) are simply not represented. In Arabic, the orthography only represents the consonants and the long vowels. This allows for a great number of ambiguities. For example كتب *ktb* could represent *kataba* 'he wrote', *kattaba* 'he made somebody write', *kutiba* 'it was written', or *kutub* 'books'. In ABUGIDAS like Devanagari or the Ethiopic scripts, each "consonant" character actually represents a sequence of a consonant and a default vowel (the "inherent" vowel). If the consonant is written with a vowel diacritic, that vowel "replaces" the default vowel. In some such scripts (e.g. Ethiopic), there is no orthographic distinction between a word-final consonant or a consonant followed by a word-final inherent vowel. In other scripts, there is a "kill" diacritic which indicates that the inherent vowel is absent. In moraic scripts, developing a theme further, characters represent one MORA, which could consist of either one consonant and one (short) vowel, or of a single short vowel. The best-known writing systems of this sort are the Hiragana and Katakana systems used in writing Japanese. Syllabic scripts, by contrast, represent whole syllables with individual characters. These scripts are used to write a number of minority languages including Cherokee (a language of North America) and Yi (a Tibeto-Burman language of China).

Syllabaries are not to be confused with LOGOGRAPHIC writing systems like Hanzi (Chinese characters). In the case of Chinese, a character generally does correspond to a syllable—a morpheme that is one syllable long. However, the characters do not directly indicate the sound of the syllables *per se*. Many different characters may be read the same way (phonologically) and a single character may have more than one reading. What the characters in Chinese really represent are morphemes. The fact that these morphemes are mostly one syllable long is epiphenomenal. It is not true that Chinese characters are completely devoid of phonological content, however. Hanzi typically have two major components: a "radical" and a "phonetic". The radical gives some hint about the meaning of the character (does it relate to water? does it relate to fabric?) while the phonetic provides information about how the character is pronounced (that is, characters that are pronounced similarly often have the same phonetic).

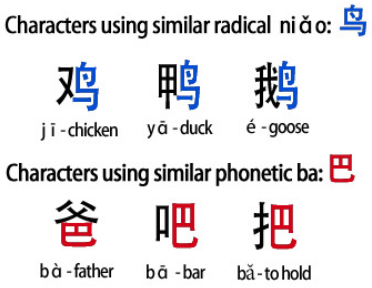


Figure 3.5: Manuscript written in the Yi syllabary

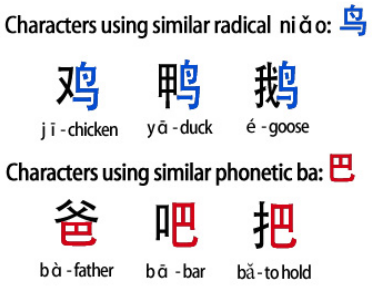


Figure 3.6: Manuscript written in the Yi syllabary

3.7.2 Orthographic depth and phoneme-grapheme correspondence

Scripts vary in how well graphemes correspond to phonemes. Chinese, despite the phonetic patterns in the characters, shows very poor correspondence

between sound and symbol. It is not alone. Abjads are systematically missing certain vowel sounds. Abugidas may make no phonological distinction between certain sequences. Even alphabetic scripts may be highly complicated in this regard, particularly if the orthography was conventionalized before major changes in the lexicon or in the phonology of the language. A prime example of this is English, which has a notoriously inadequate spelling system. It is not alone among languages written with an alphabet, however.

Many languages for which an orthography was recently introduced, or for which there has been a recent spelling reform, have very SHALLOW orthographies. That is, they show very strong phoneme-grapheme correspondence.

For various speech and natural language processing applications, it is useful to transform orthographic representations into phonological representations. For languages with shallow orthographies this is straightforward, even when orthographies are not alphabetic. For languages with deep orthographies, it can be quite difficult and large pronouncing lexicons (along with machine learning techniques) may be necessary.

3.7.3 Orthographies and Unicode

In the dark recesses of the past, there were one or more encodings for every script that could be represented computationally. Even for Latin alphabets, there were a great variety of code pages for different languages and different operating systems. Much of this chaos was born of the need to keep characters to one byte each (or perhaps two bytes). With the emergence and acceptance of Unicode, all of this changed. Now, all of the world's scripts can be represented in a single scheme, along with a collection of random garbage including poop emojis and box-drawing characters. Unicode solves many problems that existed for language scientists and language technologists previously. It raises, however, a number of issues that someone working with Unicode should know about.

Combining characters and presentation forms In Unicode, what appears as a single character may actually be a sequence of Unicode characters (or code points). Unicode has a great number of combining characters, which are rendered in combination with the preceding character(s) to produce a composite symbol. A simple example is that of the acute accent, which exists as a separate character capable of combining with a preceding character to form, for example, *é*, *á*, or *ú*. For each of these letters, it so happens, there is a precomposed form that has a single Unicode code point (one character). These composed forms exist, in part, for compatibility with previous encodings and can be transformed into the decomposed forms illustrated above. Furthermore, the shape of a character may vary based on the characters around which it occurs. This is necessary in cursive scripts like Arabic, where all let-

ters have an isolation form (how they look when there are no other Arabic characters around) and contextual forms (how they look when they connect with other characters to form a word). Ideally, the isolation form and the contextual forms of a letter have the same code point and smart font-rendering technology selects the appropriate form for each context. In practice, Unicode provides a set of "presentation forms" for each letter so that the combining forms can be rendered out of context.

Logical order versus visual order All scripts have a logical order. It corresponds to the order in which words written in the script are pronounced. Sometimes this corresponds to the visual order, whether the script is written left-to-right or right-to-left. But sometimes their are deviations between logical order and visual order. For example, in certain abugidas like the Thai script, diacritics representing vowels pronounced after consonant may be written left of the consonant, to the right of the consonant, above the consonant, or (in two components) on either side of the consonant.

3.8 Exercise: FST for Somali Morphophonology

Consider the following sets of data from Somali, a language of the Horn of Africa:

SG	SG.DEF	PL	GLOSS
daar	daarta	daaro	house
gees	geesta	geeso	side
laf	lafta	lafo	bone
lug	lugta	luɣo	leg
naag	naagta	naaɣo	woman
tib	tibta	tiβo	pestle
sab	sabta	saβo	outcast
bad	bada	baðo	sea
d͡ʒid	d͡ʒida	d͡ʒiðo	person
feeɖ	feeɖa	feezo	rib
ʕiir	ʕiirta	ʕiiro	buttermilk
ʔul	ʔuʃa	ʔulo	stick
bil	biʃa	bilo	month
meel	meeʃa	meelo	place
kaliil	kaliiʃa	kaliilo	summer
najl	najʃa	najlo	female lamb
sun	sunta	sumo	poison
laan	laanta	laamo	branch
sin	sinta	simo	hip
dan	danta	dano	affair
daan	daanta	daano	river bank

SG	SG.DEF	PL	GLOSS
saan	saanta	saano	hide
nirig	nirigta	nirgo	baby female camel
gaβaɖ	gaβaɖa	gabɖo	girl
hoɣol	hoɣoʃa	hoglo	downpour
baɣal	baɣaʃa	baglo	mule
waħar	waħarta	waħaro	female kid
irbad	irbada	irbaðo	needle
kefed	kefeda	kefeðo	pan
d͡ʒilin	d͡ʒilinta	d͡ʒilino	female dwarf
bohol	bohoʃa	boholo	hole
jirid	jirida	jirdo	trunk
ʔaajad	ʔaajada	ʔaajaðo	miracle
gaʕan	gaʕanta	gaʕmo	hand
ʔinan	ʔinanta	ʔinano	daughter

3SG.MASC	3SG.FEM	1PL.PAST	GLOSS
suɣaj	sugtaj	sugnaj	wait
kaβaj	kabtaj	kabnaj	fix
siðaj	sidaj	sidnaj	carry
dilaj	diʃaj	dillaj	kill
ganaj	gantaj	gannaj	aim
tumaj	tuntaj	tunnaj	hammer
argaj	aragtaj	aragnaj	see
gudbaj	guðubtaj	guðubnaj	cross a river
qoslaj	qosoʃaj	qosollaj	laugh
hadlaj	haðaʃaj	haðallaj	talk

Use Foma to construct an analysis of these data in terms of context-dependent rewrite rules. You should submit the following items:

1. A README file with the following items:

   (a) A list of the underlying representations you posit for each root

   (b) A list of the underlying representations for each suffix

   (c) Any special notes necessary to understand your implementation

2. A Foma script (`somali.xsft`) defining an FST that transduces between Somali underlying representations and surface representations (like those in the data tables)

3. Two lists of test-cases (1 test case per line):

(a) Underlying representations to be transduced into surface representations.

(b) Surface representations to be transduced into underlying representations.

4 The Internal Structure of Words

Words are not atoms (in the sense intended by Democritus); they have internal structure. This might not be obvious to speakers of English, where many words consist of a single meaningful unit. To an even greater degree, it may not be obvious to speakers of Chinese—most words do not have much internal structure in Chinese[1]. However, the majority of languages in the world have rich word-internal structure. This internal structure is referred to as MORPHOLOGY.

[1] Exceptions are compounds words, which are quite common in Chinese.

4.1 The morpheme as minimal sign

A fundamental concept in morphology is that of the MORPHEME. It is true that not all theories of morphology are based on morphemes, but it is hard to understand much writing about morphology without knowing what a morpheme is.

A morpheme is a minimal SIGN—a pairing of form and meaning. By form, we mean orthographic, phonological, or phonetic content. For example, the pairing of the meaning 'tree' with the phonemic string /tɹi/ is a morpheme.

Now is a convenient time to introduce another concept that will be useful in our discussion—the notion of a CONSTRUCTION. A construction is essentially a sign, a form-meaning tuple. Unlike a morpheme, though, a construction may have internal structure. A morpheme is a simple pairing of form and content that cannot be reduced to smaller signs. Just as a morpheme is an atomic sign, it is an atomic construction.

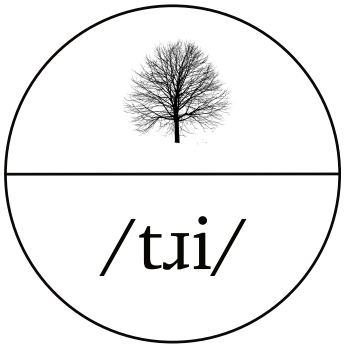


Figure 4.1: The sign pairing the concept TREE with the form /tɹi/.

4.2 Word, lexeme, and listeme

Words are made out of morphemes. It may seem obvious what a word is, but on closer examination, this is hardly the case. Perhaps the most obvious criterion for a word, in languages that delimit words with white space or punctuation, is how string is written. The sentence

(4) the dog's heroic barking frightened the evil cable car.

contains nine words because their are nine white space-delimited units. However, white space is a deceiver. Most tokenizers—computer programs

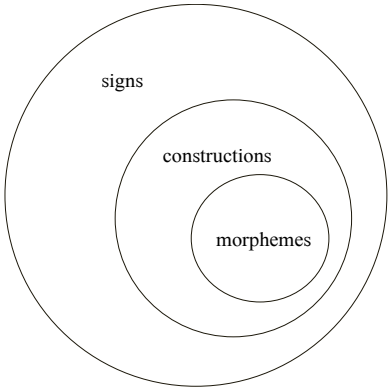


Figure 4.2: Venn diagram of morphemes, constructions, and signs

that segment text into word-sized tokens—would treat the *'s* in *dog's* as a separate word. This is because it is a CLITIC—a syntactically independent word that is phonologically dependent on another word. *Cable car* is another problematic case. It is a noun-noun compound and English is notoriously inconsistent about whether such compounds are written with no delimiter, with a space, or with a hyphen. For example, *mailman* is also a noun-noun compound but it is written as one word. German compounds are written as one word and German thus has a reputation for extremely long words; however, English also has many long words that fly under the radar because they are written as multiple words.

It is also possible to define words phonologically. Under this criterion, clitics like *'s* are not considered independent words because they are phonologically joined to the host word. A more satisfying criterion for wordhood, for our purposes, is that of syntactic word—something is a word if it acts as a unit for the purposes of syntax (the component of grammar that determines how words combine to form phrases and sentences). A useful concept, though, in understanding wordhood is that of the LEXEME. A LEXEME is a unit of contrast (like a morpheme or a phoneme). We can think of a single morphological or syntactic word as a lexeme. Like morphemes and phonemes, a lexeme can have different realizations according to context. For example, the English lexeme WALK has the realization *walked* in past tense clauses, *walks* in non-past clauses with third-person singular subjects, and *walk* elsewhere.

Lexemes are distinct from LISTEMES, which are signs that have to be listed in the LEXICON. Listemes include certain items that are not single words, including multi-word expressions with idiomatic meanings. They also include "templatic" constructions like *the Xer the Xer* (with such instances as *the bigger the better* or *the smaller the faster*). A lexicon that includes these more general constructions, instead of just morphemes and syntactic words, is called a CONSTRUCTICON.

4.3 Morphological form

A variety of different formal operations—the is, operations on the form of a word—exist in morphology. The most common of these are concatenative. They involve the concatenation of two phonological strings. If these strings are STEMS, this operation is called COMPOUNDING. If one of them is an AFFIX, a purely grammatical morpheme that cannot occur in isolation (or, in other words, is a BOUND morpheme) it is called affixation. Affixes that occur before the BASE (the phonological string to which the affix is added) are called PREFIXES and those that occur afterwards are called SUFFIXES. Affixation to both sides of the base simultaneously is called CIRCUMFIXATION.

There is also non-concatentative affixation: phonological strings can be inserted into a base in what is called INFIXATION. Another non-concatentative

process is reduplication, where all or part of the base is repeated. Reduplication may be prefixing, suffixing, or infixing. Not all morphological operations involve adding phonological material, though: APOPHONY or internal change entails "mutating" on of the segments in a base to form a different word or word form (as in English *foot* → *feet*).

In the subsections that follow, we will illustrate some of the principle formal processes with case studies from a few different languages.

4.3.1 *Tagalog*

The Tagalog language—the basis of Filipino, the national language of the Philippines—illustrates a great many formal morphological processes. These include prefixation, suffixation, infixation, and reduplication.

Stem	Singular	Plural	Gloss
laki	malaki	malalaki	'big'
ganda	maganda	magaganda	'beautiful'
bundok	mabundok	mabubundok	'mountainous'

Table 4.1: Morphological processes in Tagalog adjectives

In Table 4.1, The singular is formed with prefixation (of *ma-* and the plural is formed with both prefixation and prefixing reduplication. As illustrated in Figure 4.3, for the plural, the prefix is affixed to a stem that has already undergone reduplication.

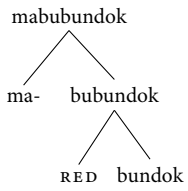


Figure 4.3: The morphological structure of *mabubundok*

Consider the adjective PARADIGMS in Table 4.1. The affixes *-in-* and *-um-* are infixes. They are used to form the perfective aspect forms of verbs depending on their INFLECTIONAL CLASS. The contemplative aspect is sometimes formed by reduplication as in *sulat* 'write' → *su-sulat*. The imperfective may involve both reduplication and infixation: *sulat* → *susulat* → *sumusulat*.

Stem	Perfective	Contemplative	Imperfective	Gloss
tapos	tinapos	tatapusin	tinatapos	'finish'
kain	kumain	kakain	kumakain	'eat'
sulat	sumulat	susulat	sumusulat	'write'
hanap	humanap	hahanap	humahanap	'seek'

Table 4.2: Morphological processes in Tagalog verbs

4.3.2 *German*

German belongs to the West Germanic branch of the Indo-European language family and is thus a close relative of English. While English once had a rich morphology, it has lost most of its morphological complexity. German has, to a great degree, kept this. Consider, for example, the following (partial) verb paradigm:

	Present	Perfect	Preterit
1SG	mache	gemacht	machte
2SG	machst	gemacht	machtest
3SG	macht	gemacht	machte
1PL	machen	gemacht	machten
2PL	macht	gemacht	machtet
3PL	machen	gemacht	machten

Table 4.3: German weak verb: MACHEN 'to make'

The ROOT in this paradigm is *mach*, which does not appear as an independent word (with a meaning like 'make') The person-number forms of the present and the preterit are formed by adding suffixes (one for each combination of person and number). The perfect, though, does not change depending on person and number. Instead, it is always formed with the circumfix *ge-X-t*.

4.3.3 *Tamil*

4.3.4 *Arabic*

Arabic and other Semitic languages (like Hebrew, Aramaic, and Amharic) employ what is called root-and-pattern or templatic morphology. This is illustrated in Table 4.4.

In these languages, roots consist of two to four consonants. Different words and word-forms are made from these roots by inserting vowels and by changing the lengths of the consonants (gemintation). The root, the pattern of specific vowels, and the long-short pattern of vowels and consonants each function like separate "morphemes".

lt

katab-a	'he wrote'
kaataba	'he corresponded'
kutib-a	'it was written'
kitaab	'book'
kutub	'books'
kaatib	'writer; writing'
kuttaab	'writers'
uktub	'write (to a male)!'

Table 4.4: Part of the Arabic paradigm for *ktb* 'with reference to writing'.

4.3.5 *Mandarin Chinese*

Mandarin Chinese has little or no inflectional morphology. One possible exception is the plural suffix 们 *-men*. Examples of using pronouns with *-men* are shown in Table 4.5.

Chinese is, however, rich in compound words (words made by concatenating two or more stems).

我	*wo*	1SG	我们	*women*	1PL
你	*ni*	2SG	你们	*nimen*	2PL
他	*ta*	3SG	他们	*tamen*	3PL

Table 4.5: Paradigms for Chinese pronouns with 们 *-men* 'PL'

客厅	'living room'	沙发	'sofa'	'living room sofa'
眼	'eye'	药	'medicine'	'eye medicine'
马	'horse'	房	'house'	'manger'
雨	'rain'	帽	'hat'	'rain hat'

Table 4.6: Chinese compounds

4.3.6 *Morphological functions*

Talking about formal morphological operations only tells part of the story. Which formal operation is performed, in a given case, is essentially orthogonal to the function of that operation. The two broadest functional categories

of morphology are DERIVATION and INFLECTION.

4.3.7 Derivation

Derivation refers to morphological operations that change the meaning or part of speech (or both) of a word.

4.3.8 Inflection

Inflection refers to morphological operations that add morphosyntactic information to a word. Inflection is essentially additive.

4.4 Patterns of exponence and theories of morphology

One way of looking at morphology is as a mapping between abstract properties, or features, and surface forms. In the most common case, across languages of the world, this is easy: roots correspond to bundles of feature and each affixal morpheme corresponds to a feature. The combination meaning and inflectional properties of the whole word are a function of these parts. This kind of pattern, where there is a one-to-one mapping between forms (affixes) and functions (morphological properties) is called simple or direct EXPONENCE. Many languages have only this pattern of exponence. As we will see, this is characteristic of both ISOLATING and AGGLUTINATING languages. But actual morphologies often deviate from this ideal. One deviation is called CUMULATAIVE EXPONENCE. This is when a single MORPH realizes multiple different morphological features. A good example of this is the English suffix *-s* in verbs. It simultaneously realizes the features [non-past], [singular and [3rd person]. It is as if all these features have "accumulated" in one morpheme. However, this contradicts, to some extent, the definition of morpheme: a morpheme is supposed to be a pairing of a form with one meaning or piece of content. Here, it is paired with three units of content, which are not necessarily related to one another. Cumulative exponence is illustrated more dramatically in 4.7, where the stem *amic-* combines with different DESINENCES[2] that cannot be straightforwardly be segmented into case and number morphemes.

Other modes of exponence are VACUOUS EXPONENCE, which an exponent appears, but does not mark any morphosyntactic property, NULL EXPONENCE, where there is a derivational or inflection change, but no formal change, and EXTENDED EXPONENCE, where exponence of a property is "spread out" over multiple formal operations. These are illustrated in Figure 4.4. This figure shows each of the modes of exponence and how properties relate to operation in such modes. For example, extended exponence displays a many-to-one relationship between formal operations and morphosyntactic properties,

[2] A DESINCENCE is an inflectional ending, as is found in many European languages.

	SG	PL
NOM	amīca	amīcae
VOC	amīca	amīcae
ACC	amīcam	amīcās
GEN	amīcae	amīcārum
DAT	amīcae	amīcīs
ABL	amīcā	amīcīs

Table 4.7: The declension of Latin *amīca* '(girl) friend.'

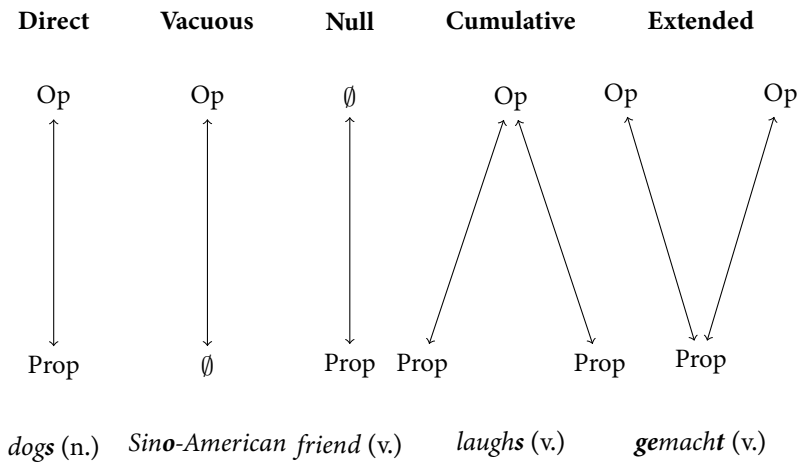


Figure 4.4: Modes of exponence

The American Structuralist linguist Charles Hockett characterized three families of morphological theories, and attempted to weigh their merits [3]. Various versions of each of these theories are still important in the language sciences today.

[3] Hockett, C. F. (1954). Two models of grammatical description. *Word 10*(2-3), 210–234

4.4.1 Item and arrangement morphology

Item and arrangement (IA) morphology treats all morphology as concatenation (arrangement) of morphemes (items). It treats each morpheme as a minimal pairing of form and meaning and specifies how these may be concatenated to derive and inflect words. Because IA morphology is simple, it is the model we assumed when doing phonological analysis. It is also the framework that is used for most computational work in morphology because it is simple to implement with FSTs, sequence-to-sequence models, and other computational models. That being said, it accounts for some facts of morphology only awkwardly. For this reason, their are a number of alternative frameworks.

4.4.2 Item and process morphology

Item and process (IP) morphology allows for two types of morphological operation: items (roots) and processes. Processes might include operations like prefixation, suffixation, infixation, reduplication, apophony, and templatic morphology. In a sense, both of these operations are morphemes. The form of these morphemes is a function, rather than a string:

- A **root** is a function from the empty string ϵ to a string[4].
- A **prefix**, **suffix**, or **circumfix** is a function from strings (bases) to strings of which they are substrings.

[4] To simplify, we have assumed that linguistic forms are strings. In fact, this is likely not the case, but it is a useful simplifying assumption.

- A **process** (more generally) is a function from strings (bases) to related string.

IP morphology is more expressive, but computational linguists have not developed good ways of building morphgological parsers/analyzers using IP morphology.

4.4.3 Word and paradigm morphology

A third model of morphology dispenses with morphemes altogether. Rather than treating word-forms as concatenations of items, it treats them as cells in a paradigm which are formed by analogical rules from words in other cells. A similar model is assumed in various paradigm completion tasks in natural language processing. Word and paradigm morphology is particularly insightful, however, when dealing with languages with indirect patterns of exponence.

4.4.4 Construction morphology

One theory of morphology that differs from the three classical morphological theories is Construction Morphology (CxM)[5]. Construction Morphology has at its heart the constructional gestalt—the notion that, rather than being COMPOSITIONAL, morphological constructions may have more content than the sum of their parts. In essence, it means that the way in which words are put together can be meaningful in an of itself, independent of the components that are combined. In this way, Construction Morphology forms a bridge between morpheme-based and word-based theories of morphology.

[5] To be perfectly accurate, Construction Morphology is a type of Word and Paradigm morphology, but it has certain properties that differentiate it from classical WP theories.

Patterns of compounding may provide the clearest illustration of morphological constructions. Consider the two sets of Chinese noun-noun compounds in Tables 4.8 and 4.9: Both sets of examples actually illustrate a number of different semantic relationships between the compounded stems (all of which consist of two compounded nouns). However, there are two broad patterns that differentiate the two sets. In Table 4.8, the first noun modifies the second noun in some way. In Table 4.9, the meaning of the compound

田鼠	*tianshu*	field-mouse	field mouse
唇膏	*chungao*	lip-ointment	lipstick
炮弹	*paodan*	artillery-bullet	artillery shell
书包	*shubao*	book-container	satchel

Table 4.8: Subordinate compounds in Chinese

父母	*fumu*	father-mother	'parents'
花木	*huashu*	flower-tree	'vegetation'
天地	*tiandi*	heaven-earth	'universe'
国家	*guojia*	country-home	'nation'
風水	*fengshui*	wind-water	'geomancy'

Table 4.9: Coordinate comopunds in Chinese

as a whole concerns both nouns symmetrically. We could express this more formally using the notation introduced by Geert Booij[6]:

(5) a. $[[x]_{N_i}[y]_{N_j}]_N \leftrightarrow$ 'a sort of SEM_j with properties of SEM_i'

b. $[[x]_{N_i}[y]_{N_j}]_N \leftrightarrow$ 'the union of SEM_i and SEM_j'

[6] Booij, G. (2010). *Construction Morphology*. Oxford: Oxford University Press

These types share something in common (the concatenation of two nouns)[7] We may say that they INHERIT from a common prototype certain properties. At the same time, each of these constructional schemas admits further specification. For example, among the subordinate compounds in Table 4.8, there are compounds where the modifier is a place where something is found ('field mouse'), a place where something is applied ('lipstick'), a type of item place in a container ('satchel'), and so on. It would be possible to write constructional schemas like those in (5) for each of these types. In Construction Morphology, we would view those more specific schemas as inheriting defaults from the more general schemas in (5).

[7] Though a careful examination will show that both types of compounds exist with other parts of speech.

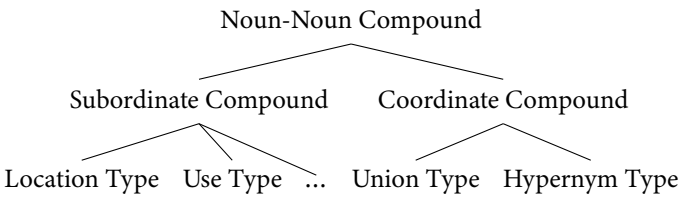


Figure 4.5: Parital inheritance hierarchy for Chinese compounding constructions

Construction Morphology is not just for compounding. It is straightforward enough to capture derivation and inflection in this framework as well. For example, the English suffix *-er* could be modelled as follows:

(6) $[[x]_{V_i}\text{er}]_N \leftrightarrow$ 'one who by profession or habit does SEM_i'

Note that it is not the case the *-er* means 'by profession or habit does X', but that strings consisting of a verb and *-er* have the meaning 'one who by profession or habit does X'. This may seem like a subtle distinction, but it is essential to the whole architecture of Construction Morphology. Affixes are not morphemes; they are only meaningful insofar as they participate in constructions.

One advantage of Construction Morphology is the clean way it allows us to model the diachronic (historical) transition between compounding and derivation. Derivational affixes often begin their lives as the HEADS of subordinating compounds. When a large number of compounds are formed with the same head, the meaning of the head may become more abstract and may start to function as a derivational affix. This process can be modelled as a transition from a general compounding construction to a specific derivational construction. Consider the case of Hmong, a language of China and Southeast Asia. In Hmong, there are compounds where verbs (including STATIVE VERBS) modify preceding nouns. Examples are given in Table 4.10: Some compounds that, historically speaking, were identical to these, have become NOMINALIZATIONS: abstract nominalizations that derive abstract nouns

tsuv-dlub	tiger-be black	'panther'
taum-mog	bean-be soft	'pea'
toj-sab	hill-be tall	'highlands'
nteeg-tuag	funeral-die	'funeral'
khoom-muag	goods-sell	'goods for sale'

Table 4.10: N-V compounds in Hmong

from verbs and nominalizations that derive a human noun that relates to the action from a verb. These are illustrated in Table 4.11: These can be grouped,

kev-noj	way-eat	'eating'
kev-haus	way-drink	'drinking'
kev-kaaj	way-bright	'brightness'
kev-zoo	way-good	'goodness'
kev-phem	way-bad	'evil'
kev-kawm	way-study	'studying'
tub-txib	son-send	'messenger'
tub-khaiv	son-send	'servant'
tub-nyag	son-steal	'thief'
tub-ncig	son-be_around	'people who handle duties at funeral'

Table 4.11: Nominalizations in Hmong

insightfully, into an inheritance hierarchy, as in Figure 4.6: This process,

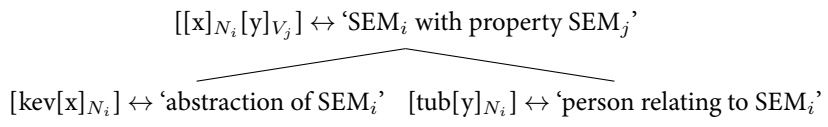


Figure 4.6: Inheritance hierarchy for Hmong derivation

by which a lexical morpheme comes to serve as a grammatical marker (the marker of a construction) is called GRAMMATICALIZATION. It is pervasive throughout language.

In a related fashion, CxM provides a useful bridge between morphology and multiword expressions (MWEs). MWEs are listemes—they must be listed in the lexicon—but they have internal syntactic structure. In an integrated theory of construction grammar, they do not differ in their status from morphological words, which are also constructions. Just as a clear pathway is provided from head-of-compound to affix, a clear pathway is given from multiword expression to compound word. *Blackboard* began its life as *black board*—an adjective modifying a noun and an instance of a syntactic construction. It acquired idiosyncratic semantics and had to be listed in the lexicon as a MWE[8]. It was then reinterpreted as a single word (with a single primary stress) through the process called MORPHOLOGIZATION. According to CxM, *blackboard* has been a construction throughout its lifetime, with only incremental changes in its representation.

[8] Note that, in CxG, the syntactic construction was *also* listed in the lexicon, or constructicon.

4.5 Morphological typology

Traditionally, languages have been divided into four types based on (1) the formal properties of the morphologies, (2) their patterns of exponence, and the freedom with which they can express whole propositions in a single word. Languages like Chinese, Vietnamese, and, to a lesser extent, English, are termed ISOLATING or ANALYTIC. These languages have relatively little morphology and very little inflectional morphology. With the exception of compounding, most words consist of a single morpheme. Languages that are like isolating languages in that they display mostly simple exponence but unlike isolating languages in that they have many prefixes and suffixes are termed AGGLUTINATIVE or AGGLUTINATING. Examples of this type include Finnish, Turkish, Malayalam, and Swahili. Languages in which cumulative and extended exponence are pervasive are termed FUSIONAL or FLEXIONAL. These include languages like German, Greek, and Russian. There is a special subclass of fusional languages labeled TEMPLATIC. In these languages, including Hebrew and Arabic, words are combined of a consonantal root, a consonant-vowel template, and a vowel melody. The template and the melody may display cumulative or extended exponence. Finally, there is a traditional class called POLYSYNTHETIC which includes languages in which whole propositions can be freely formed from a single verb. This is made possible largely through a mechanism called NOUN INCORPORATION by which nouns (typically objects) are "incorporated" into verbs through a kind of compounding.

These terms are widely used, including in literature on natural language processing, but they have fallen out of favor by LINGUISTIC TYPOLOGISTS since they actually encode an unprincipled group of parameters without filling the whole space of possibilities and without rigorous definitions.

4.6 Finite State Morphology

While there are areas of computational morphology that remain subjects of research, rule based analysis and generation for agglutinate and most fusional languages is largely a solved problem. For decades, computational linguists and NLP engineers have employed finite state methodologies to parse morphologically complex languages and to generate inflected forms based on a lemma and a set of morphological properties.

4.6.1 Concatenative morphologies as regular languages

Finite state technologies have succeeded for natural language morphologies because, as formal languages, morphologies are regular—they do not have "bracketing," embedding, or recursive structure[9] This differentiates morphology from natural language syntax, which—quite uncontroversially—is context free or "mildly" context sensitive.

[9] There are a couple of possible exceptions to the statement regarding recursion. One involves compounding. Another involves affixation in certain highly agglutinative and polysynthetic languages. Even these, though, do not stop morphologies from being regular languages.

In fact, an agglutinate language is quite easy to model with finite state techniques, and it is easy to see why this is the case. Suppose the order of inflectional affixes is fixed. This is usually the case. Constructing a regular expression that will match only the licit words in such a language is straightforward. Suppose that we also want to model derivation and that the order of the derivational affixes is not fixed but what affix can occur at a given point is constrained by what morpheme (if any) has come immediately before. It is easy to see how a finite state machine can be constructed to model such a language.

For languages with reduplication, infixation, stem change, and templatic morphology, it is more difficult to apply finite state techniques, but finite state solutions to all of these phenomena exist.

4.6.2 The Xerox approach to morphological analysis and generation

LEXICAL FORM	dog+N+Pl	dog+V+3+Sg+NPast	fish+N+Pl	fish+V+3+Sg+NPast
	↕	↕	↕	↕
MORPHEMIC FORM	#dog^s#	#dog^s#	#fish^s#	#fish^s#
	↕	↕	↕	↕
SURFACE FORM	dogs	dogs	fishes	fishes

Table 4.12: Levels of representation in Xerox-style morphological analysis

4.6.3 Implementing morphotactics with LEXC

LEXC provides a useful tool for writing FSTs that define morphotactic relations and map between the lexical form and the morphemic form. Take the following data from Inuit:

INUIT	GLOSS
aglukata:quq	'she begins to work'
agluka:quq	'she works with an intermittent stoppage'
aglunani:ʁaquq	'she stops working'
aglufaqara:quq	'she rarely works'
agluviʁa:quq	'she works with difficulty'
aqujgaquq	'she wanders about'
aqujviluxtaquq	'she walks back and forth'
igliχtipixtaquq	'she walks a lot'
igliχtikʃa:ga:quq	'she walks very slowly'
igliχtikjo:ʁaquq	'she scarcely drags herself along'
qiɬpiχta:quq	'she makes holes in something'
qiɬpiχquʁa:quq	'she makes holes in various places'
qavaχɬiqja:quq	'she sleeps fitfully'

Inuit	Gloss
qavaruga:quq	'she sleeps soundly'
qavamse:quq	'she dozes'
ku:jma:quq	'she is swimming habitually towards …'
ku:jmaʁo:ʁaquq	'she swims habitually'
ɬiŋaχtaquq	'she rings'
ɬiŋaχtaga:taquq	'she rings intermittently'

If we segment this data into morphemes, we get the following result;

Inuit	Gloss
aglu-kata:-quq	'she begins to work'
aglu-ka:-quq	'she works with an intermittent stoppage'
aglu-nani:ʁa-quq	'she stops working'
aglu-faqara:-quq	'she rarely works'
aglu-viʁa:-quq	'she works with difficulty'
aquj-ga-quq	'she wanders about'
aquj-viluxta-quq	'she walks back and forth'
igliχti-pixta-quq	'she walks a lot'
igliχti-kʃa:ga:-quq	'she walks very slowly'
igliχti-kjo:ʁa-quq	'she scarcely drags herself along'
qiɬpiχ-ta:-quq	'she makes holes in something'
qiɬpiχ-quʁa:-quq	'she makes holes in various places'
qava-χɬiqja:-quq	'she sleeps fitfully'
qava-ruga:-quq	'she sleeps soundly'
qava-mse:-quq	'she dozes'
ku:jm-a:-quq	'she is swimming habitually towards …'
ku:jm-aʁo:ʁa-quq	'she swims habitually'
ɬiŋaχta-quq	'she rings'
ɬiŋaχta-ga:ta-quq	'she rings intermittently'

A LEXC grammar for such a dataset would begin with a declaration of a series of multi-character symbols that will serve as the labels for the different MORPHOLOGICAL PROPERTIES in the language. These correspond roughly to the "content" aspect of a morpheme. Some of the names are conventional, but we will have to be creative in this case since some of

the morphemes used here do not have conventional glosses. By convention, the multi-character symbols begin with a plus sign (+) if the are for external consumption or a carat (^) if they are purely for internal use:

```
Multichar_Symbols
+Inceptive +Intermittent1 +Cessive +Rarely +WithDifficulty1
+Aimless +Oscillating
+Frequentive +Slowly +WithDifficulty2
+InSomething +VeriousLoc
+Intermittent2 +Soundly +Episodic
+Directive +Habitual
+Intermittent3
+3Subj +SgSubj
```

After the declaration of multi-character symbols, which is not terminated by a semicolon, all LEXC file must consist of a series of `LEXICON` sections, the first of which must be called `Root`. The first line of each section, which declares its name, is not terminated by a semicolon but all subsequent lines are. `Root` may actually declare the roots of the language, but only if they are the leftmost morphemes in the words of the language. We might define the following root class for Inuit:

```
LEXICON Root
aglu Derivation ;
aquj Derivation ;
iglɨχtɨ Derivation ;
qɨɬpɨχ Derivation ;
qava Derivation ;
ku:jm Derivation ;
ɬiŋaχta Derivation ;
```

The first field on each line of the lexicon section can be thought of as the label on a transition of an FST (though they can consist of strings and FSTs are often constrained so that labels consist of single symbols). If a label is `aglu`, then `aglu` is consumed on the input tape and written to the output tape. As we will see, this field can either consist of strings or pairs delimited by a colon where the string on the left is the upper side and the string on the right is the lower side. The second field is the continuation class. In this case, we are constructing a very permissive transducer and are allowing derivational suffix to occur after any root. Therefore, the continuation class for all of the lexicon entries is `Derivation`. We then must define the `Derivation` lexicon:

```
LEXICON Derivation
+Inceptive:kata: Inflection ;
+Intermittent1:ka: Inflection ;
+Cessive:nani:ʁa Inflection ;
+Rarely:faqara: Inflection ;
```

```
+WithDifficulty1:vɨʁa: Inflection ;
...
+Intermittent3:ga:ta Inflection ;
```

The labels here show that the upper side, with the morphological properties, corresponds to the lower side, with the forms of the morphemes. These correspond to the lexical form and the morphemic form, respectively. The continuation class for each of these entries is `Inflection`. This lexicon will only have one entry (though, in reality, there are many other Inuit inflection entries that could be added to this list):

```
LEXICON Inflection
+3Subj+SgSubj:quq # ;
```

The continuation class for this entry is **#**, which is the LEXC end of word symbol.

The LEXC file cannot be run directly. Instead it must be imported into Foma via the interactive interface or and XFST script. If we saved the above lexicon file as `inuit.lexc`, we could import it with:

```
read lexc < inuit.lexc
```

It would then be added to the stack. If we want to define a constant `Verbs` corresponding to it (the top FST in the stack) we can use:

```
define Verbs;
```

We can then compose the FST with another FST implementing the mapping between the morphemic morpheme and the surface form in the usual way.

```
read regex Verbs .o. SpellingRules;
```

4.7 Unsupervised morphology induction

4.7.1 Algorithms

4.7.2 Implementations

4.7.3 Limitations

4.8 Practice Exercise: Swahili

4.8.1 Part I: Swahili nouns

Swahili, like other Bantu languages, has many NOUN CLASSES or GENDERS. These are reflected both in agreement with verbs and in the morphology on nouns. In the following data set, words from one of these classes are listed (in broad phonetic transcription—you would see most of the same patterns in the orthography).

SG	GLOSS	PL	GLOSS
ubao	plank	mbao	planks
	wing	mbawa	wings
udevu	hair	ndevu	hairs
ugwe	string	ŋgwe	strings
uwati	hut pole	mbati	hut poles
uwanda	open place	mbanda	open places
uwiŋgu	heaven		heavens
ulimi	tongue	ndimi	tongues
upaŋga	sword	pʰaŋga	swords
upindi	bow	pʰindi	bows
utambi	lamp wick	tʰambi	lamp wicks
utepe	stripe		stripes
ukuta	wall	kʰuta	walls
ukuni	stick		sticks
ukut͡ʃa	finger nail	kʰut͡ʃa	finger nails
ufuŋguo	key	fuŋguo	keys
ufagio	broom	fagio	brooms
ufizi	gum	fizi	gums
uvumbi	bit of dust	vumbi	dust
usiku	night	siku	nights
uʃaŋga	bead	ʃaŋga	beads
wakati	season	ɲakati	seasons
	net	ɲavu	nets
wajo	footprint	ɲajo	footprints
wembe	razor	ɲembe	razors
wimbo	song	ɲimbo	songs

You should build an FST using Foma (the LEXC and XFST formalisms) that can parse any of these nouns as well as generate any of theme from a lexical form (e.g. "+N+Class1+Pl+bao"). The morphophological rules will be implemented in your XFST script and your morphotactics will be implemented in your LEXC file. Your initial LEXC file should look like this:

```
Multichar_Symbols
+N
+Class1
+Sg +Pl

LEXICON Root
+N:0 Noun ;
```

Use your FST to fill in the gaps in the table above. Your FST will be evaluated on its ability to perform this task

4.8.2 Part II: Swahili verbs

The second Swahili data set consists only of inflected verbs:

SWAHILI	GLOSS
atanipenda	he will like me
atakupenda	he will like you
atampenda	he will like him
atatupenda	he will like us
atawapenda	he will like them
atatupenda	he will like us
nitakupenda	I will like you
nitawapenda	I will like them
utanipenda	you will like me
utampenda	you will like him
tʰutampenda	we will like him
watampenda	they will like him
atakusumbua	he will annoy you
unamsumbua	you are annoying him
atanipiga	he will beat me
atakupiga	he will beat you
atampiga	he will beat him
ananipiga	he is beating me
anakupiga	he is beating you
anampiga	he is beating him
amenipiga	he has beaten me
amekupiga	he has beaten you
amempiga	he has beaten him
alinipiga	he beat me
alimpiga	he beat him
wametulipa	they have paid us
tʰulikulipa	we paid you

Extend your FST from Part I so that it can parse and generate verbs with these morphemes. Your initial LEXC file should look like the following:

```
Multichar_Symbols
+N +V
+Class1
+Sg +Pl
+1SgS +2SgS +3SgS +1PlS +3PlS
+1SgO +2SgO +3SgO +1PlO +3PlO
+Fut +Pres +Perf +Past

LEXICON Root
```

```
+N:0 Noun ;
+V:0 Verb ;
```

Use your FST to fill in the following table (it will be evaluated in the same way):

LEXICAL FORM	SURFACE FORM	ENGLISH GLOSS
		I have beaten them
		they are beating me
		they have annoyed me
		you have beaten us
		we beat them
		I am paying him
	atanilipa	
	utawapiga	
	walikupenda	
	nimemsumbua	

What to hand it (in a ZIP file containing a single directory with the name of your Andrew ID):

1. A README explaining any facts specific to your implementation

2. Part I

   (a) An XFST script called swahili1.xfst

   (b) A LEXC script called swahili1.lexc

   (c) A completed table of Swahili nouns.

3. Part II

   (a) An XFST script called swahili2.xfst

   (b) A LEXC script called swahili2.lexc

   (c) A completed table of lexical forms, surface forms, and English glosses, as given above.

4.9 Exercise

Karbardian is a language of the Russian Caucasus. Here are some noun paradigms from this language:

‘glass’	‘rooster’	‘death’	
aːbgʲa	—	aːʒaːla	N; NDEF; SG
aːbgʲaxa	aːdaːqaxa	—	N; NDEF; PL
aːbgʲar	aːdaːqar	aːʒaːlar	N; NOM; SG
—	—	—	N; NOM; PL
—	aːdaːqam	aːʒaːlam	N; ERG; DEF; SG
aːbgʲaxama	aːdaːqaxama	—	N; ERG; DEF; PL
aːbgʲakʲ’a	—	—	N; INS; NDEF; SG
aːbgʲaxakʲ’a	aːdaːqaxakʲ’a	—	N; INS; NDEF; PL
aːbgʲamkʲ’a	aːdaːqamkʲ’a	—	N; INS; DEF; SG
—	—	—	N; INS; DEF; PL

The inflectional features are defined below:

- POS
  - **N** Noun.
- Definiteness
  - **NDEF** Indefinite—appears in noun phrases that are indefinite (like those with *a*/*an* in English
  - **DEF** Definite—appears in noun phrases that are definite (like those with *the* in English
- Case
  - **NOM** Nominative—appears in noun phrases that have nominative case (appearing as the subject of a transitive or intransitive verb) like *the sleepy cat* in *The sleep cat ran.*
  - **ERG** Ergative—appears in noun phrases that have ergative case (appearing as the subject of a transitive verb but not an intransitive verb) like *the sleepy cat* in *The sleepy cat caught the mouse.*
  - **INS** Instrumental—appears in noun phrases than have instrumental case (usually used for noun phrases that express the means by which something is accomplished) like *the sleepy cat* in *I caught the mouse with the sleepy cat.*
- Number
  - **SG** Singular—appears in noun phrases that are singular.
  - **PL** Plural—appears in noun phrases that are plural.

Submit the following materials:

- Part 1
  - Construct an FST with Foma **not using flag diacritics** that generates all of the missing forms in the table above. Do not treat strings of affixes as a unit unless there is no other option.
  - Apply your FST to *aːbgʲaxamkʲʔa* and report the result.
  - What limitations do FSTs impose in how you model this data? What redundancies are your forced to introduce?
- Part 2
  - Construct a transducer that models these data using Foma flag diacritics.
  - How does this analysis improve on your last analysis, which used pure FSTs?

5 Encoding Meaning with Morphosyntax

5.1 Introduction

In this part of the course we will examine basic types of meaning and how they are expressed using morphology and the order of words. The meanings are so basic that you probably take them for granted. We will observe that different languages use morphology and word order in different ways to express these meanings. Because word order is one aspect of syntax, this area of linguistics is called morphosyntax, the expression of meaning using morphology and syntax.

In the first section of this chapter, we will introduce some basic meanings. In section 2, we will introduce *interlinear gloss*, a notation that linguists use to explain the morphosyntax of a language to other linguists who don't know the language. Then we will return to the basic meanings and, using interlinear gloss notation, introduce the concepts of *morphosyntactic strategies* and *morphosyntactic typology*, how languages use morphology and syntax in different ways to express these meanings.

TODO: we need to make separate chapters on parts of speech and subj-obj vs agent-patient

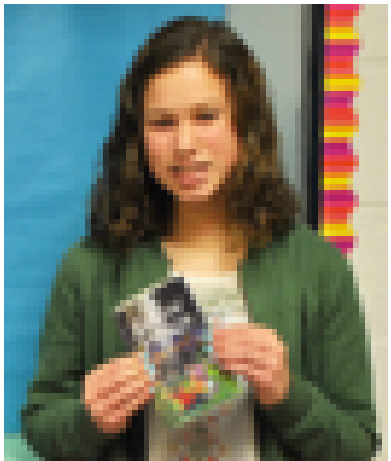
5.2 Basic meanings

Agent acts on patient: Figure 5.1 shows two examples of an *agent* acting on a *patient*. In the top picture, a fencer (agent) is stabbing a monster (patient). In the second picture, a mother (agent) is kissing a baby (patient).

Figure 5.1: Agent acts on patient

Possession: someone has something: In Figure 5.2, we see a girl who has a book and a baby who has a toy.

Location: A figure is located on a ground: Some sentences are not about doing or having, but about being located or moving. The pictures in Figure 5.3 are about a FIGURE (book or picture) being located against a GROUND (table or wall). Note that we conceptualize the situation as a book being on a table and not as a table being under a book. So before we even express the sentence using morphosyntax, we have decided which entity is the figure and which is the ground.



Movement: Figure 5.4 shows figures moving with respect to grounds. A car is moving through a tree and people are moving with respect to a bridge. Notice that we don't conceptualize the situation as the tree moving over the car or the bridge moving under the people.

Something causes an emotion: In Figure 5.5 we see a popsicle causing happiness in a child and a cucumber causing fear in a cat. Alternatively, these pictures can be interpreted as an animate experiencer having a reaction (happiness or fear), possibly in response to something (popsicle, cucumber).

TODO: fill in copula (identity, definition, property); existence; comparison; comitative; instrumental (maybe move the kid stabbing the monster with an epee).

TODO: mention other types of meaning such as cognition, perception, changing state, appearing and disappearing, doing an activity like working or singing, body functions like breathing or bleeding, etc.

Figure 5.5: Stimulus causes emotion for experiencer

5.3 *The Leipzig Glossing Rules*

Please read the document at this URL: `https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf`. The Leipzig Glossing Rules are very important because this is the format that linguists use for communicating with each other about languages that they might not speak. When you read a linguistics article or textbook, you read the example sentences and glean facts from them. Skipping the examples in a linguistics text is like skipping the formulas in a math or CS text! And as with mathematical formulas, you have to learn a bit of notation to understand them. Here is the first example from the Leipzig glossing rules document:

(7) Mereka di Jakarta sekarang.
they in Jakarta now
'They are in Jakarta now.'

The first line, *Mereka di Jakarta sekarang*, is in the OBJECT LANGUAGE, Indonesian. The third line, *They are in Jakarta now*, is in the METALANGUAGE, English. The second line is the important one. It is called the GLOSS or INTERLINEAR GLOSS (because it is between the first and third lines). The second line uses English words corresponding to Indonesian words, keeping the order of the Indonesian words. It is very important that the words in the second line are tabbed to line up with the words in the first line.

What makes the second line the most important? It tells you about the grammar of Indonesian, in this case, the order of the words and the fact that there is not a COPULA (*be*-verb) in Indonesian.

A common error for beginning linguistics students is to use line three to draw conclusions about morphosyntax of the object language. In this example, it would be a mistake to assume that Indonesian has a copula based

on the occurrence of the word *are* in English in line three. However, it is not a mistake to compare lines two and three in order to conclude that English and Indonesian have different MORPHOSYNTACTIC STRATEGIES for expressing the location of something, for example, that they are located in Jakarta. English uses a *copula strategy* and Indonesian uses a *zero-copula strategy*.

The previous example could also have been glossed differently:

(8) Mereka di Jakarta sekarang.
3.PL LOC Jakarta now
'They are in Jakarta now.'

Here, we have used morpheme labels (or tags) in place of some of the English words. The morpheme labels are formatted in small caps to distinguish them from metalanguage (English) words. Every linguistics book or article you read will have a list of abbreviations so that you can interpret the labels. There is a list of labels at the end of the Leipzig document. Many publications require the Leipzig labels, but not all do. Also, there are many meanings that are not covered by the Leipzig labels.

In the example above, 3.PL means *third person plural*. The dot between 3 and PL indicates that there are two meanings, 3 and PL, expressed by one morpheme. LOC means "locative", indicating that the Indonesian word *di* is a locative marker, a word that indicates location. Glossing *di* as *in* is simpler, but a bit misleading. If you speak more than one language, you know that there is almost never a perfect correspondence between words. It will not be the case that *di* is always used where English *in* is used. The label LOC is, therefore, a little less misleading.

Here is one more example from the Leipzig Rules document. The object language is Russian, which is written in IPA or some roughly Romanized form rather than Cyrilic.

(9) a. My s Marko poexa-l-i avtobus-om v Peredelkino.
1.PL COM Marko go-PST-PL bus-INS ALL Peredelkino
'Marko and I went to Peredelkino by bus.'

b. My s Marko poexa -l -i avtobus -om v Peredelkino.
1.PL COM Marko go -PST -PL bus -INS ALL Peredelkino
'Marko and I went to Peredelkino by bus.'

Here is a list of abbreviations used in the Russian example:

1	first person
ALL	allative (going toward)
COM	comitative (doing something together)
INS	instrumental (using an instrument)
PL	plural
PST	past tense

The (a) example above is copied from the Leipzig document. For the (b) example, we have changed the Leipzig rules in one way: we have used a space before each morpheme so that each morpheme label in line two is tabbed to line up with a morpheme in line one. In your homework, we want you to use spaces between morphemes because it is easier to grade. A dash indicates that something is an affix rather than a free-standing word. In this example *-l* and *-i* are affixes of *poexa*. In other words, the third word of the Russian sentence is *poexali*. *-om* is an affix of *avtobus*, forming the word *avtobusom*.

Looking at the Russian example, we can identify a few morphosyntactic differences between Russian and English. First, there is a difference in expressing COORDINATION. The English phrase *Marko and I* can be paraphrased from Russian as *we with Marko*. The Russian word *s* is glossed as COM for COMITATIVE. The term *comitative* refers to situations where people do things together.

Next, lets look at the the Russian word *avtobusom*, where the morpheme *-om* is glossed as INS for INSTRUMENTAL. The term *instrumental* prototypically refers to situations where a human manipulates an instrument such as a spoon or a knife. So, *avtobusom* would translate most literally into English as *with bus*, but in the third line of the example, we say *by bus* because that is the natural way to say it in the metalanguage, English.

Here is a summary of two meanings and the morphosyntactic strategies used to express them in English and Russian:

Meaning	**English Strategy**	**Russian Strategy**
People do something together	*and*, *with*	COM (*s*)
Use a form of transportation	*by*	INS (*-om*)
Person manipulates tool	*with*	INS (*-om*)

Here are English examples illustrating the three meanings:

Meaning	**English Strategy**
People do something together	I went to Pittsburgh *with Marko*.
People do something together	*Marko and I* went to Pittsburgh
Use a form of transportation	I went to Pittsburgh *by bus*.
Person manipulates tool	I ate *with a spoon*.

A final morphosyntactic difference we can observe between English and Russian is that the Russian verb agrees in number with the subject using the morpheme *-i*, which is glossed as -PL.

5.4 An example of glossing in English

(10) a. The North Wind and the Sun were argu -ing
DEF North Wind and DEF Sun COP.PST.3.PL argue -PROG
'The North Wind and the Sun were arguing'

b. about which one of them was strong -er,
about which one of 3.PL COP.3.SG strong -COMPAR
'about which one of them was stronger,'

c. when a traveler came by
when INDF traveler come.PST by
'when a traveler came by'

d. wear -ing a heavy coat.
wear -PROG INDF heavy coat
'wearing a heavy coat.'

e. They agree -d
3.PL agree -PST
'They agreed'

f. that whoever got the traveler to take off
COMP REL.INDF get.PST DEF traveler COMP take off
his coat first
3.SG.GEN coat first
'that whoever got the traveler to take off his coat first'

g. would be consider -ed strong -er
FUT.HYPOTH COP.INF consider -PASS strong -COMPAR
'would be considered stronger'

We will look how different languages use different STRATEGIES to express the same meaning. For example, to express the meaning of possession, languages can use a "have" strategy (*I have a book*) or a "be" strategy that can be paraphrased as *A book is at me* or *A book is to me*. You will learn to identify the strategies in languages that you do not speak by reading *reference grammars*, books written by linguists for linguists. In other words, you will learn the terminology and notation that linguists use to talk to communicate with each other.

5.5 An example

Compare morphosyntactic strategies in English and Hebrew for clausal possession, including tense, case, and negation. How similar is your language's strategy to English or Hebrew?

First, let's look at three basic sentences in Hebrew and English. In English, we can see three ways of referring to the speaker of the sentence, the first person singular pronoun, *I*, *me*, and *to me*. In Hebrew, there are also three forms of the first person pronoun, *ani*, *oti*, and *li*.

(11) a. I saw her.

b. I saw a book.

c. She saw me.

d. She gave a book to me.

(12) a. (Ani) ra'iti otah
1.SG see.PAST.1.SG 3.SG.FEM.ACC
'I saw her.'

b. (Ani) ra'iti sefer
1.SG see.PAST.1.SG book
'I saw a book.'

c. (Hi) ra'ata oti
3.SG.FEM see.PAST.3.SG.FEM 1.SG.ACC
'She saw me'

d. (Ani) natati lah sefer
1.SG give.PAST.1.SG 3.SG.FEM.DAT book
'I gave a book to her'

e. (Hi) natna li sefer
3.SG.FEM give.PAST.3.SG.FEM 1.SG.DAT book
'She gave a book to me.'

Now let's look at sentences expressing possession. We will distinguish two roles, the *the possessor* and *the possessed*. In the English sentence, *I have a book*, the possessor is *I*. The sentence has the same syntax as *I saw a book*, a subject, a verb, and an object. In Hebrew, however, the first person pronoun used for the possessor is *li*, in contrast to the subject of a sentence, for which we would use *ani* as the first person pronoun. In fact, the syntax of Hebrew possessive sentences is similar to Hebrew existential sentences. The English existential sentence *There is a book* is not similar to the English clausal possessive sentence *I have a book*.

```
Yesh    li        sefer
EXIST   DAT.1.sg  book
'I have a book'

Yesh   lah          sefer
EXIST  DAT.3sg.FEM  book
'She has a book'

Yesh  sefer
EXIST book
'There is a book.'
```

5.6 Conceptual Foundations

5.6.1 Grammaticalization and conventionality

The concept of *grammaticalization* was introduced in the last chapter. Grammaticalization is a historical process by which *open-class* words can become grammatical morphemes. Here are some interesting examples from *The World Lexicon of Grammaticalization*. Words for small things like steps and drops become negative morphemes, as in French *Je ne sais pas (step)*. You can imagine saying something like *I didn't go, not even a step*. This becomes so conventional that after a long period of time, the original meaning is lost. An example that we can see in English is "go" becoming a marker of future time as in *I am going to do it*. When things are fully grammaticalized, they start to be phonologically reduced as in *I'm gonna do it* or *Ama do it*.

In this chapter, we are talking about grammaticalized morphosyntactic constructions. There may be less conventional ways to say the same thing, but we are not concerned with those in this chapter. However, the boundary between conventional, grammaticalized constructions and non-conventional ways of saying things is not always clear:

Grammaticalized: Pat is taller than Chris.
Normal: Pat is tall compared to Chris.
Almost normal: Pat's height is more than Chris's.
Really indirect: As for height, Pat excels, but Chris is lacking.

Grammaticalized: I have a book.
Indirect: A book is in my proximity.

In this chapter as we look at morphosyntactic constructions in different languages, we will restrict ourselves to the most grammaticalized ones. We will see that grammaticalization can be very different in different languages. For example, in many languages, the main grammaticalized strategy for expressing comparison can be paraphrased as *Pat is tall, exceeds/surpasses Chris*. Let's call this the *verb strategy*. (Although if it is very grammaticalized, the speakers of the language may not think of *exceed*, or *surpass* as a verb.) Let's call the English strategy, *Pat is taller than Chris*, the comparative adjective strategy. Languages that use the verb strategy may not have a comparative morpheme that attaches to adjectives. Languages that use the comparative adjective strategy may not have *serial verbs*, strings of predicates like *take knife, cut bread* or *is tall, surpasses Chris*.

5.6.2 Constructions

In the previous chapter, we introduced *constructions*, pairs of form and meaning. Constructions can be a result of grammaticalization. Constructions can be arbitrary constellations of lexemes, morphemes, and syntactic structures.

Here are some well-known examples from Construction Grammar:

What is he doing? (literal or incongruity)
What is this fly doing in my soup? (incongruity)
What is she doing going to the movies? (incongruity and non-core syntax)
Why not go? (suggestion/advice and non-core syntax)
What a nice/horrible dress! (predication and non-core syntax)

Constructions follow a continuum from lexicon to grammar and from compositional (meaning is predictable from the meaning of the parts) to idiomatic or non-compositional. In other words, it's normal not to be normal.

A true construction grammar is *constructions all the way down*[1], meaning that even individual words are constructions and that every construction is built from smaller constructions.

[1] The world rests on the back of a giant turtle. *What is holding up that turtle?* Another turtle. *And what is holding up that turtle?* You can't fool me. It's turtles all the way down.

5.6.3 Radial categories

The meanings of morphosyntactic constructions are usually not simple. You can see this by looking at any ESL (English as a Second Language) web page. If you look up the usage of *will* you will find the following:

I will go tomorrow. (future or promise)
She won't wash the dishes. (future or refusal)
Phone rings. That will be my son.
This bottle will hold three liters. (capacity)
Server: What will you have? *Customer:* I will have coffee.

The theory of *Cognitive Grammar* shows that languages are organized around prototypes and *radial categories*. The first and second examples above are probably the most prototypical uses of "will". The others are clearly related meanings, but they are less predictable, and you would not take it for granted that the morpheme that means "future" in another language would carry all of the same meanings. In this chapter we will see prototypes and radial categories in many grammaticalized meanings such as possession and definiteness. We will also see prototpyes and radial categories in the linguistic categories like *noun* and *subject*.

5.6.4 Grammaticality judgments

Imagine for a moment that human languages are like formal languages, possibly infinite sets of sentences that can be characterized by a finite set of rules.[2] Also imagine that, inside the brain of each native speaker, there is an automaton of some sort that accepts all of the members of the set and rejects any

[2] In the previous chapter, you modelled morphology using finite state automata. In the next chapter we will model human languages with context free grammars and other formalisms that are more powerful than context free grammars.

sentence that is not in the set. And imagine that you access this automaton by saying, “Is this grammatical?” As you observe the behavior of this automaton, you keep track of the results by putting an asterisk on sentences where the native speaker says, “no”.

```
She drank some coffee.
She drinks coffee.
*She some coffee drank.
*Coffee some drank she.
*Drank some she coffee.
*She drink coffee.
*She drinking coffee.
*This books is interesting.
*Me drinks coffee.
```

Sometimes the native speaker has trouble deciding, and you add a question mark instead of an asterisk.[3]

[3] Manning (2003) showed that the question mark sentences here occur in corpora but with much lower frequency.

```
I consider her to be a friend.
I regard her as a friend.
?I consider her as a friend.
?I regard her to be a friend.
```

Sometimes people from different places or age groups give different answers and sometimes you get the impression that the answers were different a few hundred years ago.

```
The car needs washed.
I ain’t seen him.
We go to the movies a lot anymore.
I know not what others prefer.
```

Sometimes the native speakers don’t understand what kind of answer you want. For example, the language includes sentences that are false and sentences that could only be true in a cartoon world, but the native speakers want to reject them. They also might want to reject things that they say all the time because an English teacher told them they were wrong.

```
Colorless green ideas sleep furiously.
Me and her went to the movie.
Cancer causes smoking.
```

So, you have to work around all of these problems as you try to characterize which sentences are members of the language and which are not.

5.7 Building blocks

5.7.1 Categories of words (Parts of Speech)

As we try to model a human language as a set of rules or as a set of constructions, we sometimes have to refer to specific words. For example, to describe the English *correlative* construction we need to state that each half of the construction starts with the word *the*:

```
The more I read the smarter I get.
The smarter I get the more people respect me.
*A/this more I read a/this smarter I get.
```

More commonly, rules and constructions are described in terms of categories of words such as *noun*, *verb*, and *adjective* and categories of phrases such as *noun phrases* and *verb phrases*.[4]. For example, a prototypical English sentence consists of a noun phrase and a verb phrase. A noun phrase can contain a determiner, an adjective, and a noun. It would be silly to state a specific rule for each noun and verb. We keep the rules clean and concise by stating them in terms of categories. But we do need to know which words are in each category, so we make a list of words and their categories and call it a *lexicon*.

[4] Koeneman and Zeijlstra

Now lets look more closely at the definitions of categories such as *noun* and *verb*. We will call categories of words *parts of speech*. For some tasks in natural language processing (NLP), it is important to have operational definitions of categories. A definition is operational if two people can ask themselves, "Is this word in category X?" and have a good chance of coming up with the same answer.

In linguistics, definitions of parts of speech are operationalized by describing the morphosyntactic constructions they can and can't occur in. This type of definition is called a *distributional test*.

Tests for determiner:	
Makes a singular count noun good	*Book is good. The/this/each/a book is good.
Only one determiner per noun phrase	*The this book *Each a book
Quantity determiners in partitive construction	All/each of the books
Determiners precede all adjectives	The blue book *blue the book

We use tests in the following way. Suppose we want to know which of these words are determiners: *which, both, every, each, all, enough, much, many, my*. Let's start with *which*. The first test is about *count nouns*. These are nouns that can be counted without a classifier word in English. *Book* is a count noun because it can be counted: *one book, two books. Information*, on the other hand, is a mass noun: **one information, *two informations, one piece of information*. Singular count nouns in English do not occur without determiners: **book is good, the book is good*. Here is how we use the test:

Step 1: Be clear about what the test predicts. The "makes a count noun good" test states that if *which* is a determiner, you can put it in front of a singular count noun and the result will be grammatical. But the "only one determiner" test states that if "which" is a determiner and you put it with another determiner the result will be ungrammatical.

Step 2: Construct an example that you will present to a native speaker (possibly yourself). We can construct this sentence: *Which book is good*. Constructing a sentence can sometimes be tricky because you can accidentally change something important and come up with a sentence that does not actually test what it is supposed to test. Also, some words are invalid in some tests. For example, the first test only works with words that go with singular count nouns. The test is therefore invalid for *these, those,* and *all*. The last test is only valid for determiners that are about quantities, and therefore does not apply to *my*.

Step 3: Get a grammaticality judgment. In this case, *Which book is good* is grammatical.

Step 4: Check with Step 1. Did the grammaticality judgement come out as predicted by the test? In this case it did and we say that *which* passed the *Makes a singular count noun good* test for being a determiner.

Distributional tests are used in *annotation manuals*. The Penn Treebank

part of speech manual for English is about thirty pages long and contains many tests for each part of speech. (Human part-of-speech annotators keep hundreds of tests in their heads!) If the annotation manual is well-written, human annotators will have high *intercoder* or *interannotator* agreement.

However, distributional tests for parts of speech are not perfect.

```
Many of the books  (Passes the partitive construction test.)
Many a book (Fails the 'one determiner' test.)
The many books written by Jane Austen (Fails 'one determiner')

*Every the book (Passes the 'one determiner' test.)
*Every of the books (Fails the partitive construction test.)
My every wish came true. (Fails 'one determiner')

All/both of the books (Passes the partitive construction test.)
All/both the books (Fails the 'one determiner' test.)
```

Some of the problems can be reconciled by positing new parts of speech. *All* and *both* are determiners, but they also belong to another part of speech called *predeterminer*. Other problems can be resolved by dividing a class into sub-classes such as transitive verb and intransitive verb or count noun and mass noun. Still others can be attributed to frozen constructions such as *My every wish*.

In line with a constructional approach to language, we will also simply allow for exceptions: individual words fail to occur in a construction they should occur in or occur in a construction they should not occur in.

There are three looming questions remaining: how many parts of speech are there? Are parts of speech universal? And what about the definitions you learned in school? Let's start with the last question.

What about the definitions you learned in school? The definitions you learned in school are not operationalized. Suppose you tell two people that adjectives denote properties and prepositions denote times and locations. Will they make the same decision about whether *like, worth*, and *near* are adjectives or prepositions. (In fact, linguists like to fight about this.)

```
She is like her mother.
She is near her mother.
She is worth her weight in gold.
```

The definitions of noun and verb as things and events also crumbles under

close inspection (Koeneman and Zeijlstra; Pinker).

```
He is dancing gracefully.
His graceful dancing drew applause.
```

```
Fungi interest many people.
Their interest in fungi can become an obsession.
Interesting fungi can be found everywhere.
```

So semantic or metaphysical definitions of parts of speech are not precise enough for the fields of linguistics and NLP. However, a similar type of definition might be workable (Bender, Croft): when you use a word to refer, it is a noun; when you use a word to predicate, it is a verb. This is a functional definition of parts of speech. For this class, we will stick with distributional definitions of parts of speech, with the understanding that some words may go rogue in some constructions.

How many parts of speech are there? There is no answer to this question. As linguists investigate the constructions of each language, they may identify many subcategories: transitive and intransitive verbs, count and mass nouns, gradable and non-gradable adjectives, etc. No linguist can say exactly how many categories are necessary to describe a language.

In NLP we find a contrast between around forty parts of speech for English in Penn Treebank and twelve in the Google Universal Dependency approach. The Google Twelve were selected to optimize cross-lingual training of part of speech taggers. They are not "correct" in any way. In fact, we are not sure whether it has really been proven that these twelve are optimal for all tasks.

```
THE GOOGLE TWELVE

Noun
Verb
Adjective
Adverb
Pronoun
Determiner
Adposition
Conjunction
```

Number
Particle
Punctuation
Other

PENN TREEBANK ENGLISH POS TAGS

Coordinating conjunction
Cardinal number
Determiner
Existential-there
Foreign word
Preposition/subordinating conjunction
Adjective
Comparative adjective
Superlative adjective
List item marker
Modal
Singular noun or mass noun
Plural noun
Singular proper noun
Plural proper noun
Predeterminer
Possessive ending
Personal pronoun
Possessive pronoun
Adverb
Comparative adverb
Superlative adverb
Particle
Symbol
To
Interjection
Base form verb
Past tense verb
Gerund or present participle verb
Past participle verb
Verb not 3rd person singular present
Verb 3rd singular present
Wh-determiner
Wh-pronoun
Possessive wh-pronoun
Wh-adverb

Are parts of speech universal? Yes and No. Linguistics as a field has several goals: to describe individual languages as accurately as possible, to discover the universal nature of human language, and to compare similarities and differences between languages. Describing each language accurately for posterity may require creation of idiosyncratic categories because each language is not exactly like any other. But comparing languages requires some common basis for comparison.

The operationalized tests for major categories like noun and verb may be quite different for different languages. For example, a test for nounhood in English is that nouns occur with determiners, but some languages do not have determiners. So occurring with a determiner would not be a test for nounhood in those languages.

So how can we equate the category of Noun in English with the category of Noun in another language? We look at interchangeability and prototypes. First we find clusters of words that are interchangeable in some basic constructions: The/a/this rock/stick/child fell/rolled/floated. We do not expect the clusters to be isomorphic in any two languages. Then we look at the most prototypical members of each cluster. If the most prototypical member of a cluster is a word whose primary function is to refer to a time-stable physical object, we call that cluster Noun.

Why linguists say some categories don't exist in some languages:

Hmong

Nws yog ib tug neeg phem
3sg be one clf human evil
'She is an evil person.'

Neeg phem heev.
human evil very
'People are very evil.'

Neeg phem dua cuam.
human evil surpass ape
'People are more evil than apes.'

Nws yog ib tug neeg paub
3sg be one clf human know
'She is a person who knows'

Neeg paub heev.
human know very

```
'People know a lot.'

Neeg paub dua cuam.
human know surpass ape
'People know more than apes.'
```

5.7.2 Verbs and arguments: subcategorization

Linguists use many metaphors to talk about verbs: a mini script with roles for actors, a function that takes arguments, or an atom with valency for combining with other things. Let's work with the "function" metaphor for now. If you think of verbs as functions, you can imagine them having different numbers of arguments.

```
Intransitive:
The student yawned.
Transitive:
The student bit the dog.
Ditransitive:
The student handed the teacher a book.
    The student handed a book to the teacher.
```

Using a verb with the wrong number of arguments can result in ungrammatical sentences:

```
*The student yawned the lecture.
*The student bit the dog a treat.
*The student handed the teacher.
*The student handed the book.
```

Words that have the same part of speech are supposed to have the same distribution (be able to occur in the same types of sentences and constructions). But intransitive, transitive, and ditransitive verbs may not be interchangeable as you can see by looking at the ungrammatical sentences above. Linguists therefore talk about intransitive, transitive, and ditransitive as *subcategories* of verbs. They use a lot of terminology such as *the verb's subcategorization*, meaning what subcategory the verb is in. They also say things like, "This verb subcategorizes for an object" or, "This verb is subcategorized for an object."

A clause is one verb and its arguments. Some sentences have more than one clause: *Having slept well, she was feeling refreshed; If it rains, I will not go; The*

party was cancelled because it rained; We want them to vote; I think that the students will yawn. Each verb defines its own clause.

5.7.3 Semantic Roles

Now let's switch metaphors and think of a verb as a mini script with roles for different players. In a play, there are many ways to refer to the roles. For example, there are heroes and villains. But in some plays the villain is a murderer and in others the villain is a banker. There are also many ways to refer to the semantic roles associated with verbs, and none of them are "right".

Macro Roles: One approach to naming semantic roles is to use very few role names that are very general. When a verb has two arguments such that one acts and the other is affected, the two roles can be called *agent* and *patient* or *actor* and *undergoer*. These roles are useful for verbs like the following:[5]

[5] *Theme*, in case you've heard the word, is a macro role that comes from a particular theory generally attributed to Ray Jackendoff. Most of the patients/undergoers in these examples are themes. But subjects of many intransitive verbs are also themes. Adherents of this theory call semantic roles *thematic roles*.

```
Contact:  hit, kick, smack
Creation:  make, create, sew a dress, cook, build
Attachment: attach, sew a button on a dress
Incremental change:  mow, eat, draw
Change of location:  push, move, put, take
Change of possession: give, hand, present
And many others
```

Other macro roles include source, goal, path, location, experiencer, and stimulus.

```
We ran to the house from the road along the path.
I fear snakes.
Snakes frighten me.
Put it on the table.
It sits on the table.
```

Proto Roles: Macro roles show a lot of diversity. The agent of giving is very different from the agent of building. David Dowty attempted to define proto agent and proto patient. For example, in the sentence *I built a house*, I existed before the building event, I did something volitional, and at the end of the building event, I may have been stronger or happier, but that is really not relevant to the building event. On the other hand, the house did not exist before the building event, did not do anything volitional, and changed a lot during the event in ways that are relevant – it went from not being built to being partly built to being completely built. Proto-agent properties include

animacy, volition, not being affected by the event, and existing before the event. Proto-patient properties include changing incrementally during the event (being built, painted, or mowed) or coming into existence as a result of the event. In any given sentence, the agent may not have all of the proto-agent properties and the patient may not have all of the proto-patient properties. But typically one argument has more proto-agent properties and the other has more proto-patient properties. In PropBank, the most agent-like argument is called Arg0 and more patient-like arguments are called Arg1.

Verb-specific role names: Sometimes, you just give up trying to make the macro role names work for every verb and just use verb-specific role names like *builder*. *giver*, *seer*, *seen*, and *attachee*.

Frame-based role names: The FrameNet lexicon is based on "frames" such as commercial transaction as in *I bought a book from you for ten dollars* or *You sold me a book for ten dollars*. In non-frame-based approaches, the agent is *I* (the buyer) in the first sentence and *you* (the seller). But in a frame-based approach both sentences could have the same role names.

So can you make up any role names you want and use them any way you want? No. There is no "correct" set of role names, but each set is carefully defined and operationalized. If you use the standard role names, you have to use them in a way that is consistent with how other people use them.

5.7.4 Grammatical Relations

Subject (SUBJ) and *object* (OBJ) are grammatical relations. **How a language distinguishes who bit who is sometimes called Grammatical Encoding. Grammatical Encoding is about morphosyntactic mechanisms for distinguishing SUBJ from OBJ.** Now you will have to forget what you learned in school about SUBJ and OBJ. First we will tell you what SUBJ and OBJ are not. Then we will try to define what they are.

SUBJ and OBJ are not agent and patient: The following sentences describe a situation in which a door (the patient or theme) undergoes a change of state from being closed to being open. An agent, Alex, can bring about the change of state, and the agent may manipulate an instrument (the key) in bringing about a change in state. A force (the wind) may also cause the change in state. You can see in the sentences below that each of the participants (Alex, the wind, the key, and the door) can be the subject of a sentence. In English, the subject comes before the verb, and third person subjects trigger an agreement marker *-s* on present tense verbs.

```
Alex opens the door (with a key).
The wind opens the door.
The key opens the door.
The door opens.
The door was opened (by Alex) (with a key).
```

OBJ in English immediately follows the verb, and can become the subject of the passive voice version of the verb:

```
Alex kicked the ball.
The ball was kicked (by Alex).
```

Now, let's consider ditransitive verbs, which have an agent, a patient, and a recipient. You may have learned that *to Alex* is the "indirect object" in the sentences below. But we are going to look at it from a different perspective in which we distinguish grammatical relations from semantic roles.

```
The student gave a book to Alex.
The student told a story to Alex.
```

The two sentences below have roughly the same meaning, but the words are in a different order, and a preposition, *to*, appears in one sentence, but not the other. What the two sentences have in common is the semantic roles of *a book* and *Alex*. Alex is the *recipient* and *book* is the patient. The two sentences differ in their grammatical relations. In the first sentence, *book* is the OBJ, whereas in the second sentence, *Alex* is the OBJ. The alternation shown in these sentences is called *Dative Shift* because the preposition *to* corresponds to dative case in some languages. Not all languages have Dative Shift.

```
The student gave a book to Alex.
The student gave Alex a book.
```

It may be mind bending to think of *Alex* as an OBJ. You may have always thought of it as an indirect object. However, consider the following. The way we identify OBJ in English is that it is immediately after the verb and can become the subject of a passive voice sentence. By these criteria, *Alex* is an OBJ in the second set of sentences below!

```
The student gave a book to Alex.
Alex was given a book by the student.

The student gave Alex a book.
Alex was given a book by the student.
```

What are SUBJ and OBJ? The definition of SUBJ and OBJ is the topic of thousands of books and papers. They are a kind of prototype: a typical subject is an animate agent that has been previously mentioned in the text or dialogue. However, subjects may be inanimate, may be patients, and may be newly mentioned in the text or dialogue.

How do you identify SUBJ and OBJ? An easier question to answer is how to identify SUBJ and OBJ in any given language. In many languages you can find a cluster of behaviors associated with SUBJ or OBJ. For example, in English SUBJ is before the verb, agrees with a present tense verb, and is in *nominative case* if it is a pronoun and the verb is *finite*: *She runs/*her runs.* The cluster of behaviors associated with OBJ in English is that it is immediately after the verb and can become the SUBJ in a passive voice sentence. The cluster of behaviors associated with SUBJ and OBJ are different for each language.

Do all languages have SUBJ and OBJ? Are they universal? In most languages it is possible to find a clusters of behaviors. If a cluster of behaviors is prototypically associated with animate, discourse-old agents, you say that they are SUBJ behaviors. If the cluster of behaviors is prototypically associated with the semantic role of patient, you say they are OBJ behaviors. However, there are some languages where there is considerable controversy over SUBJ and OBJ and it is not clear that there are identifiable clusters of behaviors.

5.8 Grammatical Encoding: Who bit who?

Grammatical Encoding is the cluster of behaviors that helps you distinguish SUBJ from OBJ. This helps you understand whether the student bit the dog or whether the dog bit the student. English shows one very strong strategy for grammatical encoding: when a sentence is in its most canonical word order, the subject is before the verb and the object is after the verb. (Of course, there are many non-canonical word orders, which is what makes grammar interesting.) English also demonstrates two other mechanisms for grammatical encoding rather weakly: case marking and agreement.

English pronouns show distinctions in *case*. Case is a feature associated with nouns. Nominative case (English *I*, *we*, *you*, *she*, *he*,*it*, *they*) is associated with SUBJ. English pronouns also have three non-nominative forms. One of them can be called accusative or oblique (English *me*, *us*, *you*, *him*, *her*, *it*, *them*). Another is for possessors of noun phrases. The case for possessors is known as genitive: (English *my*, *our*, *your*, *his*, *her*, *its*, *their*). And finally there is a form for possessive (genitive) pronouns that occur alone or at the end of a phrase like *a friend of mine* (English *mine*, *ours*, *yours*, *his*, *hers*, *its*, *theirs*.

English verbs show one distinction in *agreement*. In the present tense, with a third person singular subject, verbs show a suffix *-s*.

```
Word order:
   The student bit the dog vs. The dog bit the student
Case marking:
She (biter) bit her (bitee).
*Her bit she.
Agreement:
    The student bites (agrees with third person singular) the dogs.
```

We will see that English depends primarily on word order and that other languages depend primarily on case marking and agreement. We will also see that grammatical encoding systems can be very complicated and that many of the complications are related to *animacy*, *definiteness*, *completedness*, and other semantic properties of the sentence and its components.

5.8.1 Case Marking

Tangkhul is a Tibeto-Burman language of Northeastern India (Manipur State) It has three uncontroversial case markers: *-na* (nominative), *-li* (accusative/dative/locative), *-wui* (genitive).

```
ā-na   i-li leishi
3-nom  1-acc   love
‘He loves me.’
```

In Tangkhul, nominative marking is often omitted if the subject is unambiguous.

```
ā lairik pa-li
3    book    read-prog
‘He is reading a book.’

ā-na lairik  pa-li
3-nom book   read-prog
‘He is reading a book.’
```

Cases are not grammatical relations: It may be tempting to view nominative case as the same as SUBJ. Indeed, where nominative case marking occurs, it almost always indicates a subject. There are some subjects in some languages that illustrate all behaviors of SUBJ except case. There are also languages where SUBJ is associated with more than one case, as we will see with Hindi below. *Case is a feature/behavior associated with grammatical relations, but cases are not grammatical relations.*

5.8.2 Agreement

Words have morpho-syntactic features such as number (e.g., singular, plural, dual, paucal), person (e.g. first person, second person, third person), case, and gender. Agreement is a phenomenon in which two words have to match features in some way. There can be agreement between adjectives and nouns, determiners and nouns, nouns and their possessors, and other things.

Here we will talk about agreement between a verb and its arguments and how it helps you identify who bit who. We have seen a small example of agreement between a verb and its SUBJ in English. Here is a more extreme example of agreement. Chichewa is a Bantu language of Malawi, Zambia, Zimbabwe, and Mozambique Like other Bantu languages, it has a large number of "genders" or noun classes. By the traditional count, in which singular and plural classes are counted separately, there are 18. Chichewa verbs agree with both subject and object in gender/number.

In the sentence below, *alenje* (hunters) is in noun class 2, the plural human noun class. *Njûchi* (bees) is in noun class 10, the plural of class 9. The verb has five morphemes. The first morpheme indicates the noun class of the SUBJ of the sentence, class 10. The second morpheme indicates the tense of the verb. The third morpheme indicates the class of the OBJ of the sentence, class 2. This is how we know that the bees bit the hunters. Since the verb tells us that a class 10 noun bit a class 2 noun, the nouns themselves can be in different orders with respect to the verb. Regardless of the order of *hunters*, *bit*, and *bees*, we know that the bees bit the hunters.

```
Njûchi zi-ná-wá-lum-a alenje
10-bees 10SM-past-2OM-bite-fv 2-hunters
'The bees bit the hunters.'
```

```
SVO Njûchi zi-ná-wá-lum-a alenje
OVS Alenje zi-ná-wá-luma njuchi
VOS Zi-ná-wá-lum-a alenje njuchi
SOV Njuchi alenje zi-ná-wá-lum-a
```

Hindi Case Marking: a case study

```
?Ladki ladka dekh     rah-i     hai.
 girl  boy   see.PART PROG-FEM  COP
''The girl sees a boy''
```

```
Ladki ladke ko    dekh      rah-i     hai.
girl  boy   CASE  see.PART  PROG-FEM  COP
''The girl sees a boy.''
```

*Ladki ladke ko   dekh   rah-a   hai.
 girl boy   CASE see.PART PROG-MASC COP
’’The girl sees a boy’’

?Ladka ladki dekh   rah-a   hai.
 boy   girl see.PART PROG-MASC COP
 ’’The boy sees a girl.’’

Ladka ladki ko   dekh   rah-a   hai.
boy   girl CASE see.PART PROG-MASC COP
’’The boy sees a girl’’

*Ladka ladki ko   dekh   rah-i   hai.
 boy   girl CASE see.PART PROG-FEM COP
 ’’The boy sees a girl.’’

The gender of ’’chuha’’ (mouse) is masculine.

Ladki chuha dekh   rah-i   hai.
girl mouse see.PART PROG-FEM COP
’’The girl sees a mouse.’’

*Ladki chuha dekh   rah-a   hai.
 girl mouse see.PART PROG-MASC COP
 ’’The girl sees a mouse.’’

Ladki chuhe ko   dekh   rah-i   hai
girl mouse CASE see.PART PROG-FEM COP
’’The girl sees a mouse.’’

*Ladki chuhe ko   dekh   rah-a   hai.
 girl mouse CASE see.PART PROG-MASC COP
’’The girl sees a mouse.’’

The gender of ’’chidiya’’ (sparrow) is feminine.

Ladki chidiya dekh   rah-i   hai.
girl sparrow see.PART PROG-FEM COP
’’The girl sees a sparrow.’’

*Ladki chidiya dekh rah-a hai.
girl sparrow see.PART PROG-MASC COP
''The girl sees a sparrow.''

Ladki chidiya ko dekh rah-i hai.
girl sparrow CASE see.PART PROG-FEM COP
''The girl sees a sparrow.''

*Ladki chidiya ko dekh rah-a hai.
girl sparrow CASE see.PART PROG-MASC COP
''The girl sees a sparrow.''

Ladka chuha dekh rah-a hai.
boy mouse see.PART PROG-MASC COP
''The boy sees a mouse.''

*Ladka chuha dekh rah-i hai.
boy mouse see.PART PROG-FEM COP
''The boy sees a mouse.''

Ladka chuhe ko dekh rah-a hai.
boy mouse CASE see.PART PROG-MASC COP
''The boy sees a mouse.''

*Ladka chuhe ko dekh rah-i hai.
boy mouse CASE see.PART PROG-FEM COP
''The boy sees a mouse.''

Ladka chidiya dekh rah-a hai.
boy sparrow see.PART PROG-MASC COP
''The boy sees a sparrow.''

*Ladka chidiya dekh rah-i hai.
boy sparrow see.PART PROG-FEM COP
''The boy sees a sparrow.''

Ladka chidiya ko dekh rah-a hai.
boy sparrow CASE see.PART PROG-MASC COP
''The boy sees a sparrow.''

*Ladka chidiya ko dekh rah-i hai.

boy sparrow CASE see.PART PROG-FEM COP
''The boy sees a sparrow.''

The gender of ''chidiya'' (sparrow) is feminine.

Ladki ne chidiya dekh-i.
girl ERG sparrow see-FEM.PERF
''The girl saw a sparrow.''

Ladki ne chidiya ko dekh-a.
girl ERG sparrow CASE see-MASC.PERF
''The girl saw a sparrow.''

*Ladki ne chidiya dekh-a.
girl ERG sparrow see-MASC.PERF
''The girl saw a sparrow.''

*Ladki ne chidiya ko dekh-i.
girl ERG sparrow CASE see-FEM.PERF
''The girl saw a sparrow.''

The gender of ''chuha'' (mouse) is masculine.

Ladki ne chuha dekh-a.
girl ERG mouse see-MASC.PERF
''The girl saw a mouse.''

*Ladki ne chuha dekh-i.
girl ERG mouse see-FEM.PERF
''The girl saw a mouse.''

Ladki ne chuha ko dekh-a
girl ERG mouse CASE see-MASC.PERF
''The girl saw a mouse.''

*Ladki ne chuha ko dekh-i.
girl ERG mouse CASE see-MASC.PERF
''The girl saw a mouse.''

The boy saw a sparrow.
Ladke ne chidiya dekhi.
*Ladke ne chidiya dekha.
Ladke ne chidiya ko dekha

```
*Ladke ne chidiya ko dekhi.

The boy saw a mouse.
Ladke ne chuha dekha.
*Ladke ne chuha dekhi.
Ladke ne chuhe ko dekha.
*Ladke ne chuhe ko dekhi.
```

5.9 World Atlas of Language Structures

WALS (World Atlas of Language Structures), `https://wals.info/`, has about one hundred fifty chapters on morpho-syntactic variation. Each chapter is about one parameter (for example, word order of verb and subject) of variation and comes with data about a sample of languages that have different values for the parameter (for example VS, SV, or no dominant order). The values are plotted on a map with colored dots representing the different values of the parameters for the sample of languages. The entire atlas is downloadable as a matrix of language names (about 2500 languages) and parameters (about 150). Since each chapter covers a different sample of languages, the matrix is sparsely populated.

Following are some links to WALS chapters.

5.9.1 Word Order

Chapters 81 to 97 are about word order, and whether or not languages are consistently head-initial or head-final in clauses and noun phrases.

5.9.2 Locus of Marking

Chapter 23 (`https://wals.info/chapter/23`) contrasts *head marking* and *dependent marking* in clauses. Head marking corresponds roughly to what we called *agreement* above. The verb is the head of the clause and the markers (the agreement morphemes) are on the head. Dependent marking corresponds roughly to what we called *case marking*. Nouns are dependents of verbs and case markers are attached to nouns.

Chapter 24 (`https://wals.info/chapter/24`) contrasts head marking and dependent marking in possessive noun phrases. In the possessive noun phrase *the student's book*, *book* is the head and *student* is the dependent. A case marker on a possessor dependent like *student* is called *genitive case*. A marker on *book* is called a possessive marker.

Chapter 25 (`https://wals.info/chapter/25`) examines whether languages are consistently head marking or dependent marking in clauses and noun phrases.

5.9.3 Alignment (Ergative-Absolutive and Nominative-Accusative)

Chapters 98 (`https://wals.info/chapter/98`) and 99 (`https://wals.info/chapter/99`) are about Nominative-Accusative and Ergative-Absolutive case marking in full noun phrases and pronouns.

5.9.4 Possession in clauses and noun phrases

Chapter 117 (`https://wals.info/chapter/117`) is about clauses that express possession. In addition to the two strategies we have looked at (*I have a book* and *A book is at me*), there are others that can be paraphrased as *I am with book*, *My book exists*, and *As for me, there is a book*.

Chapter 24 (`https://wals.info/chapter/24`) is about head and dependent marking in possessive noun phrases.

Chapter 86 (`https://wals.info/chapter/86`) is about the order of the possessor and the possessed.

Chapter 59 (`https://wals.info/chapter/59`) is about whether the language has a different morpho-syntactic expression of alienable and inalienable possession or kinship in contrast to non-kinship possession.

5.9.5 Comparative sentences

Chapter 121 (`https://wals.info/chapter/121`) is about morpho-syntactic strategies for comparing two entities with respect to a scalar value such as height, width, or price. The English strategy (X is Z-er than Y, or X is more Z than Y) is rare. More common strategies can be paraphrased as *X is Z, surpasses Y*, *X is Z from Y*, *X and Y, X is Z*, and others.

5.10 Pear Film assignment

The Pear Film (`http://pearstories.org/`) is a six-minute film without words that was created by linguists for the purpose of studying discourse and morpho-syntax. Experimental subjects narrated the film without knowing the purpose of the experiment.

6 Formalisms for Syntax

6.1 Introduction

In this chapter we turn to the discussion of *syntactic formalisms*. We have previously encountered syntax in Chapter 5 and we have previously used formalisms such as finite state transducers for describing phonological and morphological rules. This chapter differs in some ways. In Chapter 5 we looked at constructions that pair form and meaning, and we presented examples of morphosyntactic strategies for pairing form and meaning. In discussing syntactic formalisms, we will not focus on meaning, but we are not forgetting about it: we are looking at how words are arranged in hierarchical structures so that they support the expression of meaning. We will present formalisms for describing the hierarchical structure of words in sentences and show that they need to be more powerful (in a way we will define) than the formalisms that describe phonology and morphology.

What to look for in this chapter:

- words are grouped into units in sentences and the units are arranged in hierarchical structures
- Formal languages represent a potentially infinite set of sentences that are built from a finite vocabulary, using some kind of production rules, often called rewrite rules, and other legal moves to derive sentences.
- Syntactic formalisms act as accept-reject machines: providing a well-formed derivation for each sentence in the language and a failed derivation for sequences of vocabulary items that are not in the language.
- Syntactic formalisms are more powerful that finite state formalisms that we use for phonology and morphology.

6.2 Does syntax exist, and if so, why?

Until the mid 20th century, generations of people were traumatized by the school subject of sentence diagramming – breaking sentences into parts such a subjects, predicates, and modifiers. There were good reasons for teaching sentence diagramming. For example, if you can name the parts of a sentence,

you have a vocabulary for talking about how to structure sentences for good writing. For example, a writing teacher might be able to better define a run-on sentence: a run-on sentences has two finite verbs, each with a nominative subject and no subordinating or coordinating conjunctions. This is much more precise than saying that a run-on sentence is bad because it contains two complete thoughts. Also, understanding terms like subject and predicate can help in learning other languages, especially those with case marking. However, people suffered from sentence diagramming and wondered what it had to do with clear communication. Why was it necessary to learn it? And eventually it was dropped from school curricula (at least in the US).

[Insert graphic of a traditional sentence diagram.]

Similiarly, the field of natural language understanding, often questions the necessity of syntax. In the early days of NLU, proponents of *semantic grammars* ([1]) claimed that they could write computer programs that could analyze the meaning of a sentence without using any syntax. (In fact, they did use syntax covertly in referring to the order of words and other structural notions. They just did not use syntax in the way that syntacticians did at the time.) More recently in NLU, there are many tasks that can be performed by neural networks that are not explicitly trained on syntactic structures. Raising the question of whether syntax is in the hidden layers as an epiphenomenon, or whether it is just not necessary.

[1]

Even among human language scientists, the existence of syntax is controversial. There is ongoing debate between *functionalists* and *generativeists* with respect to syntax. Functionalists believe that syntax is a side effect of human communication and that it cannot be separated cleanly from the process of communication.

Generativists, on the other hand, tend to believe that syntax is an autonomous component of human language and that it exists independently of meaning and communication. This is called the Autonomy of Syntax hypothesis. Medical evidence supporting the generative view includes the existence of aphasias (brain injuries) that affect only syntax or only meaning.

In addition, generativists pose some apparent paradoxes. One kind of paradox has to do with sentences that seem to have similar meanings but... (*Who did you see Sam and, *Who do you think that left) To be completed.

The second type of paradox is the Logical Problem of Language Acquisition: how can a baby learn complex structures with imperfect input (baby talk, fragments, and other disfluent language). (*Is the student who tall is smart) To be completed.

6.3 *Chunks of sentences*

We will not take a position on Autonomy of Syntax Hypothesis in these lecture notes. (we are only inclined to believe it a tiny bit, but even if syntax is epiphenomenal, investigating it scientifically is still important.) We will

present syntax as a mechanism for grouping words into chunks that (1) result in human languages being *productive* and usually *recursive* and (2) that can be combined to make meanings in a predictable way *compositionality*.

Consider something you might say to your smartphone: *Send a text to mom*. The speech processing software will recognize *send a text* as a chunk of words that invokes the texting app. The next chunk of words, *to mom*, will indicate a recipient from your contacts list. So maybe there is a kind of grammar that allows an app invocation followed by a contact. Now any app invocation such as *send a text*, *send an email*, or *call* can be combined with any name from your contacts list, allowing your phone to correctly interpret things you have not said before. (Relate better to the property of productivity.)

Now consider the two sentences below. In the first one, *the old man* is a chunk of words that refers to something. In the second sentence, *the old* refers to something and *man the boats* is an action. (To be completed with pictures that I saved in Downloads on my home computer old-man-boat-<1-4>.jpeg. Also look up "old man sleeping".)

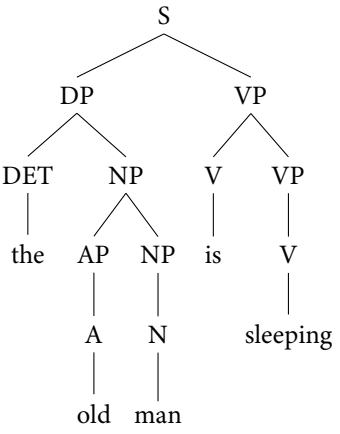
Insert pictures

(13) a. The old man is sleeping.

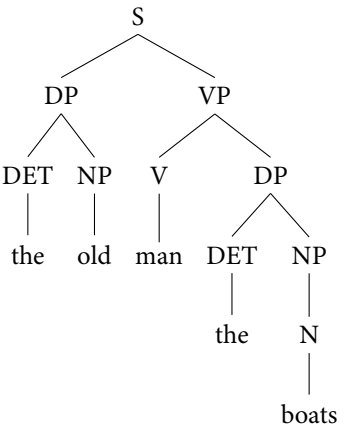
b. The old man the boats.

6.4 *Hierarchical Structures and Rewrite Rules*

(14) a.



b.



The chunks of sentences are called *constituents* and are represented hierarchically in trees such as the ones shown here. In the first tree, *the old man* and *is sleeping* are the main constituents. In the second tree the main constituents are *the old* and *man the boats*. We will discuss constituents again later in this chapter.

The nodes in the trees are labelled. Labels that are in uppercase letters are called *non-terminal symbols*. The non-terminal symbols represent either parts of speech (categories of words) or categories of phrases like noun phrases and verb phrases. This table defines the non-terminal symbols in these trees.

S	Sentence
DP	Determiner Phrase
NP	Noun Phrase
AP	Adjective Phrase
VP	Verb Phrase
DET	Determiner
A	Adjective
N	Noun
V	Verb

What makes a symbol non-terminal is that it can *expand* or *rewrite* as something else. For example, S rewrites as DP followed by VP. VP rewrites as V followed by VP or as V followed by DP.

Labels that do not start with uppercase letters are called *terminal symbols*. Terminal symbols do not expand or rewrite. For the purpose of human language syntax, terminal symbols are words. A complete tree has an S at the top and each branch of the tree ends in a terminal symbol.

We can model hierarchical trees with *rewrite rules* such as the ones below. Don't worry if you don't like the way we have drawn the trees and the rules or if you would do it differently. The whole idea is that you can make whatever trees you want by changing the rewrite rules. However, in this chapter, we will explain how linguists make decisions about node labels and tree structures,

and you may change your mind by the end of the chapter.

```
S --> DP  VP
DP --> DET  NP
NP --> AP  NP
NP --> N
AP --> A
VP --> V  VP
VP --> V  DP
```

Some terminology: You pronounce these rules as follows: "S rewrites as DP VP", "DP rewrites as DET NP", etc. You could also say "S expands to DP and VP". However, linguists usually pronounce the rewrite arrow with the words "goes to". So they say "S goes to DP VP". Pronouncing the rewrite arrow as "goes to" is just something that happened over time as linguists worked together.

In addition to thinking of the rewrite rules as expansion rules, you can think of them as reduction rules. So you could say that DP followed by VP reduces to S.

6.5 *The Chomsky Hierarchy and Generative Capacity/Power*

This section is crash course on formal language theory. A formal language is something like a regular expression or a programming language. There is a whole field of formal language theory. In this chapter we will look at how to apply formal language theory to human languages.

In 1957, Noam Chomsky (building on the work of others?) discussed three categories of formal languages: finite state, context free, and context sensitive. We will look at each of these, and we will also look at the idea of *generative capacity*. Not all formal languages work for human languages. By looking at the idea of generative capacity, we will see which of the three types of formal languages has the generative capacity for human language.

You can think of a formal grammar as a mechanism for re-writing a string of characters/words or building a tree.

A grammar for a formal language consists of *terminal symbols* and *non-terminal symbols*.

When we use formal grammars to parse sentences in human languages, the terminal symbols are (usually) the words of the language. But for now, the only "words" we will use as terminal symbols are the letters "a" and "b".

The non-terminal symbols will be written in capital letters. A special non-terminal symbol is S, the starting symbol, which can also mean *sentence*.

6.5.1 *Finite State Grammars*

We will start with finite state grammars. To make a *rule* you write an arrow pointing to the right. We will call this a *rewrite arrow* or just an arrow. To

the left of the arrow, you write one non-terminal symbol. To the right of the arrow, you put a combination of terminal and non-terminal symbols, but you have to follow very strict guidelines. You can have one non-terminal symbol and the terminal symbols need to be all before or all after the non-terminal symbol.

If S, A, and B are non-terminal symbols, and *a* and *b* are terminal symbols, and we decide to put the terminal symbols to the left of the terminal symbol on the right side of the arrow, here are some legal finite state grammar rules:

```
S --> b A
A -->  a  A
A -->  a
```

Now we will look at the concept of a grammar *generating* a sentence, a grammar generating a language, and at the idea of a *derivation* of a sentence.

The three rules above happen to work together as a *grammar*. The grammar can generate *sentences*. You generate sentences by *rewriting* a terminal symbol using the rules.

To generate a sentence you always start with S. The arrow means that you can re-write the left side of the arrow with the right side of the arrow. The first rule tells you that you can rewrite an S as a *b* followed by an A. The sequence of re-writing operations is called a *derivation*. Our first derivation starts like this:

```
Step 1:  S
Step 2:  b  A
```

A derivation is finished when there are no non-termial symbols left. We have so far rewritten an S into a b and an A. But A is a non-terminal symbol, so we aren't done yet. The simplest way to end the derivation is to use the third rule to rewrite an A as an *a*.

```
Step 1:  S
Step 2:  b  A
step 3:  b  a
```

Now we have a complete derivation of the "sentence" *b a*.

We can derive longer sentences by using the second rule.

```
Step 1:  S               start symbol
Step 2:  b  A            rule 1
step 3:  b  a  A         rule 2
Step 4:  b  a  a         rule 3
```

Let's use Rule 2 twice:

```
Step 1:  S               start symbol
```

```
Step 2:  b  A          rule 1
step 3:  b  a  A       rule 2
Step 4:  b  a  a  A    rule 2
Step 5:  b  a  a  a    rule 3
```

So, now we have derived the sentences *b a*, *b a a* and *b a a a*. If you think about it, you can see that this grammar generates a specific *language*, one where all the sentences start with a *b* and after the *b* there are one or more *a*'s.

The language is the set of sentences you can generate or derive using the grammar. The language generated by this grammar contains an infinite set of sentences. You can keep using Rule 2 to get as many *a*'s as you want. We can call our language ba^n (one b followed by one or more a's) where n is any positive integer.

It is important to notice that our three rules do not generate any possible combinations of a's and b's. For example, the grammar cannot generate a sentence that starts with an *a* or a sentence that starts with more than one *b*. When you write a grammar for a human language, you write rules such that you can derive all of the sentences that are grammatical and there is no way to derive sentences that are not grammatical. We will continue to look at this idea of deriving all and only grammatical sentences throughout this chapter.

Generative Capacity or Power refers to the *kinds of languages* you can derive/generate/describe. But we won't talk about that just yet. We will move on to Context Free Grammars. When you see what kinds of languages Context Free Grammars can generate, you will be able to see what kinds of languages Finite State Grammars cannot generate. However, if you have done any programming, you might be interested to know that Finite State Grammars can describe any language you can write with a regular expression. And of course, anything you build with a finite state automaton can be generated by a Finite State Grammar.

6.5.2 Context Free Grammars

Like Finite State Grammars, Context Free grammars also have terminal symbols, non-terminal symbols, and rewrite arrows. But the guidelines allow more types of rules. It is still the case that each rule can have only one non-terminal symbol to the left of the rewrite arrow. However, on the right side of the rewrite arrow, you can have any combination of terminal and non-terminal symbols in any order.

7 Discourse and Pragmatics

Phonology is highly constrained. In any language, there are a fixed number of phonemes and they can combine in a fixed number of ways. There are clear rules that govern the relationship between more abstract representations and surface representations. Morphology may seem a little less structured, but there are a fixed set of derivational and inflectional processes with restricted application. Syntax is freeer still—it allows a finite lexicon and a finite grammar to generate and infinite set of sentences. On the other hand, these sentences are governed by well-defined grammatical principles that structure most aspects of phrases. By the time we transcend sentences and move up to the level of discourse, we might assume that there is no structure at all—that anything goes. However, a cursory examination of the issue shows that this is not the case.

Consider the following sentences:

- However, the incompetence of his managers insures him a steady, six-figure income.
- He only knows Java.
- Worse still, he always optimizes the outermost loop first.
- Eric is a pathetic programmer.

There are $4! = 24$ ways to permute these sentences, but only 2 of them are acceptable as discourses:

1. Eric is a pathetic programmer.
2. He only knows Java.
3. Worse still, he always optimizes the outermost loop first.
4. However, the incompetence of his managers insures him a steady, six-figure income.

and

1. Eric is a pathetic programmer.

2. However, the incompetence of his managers insures him a steady, six-figure income.
3. He only knows Java.
4. Worse still, he always optimizes the outermost loop first.

Contrast this, for example, with the following:

1. Eric is a pathetic programmer.
2. Worse still, he always optimizes the outermost loop first.
3. However, the incompetence of his managers insures him a steady, six-figure income.
4. He only knows Java.

Or this:

1. However, the incompetence of his managers insures him a steady, six-figure income.
2. Worse still, he always optimizes the outermost loop first.
3. Eric is a pathetic programmer.
4. He only knows Java.

Clearly, there is some structure principle that governs how these sentences—all of them syntactically well-formed by themselves—can be ordered. There is something that distinguishes discourse, a coherent sequence of sentences or utterances, from a randomly ordered set of sentences (even one that makes semantic sense). We call this COHERENCE.

Discourse is, in some ways, harder to study than syntax, morphosyntax, morphology, phonology, and phonetics because, at some level, one cannot get it right without reference to all of the more constrained "lower" levels. On the other hand, it often seems less formally challenging and more hospitable to beginners than other fields in the language sciences. Sometimes the term DISCOURSE ANALYSIS is even used to refer to unprincipled but opinionated commentary on what other people have said. However, there is real content in the best research on discourse and it is of vital importance to artificial intelligence. It is not sufficient for intelligent systems to generate well-formed sentences with the right meanings; it is essential that they be able to produce well-formed discourses (including well-formed dialogues). Likewise, no artificial intelligence will really pass the turning test until it can respond appropriately to the subtleties of discourse structure.

However, there is more to communicative context than stringing sentences together in an a coherent way. What we do with language often extends far beyond the semantic interpretation of what we say. The field that studies what we do with language, which lies at the intersection of linguistics and philosophy of language, is called PRAGMATICS.

7.1 Cohesion and Coherence

There are two ways that sentences hold together. In line with their general penchant for confusion terminology, linguists decided to refer to both of these with nouns related to *cohere*. These are COHESION and COHERENCE. Cohesion refers to the way that some pairs of contiguous sentences adhere to one another to a greater degree than others. The existence of discourse segments like paragraphs is an example of cohesion. In a sense, cohesion is like constituency—sentences that have a very high degree of cohesion form constituents and these can, in turn, be grouped into higher-order units until one reaches a node that dominates the whole discourse. But cohesion only concerns the *degree* of relationship between discourse segments. It does not concern the *type* of relation between such segments. The property of having meaningful relationships among its parts makes a discourse COHERENT.

7.1.1 Discourse segmentation and cohesion

When you are writing, you may sometimes have difficulty deciding where to divide text into paragraphs. On the other hand, it is not true that sentences are uniformly related in equal degree to the preceding and following sentence. What makes two sentences related? That is to say, what gives the cohesion? Here are some factors (linguistic devices that tie sentences together):

- LEXICAL COHESION. Two sentences go together if they share vocabulary items, synonyms, or hypernyms.

  (15) Peel, core and slice the **pears and the apples**. Add **the fruit** to the skillet.

- ANAPHORA. A different kind of "back-reference" where a pronoun refers back to, or is co-referent with, an earlier pronoun or noun phrase.

  (16) a. David$_i$ taught the lecture with grace and aplomb. He remained calm despite the fact that he$_i$ didn't know what he$_i$ was talking about.

  b. Thomas$_i$ left Amazon$_j$ to work at Google. He decided that they$_j$ were a cult. They$_j$ also didn't pay him$_i$ as well.

- COHESION CHAINS. Multiple cases of lexical cohesion and anaphora may tie more than one sentence together in a cohesive group.

- DISCOURSE MARKERS. Certain words and phrases that signal either continuations or discontinuities in topic. For example, consider *however* in the following example:

  (17) He ultimately accepted the job at Microsoft Research. **However**, he always wished he had taken a chance with the start-up.

The discourse marker signals that the second sentence is directly connected, in terms of topic, to the first. There are many discourse markers in English, some of which are single words (like conjunctions) and some of which are multi-word expressions.

Boundaries between discourse segments exist when cohesion, as measured by these metrics, dips.

7.1.2 Coherence relations

Cohesion is about how related to discourse segments are; coherence is about how they are related. It can be characterized by meaningful relations between sentences or other units of discourse. Take the following examples [1]:

(18) a. John hid Bill's car keys. He was drunk.

b. John hid Bill's car keys. He likes spinach.

The two examples differ in coherence. It is easy to find a meaningful relationship between John hiding Bill's car keys because he (Bill) was drunk[2]. The second sentence is an EXPLANATION for the first sentence. In contrast, it is difficult to decide what relationship could possible hold between John's hiding of Bill's car keys and one of them liking to eat the leaves of *Spinacia oleracea*. That is to say, the first example is a coherent discourse, but the second example is not.

Some coherence relations from Hobbs (1979):

Result Infer that the state or event asserted by S_0 causes or could cause the state or event asserted by S_1.
Lakshmi had an original thought. She was fired.

Explanation Infer that the state or event asserted by S_1 causes or could cause the state or event asserted by S_0.
Lakshmi was fired. She had had an original thought.

Parallel Infer $p(a_1, a_2, ...)$ from the assertion of S_0 and $p(b_1, b_2, ...)$ from the assertion of S_1, where a_i and b_i are similar, for all i.
Lakshmi was fired; Krishna was promoted.

Elaboration Infer the same proposition P from the assertions of S_0 and S_1.
Lakshmi was fired. Her employment was terminated abruptly and without warning.

Occasion A change of state can be inferred from the assertion of S_0, whose final state can be inferred from S_1, or a change of state can be inferred from the assertion of S_1, whose initial state can be inferred from S.
Xudong opened a text editor. He starting pounding out a script similar to those he had written a thousand times before.

[1] Jurafsky, D. and J. H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). Upper Saddle River, New Jersey: Prentice-Hall

[2] It would also be easy enough to find a reasonable relationship between the two sentences if it was John who was drunk.

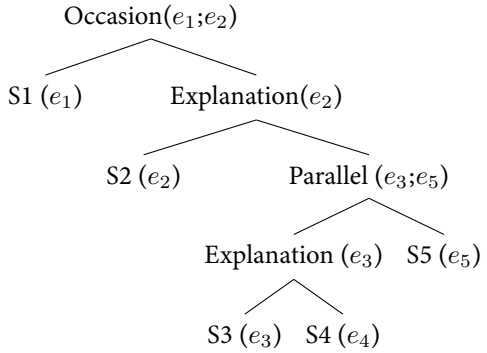
These are not all of the possible coherence relations. Indeed, in the Rhetorical Structure Theory treebank, there are a great many more. However, these are sufficient to illustrate the idea of discourse parsing. The insight behind this task is that coherence relations hold not just between sentences but between larger segments of discourse and that these relationships have a hierarchical structure. Consider the following discourse[3]:

[3] Jurafsky, D. and J. H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). Upper Saddle River, New Jersey: Prentice-Hall

(19) John went to the bank to deposit his paycheck. (S1)
He then took a train to Bill's car dealership. (S2)
He needed to buy a car. (S3)
The company he works for now isn't near any public transportation. (S4)
He also wanted to talk to Bill about their softball league. (S5)

The structure of this discourse is not linear, but it is coherent. The following parse tree illustrates the hierarchical relationships between the "constituents" in this discourse:

(20)



S1 is the occasion for the rest of the discourse. Sentences S3, S4, and S5 form, in turn, the explanation for S2. The remainder of the discourse gives two parallel reasons for S2: the need to buy a car and the desire to talk to Bill. The first of these, consisting of S3 and S4, is internally complex—S4, about public transportation, serves as an explanation for S3, about needing to buy a car.

When discourse coheres, it is possible to find relations like these biding together all parts of the discourse. In other words, discourse is not as free and easy as the layperson may suppose. Intelligent systems whose behavior does not reflect correct coherence relations will be perceived by human observers as talking nonsense even if the individual sentences that they produce are well-formed.

7.1.3 Rhetorical Structure Theory

+ Rhetorical Structure Theory (RST) is an elaborate system for annotating coherence relations in discourse. It has served as the basis for an extensive discourse treebank (the RST Treebank) and for various computational approaches to discourse.

7.2 Old and New Information

All languages have ways of distinguishing between entities and propositions that have been referenced earlier in discourse and those that are newly introduced by the current discourse unit (like a sentence or utterance). Old information can be marked directly as DEFINITENESS, indicated either by an affix or a determiner. It may be marked by the use of pronouns. It may be expressed via word order, including so-called topicalization. It may also be encoded by other syntactic devices. In fact, it is often encoded by a combination of these devices. However it is encoded, though, it is an essential part of discourse structure that relates to cohesion, coherence, and anaphora.

Consider the following example:

(21) The dog bit a girl.

This sentence demands a particular discourse context—one in which the speaker can presume that the listener has prior knowledge of *the dog*. This is encoded in two ways, one explicitly and one by correlation:

1. The use of the definite article *the* marks *dog* as old information (that is, an entity or proposition that has been introduced earlier in discourse).
2. The appearance of *the dog* as the subject of a sentence, which is correlated—though imperfectly—with the notion of TOPIC[4]

[4] Topics are what a discourse is about. After they are introduced in a discourse, they become old information.

By contrast, *a girl* is marked as new information. It would be very strange to say

(22) A girl walked down the street. The dog bit a girl.

if both instances of *a girl* refer to the same entity. However, both of the following two examples are fine:

(23) A girl walked down the street. The dog bit the girl.

(24) A girl walked down the street. The dog bit her.

Across languages, a common strategy for marking old information is what is called TOPICALIZATION. In this strategy, a noun phrase or clause referring to an entity or proposition that has already been introduced in discourse, is "moved" to a special location, often the right periphery of the clause. Consider the following exchange:

(25) Prantl's, I find to be the best bakery in Pittsburgh.

The response does not sound especially good as an introduction of bagels and lox into the discourse, but sounds much better if some other sentence has already performed this nefarious deed. The fronted noun phrase, in topicalization, is always discourse old.

Other languages have other devices for encoding old and new information. In Chinese, there are (at least) two ways of saying 'I ate rice':

(26) a. 我 吃 饭 了
1SG eat rice PERF
'I ate rice.'
b. 我把饭吃了
1SG BA rice eat PERF
I ate rice.'

In a corpus study concentrating on two verbs[5], it was shown that the best predictor of the use of the second construction is information structure: it marks the "object" (like 'rice') as old information. In light of this fact, we might be tempted to translate (26a) as 'I ate (rice)' and (26b) as 'I ate the rice'.

[5] Yao, Y. (2014). Predicting the use of ba construction in mandarin chinese discourse: A modelingstudy with two verbs. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*

7.3 Anaphora and Coreference

7.4 Dialogue

7.4.1 Human-human dialogue

7.4.2 Human-computer dialogue

7.5 Pragmatics: Language Use in Context

Semantics concerns the truth values of propositions out of context. Pragmatics, by way of contrast, concerns the felicity of utterances within context. It is about how we do things with language. Consider the following utterance:

(27) It certainly is loud in here!

On a semantic level, it seems to be a factual statement about the volume of noise in a room. However, suppose that this sentence is uttered by a person, sitting at her desk with her hands on her head, while her office mates chatter incessantly. The utterance then can be seen as a passive aggressive attempt to silence the other people in the room. Semantically, it means "the volume of sound in this room is high." Pragmatically, it means "be quiet."

7.5.1 Speech Act Theory

When we do things with language, we engage in what are called SPEECH ACTS. There are a great many types of speech acts, a few of which are listed here:

- Statement
- Question
- Command
- Promise

There are a variety of different linguistic devices that are used to mark speech acts of this kind. For example, in Hmong (to a lesser extent, Chinese) and many languages of Southeast Asia, speech acts are marked with special particles that occur at the end of sentences. In English, they may be marked by special syntactic devices (wh-movement for wh-questions, for example).

Philosophers of language and semanticists have spent the most time studying statements, which are special because they have truth values (that is to say, they may be either true or false). Other speech acts are different in that do not. You cannot very well say that a question is true or false, or that a command is true or false. These acts are significant not for what they mean but for what they do.

How do we get rigorous about speech acts, then, if most of them do not have true values? One way is with felicity conditions. A speech act may or may not be appropriate depending on the context in which it is uttered (in particular, by whom it is uttered). Take the statement:

(28) I bequeath this watch to my brother.

I cannot really get away with saying this if I am not the owner of the watch. In this case, my speech act would be said to be INFELICITOUS. However, if I say

(29) I take this man to be my husband.

and I am at a wedding, am to be married to "this man," and am at the appropriate stage in a wedding ceremony, this utterance may be FELICITOUS. Likewise,

(30) I declare war on Saudi Arabia!

Is infelicitous if most of us say it, but is felicitous (but perhaps a bit unfortunate) if a head of state says it.

These sentences are of a class that the philosopher of language J. L. Austin called PERFORMATIVES. They are utterances that perform an action. You can often tell if a sentence is a performative by adding *hereby* to it. You can say:

(31) a. I hereby name this ship the Queen Elizabeth.
b. I hereby take this man to be my husband.
c. I hereby bequeath this watch to my brother.
d. I hereby declare war.

However, you cannot say:

(32) a. Birds hereby sing.
b. There is hereby fighting in the Ukraine.

That is to say, the sentences in (31) are performatives but the sentences in (32) are not.

7.5.2 Locution, illocution, perlocution

Austin divided the pragmatic force of an utterance into thee facets:

- LOCUTION. The act of saying words.
- ILLOCUTION. The action performed *in* saying those words.
  - Ask
  - Promise
  - Command
  - Etc.
- PERLOCUTION. The effects of those words on the listener. The action performed *by* saying those words.
  - Persuade
  - Convince
  - Scare
  - Elicit an answer
  - Etc.

The philosopher of language John Searle (1975) set up the following classification of illocutionary speech acts:

- ASSERTIVES. speech acts that commit a speaker to the truth of the expressed proposition, e.g. reciting a creed
- DIRECTIVES. speech acts that are to cause the hearer to take a particular action, e.g. requests, commands and advice
- COMMISIVES. speech acts that commit a speaker to some future action, e.g. promises and oaths
- EXPRESSIVES. speech acts that express the speaker's attitudes and emotions towards the proposition, e.g. congratulations, excuses and thanks
- DECLARATIONS. speech acts that change the reality in accord with the proposition of the declaration, e.g. baptisms, pronouncing someone guilty or pronouncing someone husband and wife"

Searle also noted that speech acts can be indirect:

(33) Can you pass the salt?

This utterance has the form of a question, but the actual speech act is a directive.

7.6 Gricean Maxims

7.7 Exercise: Cohesion and Paragraph Segmentation with Linguistic Features

In this assignment, you will use linguistic features to divide Wikipedia articles into paragraphs. The ground truth will be the original paragraph segmentation. Pedagogical goals:

- Understand how formal factors contribute to discourse cohesion
- Explore the relationship between topic and cohesion on a large scale
- Discover formal factors that predict the cohesion of discourse and the continuity of topic

The task is to segment the test set into paragraphs using linguistic features and to explain what features were chosen and why they work.

7.7.1 Data Set

You will be provided with a respectably sized data set called "paraseg." It is divided into training, development, and test sets (randomly selected). Data is formatted with one sentence per line (with a blank line between documents). After each sentence is a tab, followed by a label: either `B` or `I`. `B` occurs at the beginning of paragraphs (it immediately follows a paragraph boundary) and `I` occurs elsewhere.

7.7.2 The Task

You will be required to write a program that can segment the test set into paragraphs. You need to be able to beat a random baseline on the WinDiff metric (not too hard). You are not allowed to use an end-to-end system. You must use linguistic features explicitly (otherwise, the exercise will be of no pedagogical value).

7.7.3 What to Hand In

1. The test data with your hypothesized labels
2. A write-up with the following items
   (a) Description of the operationalization of at least three linguistic features for discourse segmentation
   (b) A discussion of why each of these features should help to identify discourse boundaries or discourse cohesion

8 Social Meaning in Language

In recent years, social media services have come under increased pressure to regulate hate speech, and other kinds of offensive speech, on their platforms. This can be done by human moderators, under some conditions. However, human moderation has its limits. It is expensive, for example, and this expense means that it cannot be applied to every post (only to posts that are reported as offensive). Even if every post could be vetted, there is no guarantee that the human moderators would be consistent in their evaluation of the posts. It would be highly desirable if a computational means could be devised to automatically flag posts as hate speech. However, this too is complicated.

Hate speech or offensive language cannot be detected simply by matching lexical items. For example, the word *kike* may occur in antisemitic hate speech. It may also occur in descriptions of acts of antisemitic hate that are not actually offensive themselves:

> The book was defaced with swastikas and had the word "kike" written on several pages.

An acceptable hate speech classifier would have to classify such examples as negative (not hate speech) even though they contain a potential feature for hate speech.

This raises a deeper question: what makes an utterance or a written passage hate speech? If its not the lexicon, perhaps it is the morphology or the syntax? If not that, surely it must be semantic or discourse factors that distinguish ordinary speech from hate speech.

It turns out that, ultimately, hate speech—and language in general—is a social phenomenon. While language has a rich internal structure (a fact we hope we have illustrated in preceding chapters), the function of language is always social and it is impossible to understand what it really means without making reference to its embedding in a social context.

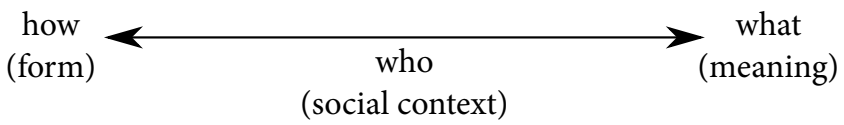


Figure 8.1: The relationship between how, what, and who

8.1 Types of Linguistic Variation

A single language can vary in every dimension in which languages can vary from one another. Certainly, varieties of a language can vary in the words that they use, but also in how those words are pronounced (on both a phonetic and a phonological level, how words are structured, and how words combined to form phrases and clauses.

8.1.1 Phonetic variation

Some of the best-studied examples of sociolinguistic variation are phonetic. This is because they are easy to measure in an objective fashion and are simultaneously conspicuous and invisible. An important example of phonetic variation comes from a study of the English of Martha's Vineyard, an island off the coast of Massachusetts, by influential sociolinguist William Labov. Labov found that some people on the island produced the diphthongs in words like *bite* and *bout* as [əj] and [əw] rather than [aj] and [aw]. To measure this, he interviewed a sociologically balanced sample of residents of Martha's Vineyard. He used interview questions that did not explicitly tell participants what he was investigating, but which were designed to elicit words that contained the diphthongs of interest. He found that the [əj] and [əw] variants were more common among people who identified as from Martha's Vineyard than those who identified as from Massachusetts and were more common in the rural part of the island than the more populated side of the island occupied by summer resorts. Men were more likely to use the non-standard variant that women. The people who used the non-standard variant most were male fisherman living on the coast of the rural side of the island.

8.1.2 Phonological variation

Dialect surveys of American English (and other languages) have shown extensive variation in the number and relationship among phonological categories in different varieties. A drinking song from the University of California, Berkeley says:

> Oh, they had to carry Harry to the ferry,
> And the ferry carried Harry to the shore;
> And the reason that they had to carry Harry to the ferry
> Was that Harry couldn't carry any more.

In the dialect of the author of this song, *carry*, *Harry*, and *ferry* all rhyme; in many other dialects of English, *carry* and *ferry* have different vowels (/æ/ and /ɛ/). In American English, these are confined primarily to the Northeast. In the rest of the US, and in almost all of Canada, /æ/ has merged with /ɛ/ before /ɹ/.

A similar type of merger has happened in the English of Pittsburgh, Pennsylvania, and also in the Great Basin of the western US. In these dialects, /i/

and /ɪ/ have merged to /ɪ/ before /l/ so that *still* and *steel* are pronounced the same.

Large-scale studies of phonological variation have typically been carried out with telephone or internet surveys wherein respondents are asked to pronounce a list of words or to assess whether pairs of words are pronounced the same.

8.1.3 Morphological variation

Languages vary morphologically in at least two ways: they may differ in the inventory of morphological constructions and they may differ in the structure of morphological paradigms.

As an example of the first type, certain varieties of English used in informal REGISTERS allow the use of noun-verb compounds like *eye-fuck*, where the noun is typically denotes a location or instrument pertaining to the event encoded by the verb. Such compounds often concern sex and violence, so it is not surprising that they are confined largely to varieties of English used in contexts where the open discussion of such topics is accepted[1]. However, there is no reason that similar noun-verb compounds referring, say, to finance, could not be used in different registers of English.

[1] See, however, examples like *Facebook-friend* (as in *I Facebook-friended her* or *Facebook-stalk*

Paradigmatic variation is also found in English. An example includes the paradigm for *to see*.

Table 8.1: Paradigms of *to see*

	DIALECT A	DIALECT B
INFINITIVE	see	see
NON-PAST	see	see
PAST	**saw**	**seen**
PRESENT PARTICIPLE	seeing	seeing
PAST PARTICIPLE	seen	seen

8.1.4 Lexical variation

Dialects, both regional and purely social, are known to differ in what words are used for particular objects or concepts. A famous example, in North America, is the word used for sweet carbonated drinks. A map of the geographic distribution of *soda*, *pop*, and *coke* is given in Figure 8.2.

8.1.5 Syntactic variation

Syntactic diversity is sometimes more subtle than phonetic, phonological and morphological diversity. It has also been less studied. Pioneering work in this area also came from William Labov in his work on African American Vernacular English (sometimes called Black English), a language variety that differs in interesting ways from other varieties of American English. He

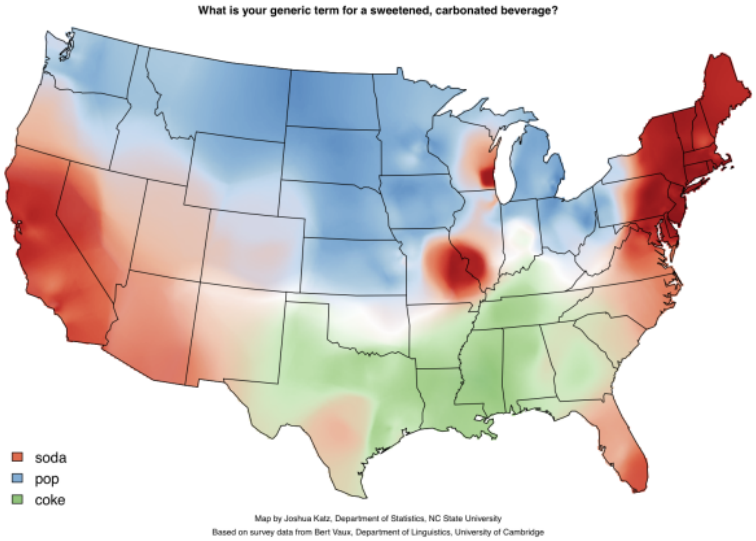


Figure 8.2: Distribution of *soda*, *pop*, and *coke* in the US

showed that the differences between AAVE and standard American English were not due to deficiencies in AAVE, any more than syntactic differences between English and Spanish were due to deficiencies in English. Rather, AAVE was systematic and internally coherent.

Recently, the Yale Grammatical Diversity Project has undertaken a project to document the syntactic diversity of American English. For example, one chapter is on the *needs washed* construction. Some speakers of English accept the sentence *Most babies like cuddled* as grammatical; many others do not. It turns out that those who like this sentence are distributed throughout the Midwest of the US but are concentrated especially heavily in the band from central Pennsylvania to eastern Ohio. For this reason, the *needs washed* construction is sometimes associated with Pittsburgh, but it's actual distribution is much wider.

The Yale Grmamatical Diversity Project relies largely on past surveys of grammatical features, which it amalgamates and synthesizes. Syntactic variation is typically assessed using large surveys, but may be studied more rigorously via corpus methods, open-ended interviews, and so on.

8.2 *Language and Social Identity*

A discussion of how language can vary is empty without an discussion of *why* it varies. The bottom line is that variation within a language is a reflection of social meaning. People talk their identities. They talk like those they identify with; they also talk in ways that reflect their position of power (or lack of power). It will be useful to examine a range of social parameters according to which people can vary and some examples of the linguistic variables that have

been found to vary with them.

8.2.1 Language and Place

The most obvious social parameter according to which languages vary internally is geographical—people tend to talk like those in proximity to them. Some of these variants are not occur below the level of conscious awareness of speakers. They may not even be aware that there are people who speak differently. Other variants may be associated with a particular locations for speakers. For example, speakers of English in Pittsburgh are very conscious that white working class Pittsburghers pronounce *downtown* as /dɑntɑn/.

8.2.2 Language and Class

Class based SOCIOLECTS are an enduring feature of societies with stratified class systems. In the English speaking world, the UK is the most salient example of this, with its well-defined "posh" and working class dialects. However, even in countries that style themselves as more egalitarian, there is significant class-based variation in speech. As in so many things, Labov pioneered this area of study with his famous Fourth Floor study.

Labov started with the hypothesis that there was class variation in the use of post-vocalic /ɹ/ in the English of New York. He decided to test this hypothesis in a department store with several levels (where the more expensive goods were located near the bottom and the more expensive goods were located on the upper floors). The floor on which shoppers were located served as a proxy for their socioeconomic class.

Labov surveyed shoppers on each of the floors by asking them which floor the televisions were on, to which they responded "the fourth floor" (either with or without postvocalic /ɹ/. He then pretended not to hear and asked them again, thus obtaining a careful pronunciation. Based on the data collected from this study, Labov the original hypothesis was supported: shoppers on the upper floors were more likely to pronounce postvocalic /ɹ/ in both words whereas those on the lower floors were more likely to do so. Speakers in the transitional zone produced no postvocalic /ɹ/ in casual speech but did produce it in careful speech.

8.2.3 Language and Formality

The casual/careless distinction relates to a more general distinction in formality. In almost all languages, there are differences between formal and informal speech. In some languages, there are even grammatically encoding distinctions in familiarity (like German or earlier forms of English) or even elaborate systems of pronouns that distinguish a variety of social relationships. Some languages like Thai even have different systems of verbs for formal, versus informal, registers.

8.2.4 Language and Power

It is impossible to talk about familiarity and formality without talking about power. In fact, many social relationships revolve around power asymmetries. These differences in power are often correlated with differences in language. Some of these differences flow naturally from asymmetrical power relationships. For example, in conversational DYADS[2], the lower power participant is more likely to use first person pronouns like *I* and *me*, while the higher power participant is less likely to do so. Other markers of power relationships are more arbitrary and involve titles, terms of address, or pronouns (as discussed in the preceding section). Power asymmetries can characterize racial or ethnic relationship as well as gender relationships. These variables are both associated with linguistic differences.

[2] A dyad is a pair.

8.2.5 Language, Race, and Ethnicity

Ethnic or racial identity, where it is strong, is significantly correlated with linguistic form. Perhaps the most famous example of this, in the sociolinguistic literature, is the case of AAVE (African American Vernacular English). For much of the history of the United States, African Americans spoke a wide variety of regionally-based dialects similar—in many respects, at least—to those of their white neighbors. However, as race, as a source of identity, grew in prominence—and as media and population movements made African Americans more aware of how other African Americans talked—most black Americans started to converge towards a single cluster of varieties of English. This process continues up to the present.

An analogous process occurred among the white residents of the Southern United States. While previously, this region was home to a whole host of dialects, the 1900s saw the emergence of a single dialect that was shared by rural whites through the Southern Region (but not by blacks and others).

8.2.6 Language and Gender

The relationship between language and gender has been studied extensively. It has been found, for instance, that for speakers of English, women have larger vowel spaces (once you control for the length of the vocal tract). Intonational patterns for men and women can also be quite different. Furthermore, women display different patterns in turn taking, politeness, questions, and so on.

There is now research on gay and lesbian speech as well as transgender speech, though these areas have not yet been investigated as well as cisgender male-female differences.

Gender raises an important issue, however: gender is not an intrinsic characteristic of humans (though sex is); gender is something that is performed in culturally defined ways. That being the case, approaches to language and

gender that treat gender as an immutable characteristic are misleading. For a long time, interactionist sociolinguists like Robin Lakoff and Deborah Tannen have criticized the methodological and conceptual approaches of this viewpoint.

8.2.7 Language and Age

Age: (Don't go doing studies on something you don't understand.)

What the hell is age?

Biological age? Physical age (puberty, menopause)? Social, based on life events you have experienced like supporting yourself? Shared experience time like high school?

What age spans are meaningful? 13-17? 18-22?

Male speech features increase with age, so if you use male speech features as a predictor of age, you will predict men to be older than they are.

Youth innovate more. Adults need to be taken seriously.

Features of youth: I, you, niiice, coool, dont, like. But will these be features of old people in 20 years? Need to take language change into account.

8.3 Methods in Variationist Sociolinguistics

8.4 Methods in Social Media Analysis

1. Social media users may not be representative of the general population and demographic information is not easily available.
2. The API doesn't give you all the data, and you have to sample, which can introduce additional bias. Then nobody wants to gather the data again, so they all keep using the same data set for secondary analysis in other studies.
3. New units of analysis: messages and threads.
4. The target audience is not clear. Colleagues and friends are all seeing the same message. Hashtags and at-signs are inaccurate approximations of the intended target audience.
5. Emphasis shifted from phonological variation to lexical variation. But most studies don't take into account that phonology influences orthography.
6. NLP tools don't work well on social media, but normalizing social media data destroys important information for sociolinguistic studies.

8.5 Methods in Interactional Sociolinguistics

Notes from CL article:

The main point is variationism vs interactionalism. In interactionalism, you use language, not as a robot of your demographics, but to effect change on your interlocutor. Such as using code switching to show superiority or solidarity. Most computational studies are variationist, and don't even confirm that they are measuring the right thing.

Variationist: Gender is something you have

Interactionactionalist: Gender is something you do

Tannen and others:

Women: rapport, social Men: report, information exchange However, small sample sizes; ignore other demographic features; don't address similarities between genders

(Scathing) Review of gender prediction studies in NLP

Where do you get the gold answers?

The older studies on Switchboard had demographics that were collected at the time of the recording.

Now, you can get answers from (1) user profiles, (2) human annotation, (3) names, which are not always gender neutral

How do you sample the Data?

Maybe you only use data from people who put their gender in their profile (introduces bias)

Random sampling: very hard to get demographic features

Focused sampling: e.g., people talk about sororities or female hair care products (introduces bias)

What features do you use?

Words: Men use more Prepositions and Determiners; Women use more pronouns, especially "I".

Pennebaker's word lists do not work.

Stylistic features: Men use longer words, sentences, and texts; Women use more emotion words, emoticons, lol, omg, etc.

Gender-specific words: e.g., I am a granny/waitress; My maiden name is ...

Problems with features:

Sometimes features you thought were predictive of gender go away when you control for other things. For example, men write more non-fiction and women write more fiction. Some "female" features are actually fiction features. Also, some "gender" features go away when you consider occupation, so they were really occupation.

Consider the interlocutor: there are more gender-specific features in same-gender conversations in Switchboard. For Tweets, however, there are not more gender-specific features for tweeters with more female followers. But on YouTube the poster does use features to match the gender of their followers.

However, when you try to predict gender on a data set that you did not train on, you do get better-than-random results, so there are some real gender differences.

8.6 Exercise: Linguistic Variation in Social Media

In this exercise you will examine sociolinguistic variation in social media by identifying a social variable with a social-media proxy (like geographic location, gender, sexual orientation, political affiliation, ethnic identity, etc.), identifying a linguistic variable that you believe to relate to this social variable, and produce a statistical analysis testing this hypothesis. Your write-up should detail the three steps:

1. Find an aspect of social identity for which there is a social medial correlate. For example, you could study geographic variation by using geotags on Twitter messages. Likewise, you could study variation based on political affiliation by looking at left-leaning and right-leaning subreddits.

2. Read or listen to samples of the media you are analyzing and try to detect linguistic variation that seems to be correlated with the social variable. You should not use the simple frequency of individual open-class lexical items (e.g. *democracy*, *libs*, or *Trump*) as linguistic variables, since these are likely to be quite trivial. You could use the frequency of other words, however, like *the* or *however*. You can also use frequency or diversity of lexical classes (noun, verb, interjection). Furthermore, you can look at spelling variation. You could also examine the frequency of proper nouns in general, or look for particular syntactic constructions using NLP tools (hard to do when working on social media text, but worth extra credit). Your linguistic variable should be something that you can quantify automatically.

3. Conduct an experiment testing whether your social variable predicts your linguistic variable and test the significance of your experimental results statistically.

Bibliography

Booij, G. (2010). *Construction Morphology*. Oxford: Oxford University Press.

Chao, Y.-R. (1934). The non-uniqueness of phonemic solutions of phonetic systems. *Bulletin of the Institute of History and Philology 4*, 363–397.

Chomsky, N. and M. Halle (1968). *The sound pattern of English*. Studies in language. New York: Harper & Row.

Halle, M. (1962). Phonology in generative grammar. *Word 18*(1–3), 54–72.

Hockett, C. F. (1954). Two models of grammatical description. *Word 10*(2-3), 210–234.

Johnson, C. D. (1972). *Formal Aspects of Phonological Description*. The Hague: Mouton.

Johnson, K. (2011). *Acoustic and Auditory Phonetics* (3rd ed.). Malden, MA; Oxford, UK; and Victoria, Australia: Blackwell Publishing.

Jurafsky, D. and J. H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). Upper Saddle River, New Jersey: Prentice-Hall.

Yao, Y. (2014). Predicting the use of ba construction in mandarin chinese discourse: A modelingstudy with two verbs. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*.