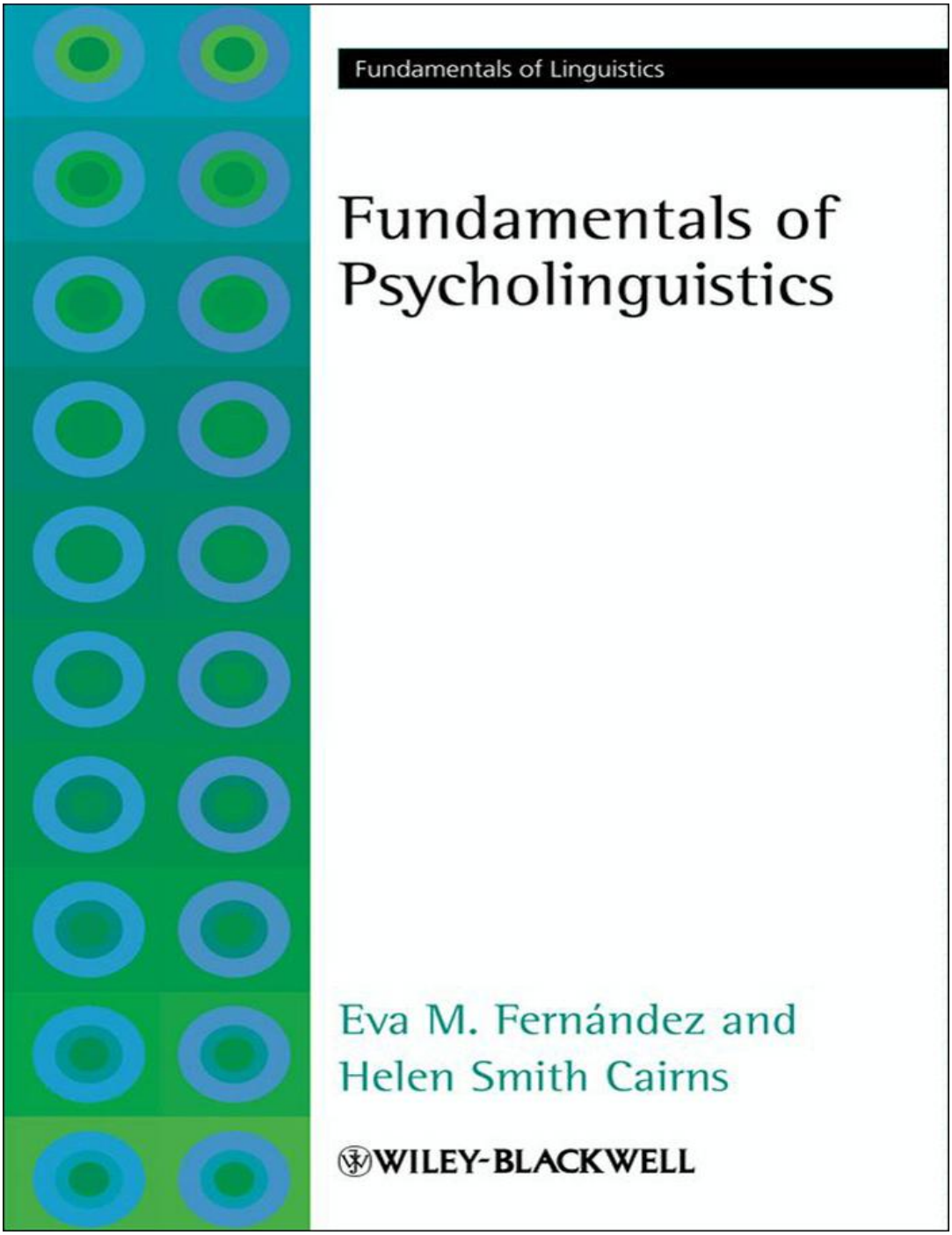

Fundamentals of Linguistics
Fundamentals of Psycholinguistics
Eva M. Fernández and Helen Smith Cairns
WILEY-BLACKWELL

3 The Biological Basis of Language

Psycholinguistics is a field primarily concerned with how language is rep resented and processed in the brain. The focus of this book is, therefore, on language as a system controlled by the brain that is different from but closely linked to general cognition. As such, language is an aspect of human biology. We will explore some of the evidence that psycholinguists – and scholars in related fields – have uncovered linking lang uage to human biology. This chapter will also help you distinguish bet ween language as a biological system and language as a sociocultural artifact.

The organization of this chapter is based on an important historical precedent. Over 40 years ago a neurologist named Eric Lenneberg adduced five general criteria that help determine whether a system is based in the biology of a species (Lenneberg 1964, 1967). These criteria, each described in the sections that follow, are as valid today as they were then, and we will use them to frame the arguments for the biological basis of language.

According to Lenneberg (1967: 371–4), a system is biological if:

- its cognitive function is species specific;
- the specific properties of its cognitive function are replicated in every member of the species;
- the cognitive processes and capacities associated with this system are differentiated spontaneously with maturation;
- certain aspects of behavior and cognitive function for this system emerge only during infancy; and
- certain social phenomena come about by spontaneous adaptation of the behavior of the growing individual to the behavior of other individuals around him.

Later in the chapter, we will describe some of the anatomical and physiological correlates for language. We conclude with a summary of a system closely related to language but decidedly not biological: reading and writing.

As you read this chapter, you might stop to appreciate the great strides that have been made in research focusing on the biological foundations of language since Lenneberg wrote about them in the 1960s. Research in this area has moved at a strikingly rapid rate, even in the past decade or two, facilitated in part by technological advances.

Language Is Species Specific

If we define *communication* loosely as a way to convey messages between individuals, we can generalize that every species has a communication system of some sort. If the system is **species specific** – that is, if it is unique to that species – the system is likely to be part of the genetic makeup of members of the species. Some communication behaviors arise in certain species even if the individual has never heard or seen adults perform the behaviors. Some kinds of crickets and other insects have such a system. Other communication systems, like language for humans and bird song for some species of birds, can be acquired only if the young animal has the opportunity to experience the system in use.

No other species has a communication system like the language used by humans. There are two ways to approach this claim, and thus meet Lenneberg's first criterion. One is rather obvious: no other animals talk, nor do any other animals have a gestural system with the organizational structure of human language. The other way to address this issue is to ask whether other animals can be taught a human communication system.

You have undoubtedly heard of experiments in which researchers have attempted to teach a form of human language to apes. That sort of experimentation is designed to test the claim that human language is species specific: if other species could learn human language, then human language would not be species specific. Primates do not have vocal tracts like those of humans, so the approach has been to teach them communication that involves gestures or manipulated objects.

For example, the famous chimpanzee Washoe was taught to sign words taken from American Sign Language (Gardner and Gardner 1969; Brown 1970). Others, like the chimpanzee Lana (Rumbaugh and Gill 1976) or the bonobo Kanzi (Savage-Rumbaugh and Lewin 1994), have been trained on a variety of computer keyboard systems. Others, like the chimpanzee Sarah, have been taught to manipulate plastic symbols (Premack 1971, 1976). This type of research has been extended beyond primates. Parrots are excellent mimics of the sounds in their environment, and are particularly good at

imitating human speech, even though their vocal tracts are very different from those of humans.

Research in interspecies communication has yielded a tremendous amount of information about the cognitive and social potential of non-human species. Some apes have been able to acquire remarkably large lexicons and use them to communicate about past events, to make simple requests, to demonstrate remarkable abilities of perception and classification, and even to lie. Apes have also demonstrated true symbol-using behavior (e.g., using a red plastic chip to stand for the color green) and the ability to recognize two-dimensional pictures of objects. The grey parrot Alex learned to label many objects, colors, and shapes, and also learned to combine sounds in ways that suggest some degree of awareness of the phonological units that make up speech (Pepperberg 2007).

Importantly, no animal has been able to learn a creative syntactic system. For example, Washoe, the chimpanzee, learned more than a hundred individual words and could combine them communicatively to request food or play. She did not, however, order them in consistent ways to convey meaning, nor was there any evidence that her utterances had any kind of structural organization (Fodor, Bever, and Garrett

1974: 443). Suppose Washoe wanted her trainer (call him Joe) to tickle her. She might sign, *Joe tickle Washoe*, *Washoe Joe tickle*, *Washoe tickle Joe*, *Tickle Joe Washoe*, or any other combination of those three gestures.

The animals that use computers have been trained to press the keys in a particular order, otherwise they do not receive a reward. Lana, a chimpanzee trained this way, would ask for a drink of water by pressing three keys indicating *please*, *machine give*, *water*. Of course, no evidence exists that demonstrates Lana knows the meaning of any of the words associated with the computer keys.
Lana has simply learned that this pattern of behavior will result in a reward of water, whereas other patterns will not. It is not news that smart animals can be trained to produce complex behavioral sequences for reward. However, their use of these sequences does not signify knowledge and use of syntax, particularly the recursive properties of syntax we discussed in Chapter 2. So Lenneberg's basic argument has not yet been falsified. None of these animals has acquired a system that incorporates anything approaching the formal complexity of human language (Hauser, Chomsky, and Fitch 2002).

Even if people had succeeded in teaching animals a communication system incorporating syntax, the claim that human language is biologically based would hardly have been damaged. Human language is certainly the only naturally occurring and naturally acquired system of its type in the animal kingdom. The fact that humans can fly under

very special and artificial circumstances does not challenge the claim that flight is biologically based in birds but not in humans.

Language Is Universal in Humans

Lenneberg's second criterion – that a biological system must be **universal** to all members of the species – is met by language in two ways. First, all human babies are born with a brain that is genetically prepared to organize linguistic information; thus, the psychological processes involved in both acquiring and using language are at play, no matter the person. Secondly, all human languages have universal properties.

There are close to 7,000 languages spoken in the world today and, on the surface, they differ greatly. However, there are profound similarities among the languages of the world – so many similarities, in fact, that *human language* can be thought of as a single entity. Language universals, some of which you read about in Chapter 2, embrace and unify all human languages. These universals do not derive from social, cultural, or general intellectual characteristics of humans.

Instead, they result from the way the human brain organizes and processes linguistic information: language universals are a product of human neurology. Thus, a person's ability to acquire and use language is as natural as a person's ability to walk or a bird's ability to fly.
Thinking of language in this way is similar to the way we think about having hair or walking bipedally, two aspects of being human that are rooted in our biology.

A fundamental goal of linguistics is to describe Universal Grammar, which consists of all the absolute universals of human languages plus a description of their parameters of variation. Universal Grammar represents the "blueprint" or "recipe" for human language that every person is born with. Chapter 2 dealt with language universals and with the type of information supplied by Universal Grammar. Every point made about the organization and functions of the grammar and lexicon is true of all human languages.

All languages have a phonology, a morphology, a syntax, and a lexicon.

All languages possess rules and principles that allow their speakers to combine meaningless phonetic or gestural segments to create meaningful words and sentences.

All languages have an inventory of phonemes, phonotactic constraints on the way words can be formed, and phonological and morphological rules.
Moreover, all languages have a recursive syntax that generates complex sentences, and because of this every human being has the capacity for unlimited linguistic creativity. Finally, all languages have a lexicon, which stores information about words by distinguishing form and meaning. Thus, the general organization of all human languages is the same.

If languages were not biologically based, there would be no necessity for them all to have a similar organization – and we would expect great variation from language to language in terms of their internal organization.

The general organization of language is not the only aspect of linguistic universality. The general properties of grammatical rules are the same for all languages. For instance, in phonology the rules for syllable structure are shared by all languages, although some languages place limitations on syllable structures that other languages do not (as we discussed in Chapter 2, with examples from Spanish and Japanese).

Similarly, in syntax there are restrictions on movement that are universal, and syntactic rules in all languages are structure dependent.

We can turn the concept of universality around and consider impossible languages and impossible rules. No human language could exist in which only simple sentences were used for communication, without the capacity to form complex ones. There are occasional attempts to categorize a language as being primitive. For example, the linguist Daniel Everett has argued along these lines for Pirahã, a language spoken by hunter-gatherers in northwestern Brazil (Everett 2005).

Everett's evidence includes a claim that Pirahã syntax lacks embedding, a charge that the language does not have complex syntax. More careful investigation of the facts about Pirahã syntax has strongly countered Everett's claims: the language does have recursive constructions (Nevins, Pesetsky, and Rodrigues 2009).

It is possible, of course, that at some point our hominid ancestors had a language that consisted only of simple sentences, but that would be speculation, because researchers do not know what the language of protohumans was like (Evans 1998); the lack of fossil evidence about protohuman language is hard but not impossible to overcome, given advances in our understanding of the neural and genetic mechanisms for language (Fitch 2005).

What is certain is that no language spoken by *Homo sapiens* – modern humans –could be so restricted as to not contain recursion. A corollary of this is that there is no such thing as a primitive human language.

The languages spoken in communities of modern-day hunter-gatherers are as rich and complex as the languages of the most industrially and technologically advanced communities, and they all possess human linguistic universals. The same is true of vernacular (non-standard) languages, of languages without writing systems, and of languages that are signed: they are organized in the ways we have described in Chapter 2.

To examine directly whether humans can acquire rules that do not conform to Universal Grammar, a group of researchers attempted to teach a possible and an (impossible) made-up language to a polyglot savant – a person with an extraordinary talent for acquiring languages (Smith, Tsimpli, and Ouhalla 1993).
For this investigation, the extraordinary language learner, Christopher, was exposed to Berber (a language spoken in North Africa, but which Christopher had never learned) and Epun (a language the experimenters invented for the study, containing rules that violated certain aspects of Universal Grammar).
The researchers found that while Christopher learned Berber easily, he found it difficult to learn certain types of rules in Epun, particularly rules that violated structure dependency.

THE BIOLOGICAL BASIS OF LANGUAGE 76&77&78&79&80

■ Language Need Not Be Taught, Nor Can It Be Suppressed

Lenneberg's third criterion is about how biological systems consist of processes that are differentiated (develop) spontaneously as the individual matures. This has two correlates in language acquisition: language does not need to be taught, and acquisition cannot be suppressed. Language acquisition in the child is a naturally unfolding process, much like other biologically based behaviors such as walking.

Every normal human who experiences language in infancy acquires a linguistic system, and failure to do so is evidence for some sort of pathology. Contrary to the belief of many doting parents, language is not taught to children. The fact that children need to hear language in order to acquire it must not be confused with the claim that children need specific instruction to learn to speak. It is probably the case, however, that children need to experience social, interactive language in order to acquire language. A case study involving two brothers, Glen and Jim, who were the hearing children of deaf parents, illustrates both of these points (Sachs, Bard, and Johnson 1981).

The boys were well cared for and did not suffer emotional deprivation, but they had little experience with spoken language other than from watching television. When discovered by the authorities, Jim (18 months old at the time) did not speak, and Glen (3 years, 9 months old) knew and used words, but his morphology and sentence structure were virtually non-existent.

(1) Glen would produce sentences such as the following:

a. That enough two wing.
b. Off my mittens.
c. This not take off plane.

Speech-language pathologists from the University of Connecticut visited the home regularly and had conversations with the children. They did not attempt to teach them any particular language patterns, but they played with them and interacted linguistically with them.

In 6 months, Glen's language was age-appropriate and Jim acquired normal language. When last tested, Glen was a very talkative school-age child who was in the top reading group of his class. The story of Glen and Jim illustrates the importance of interactive input for children during the years they are acquiring language. It also illustrates the fact that specific teaching is not necessary.

In Chapter 4, we will explore further the role of caretakers in the acquisition process.
The fact that language learning cannot be suppressed is yet another manifestation of the biological nature of language.
If language were more bound to the particular types of linguistic experiences a child has, there would be much greater variation in the speed and quality of language learning than is actually observed. In fact, people acquire language at about the same speed during about the same age span, no matter what kind of cultural and social situation they grow up in. Children from impoverished circumstances with indifferent parental care eventually acquire a fully rich human language, just as do pampered children of affluent, achievement-oriented parents.
The biologically driven processes of language acquisition even drive the creation of new languages. Judy Kegl, Ann Senghas, and colleagues (Kegl 1994; Kegl, Senghas, and Coppola 1999; Senghas, Kita, and Özyürek 2004) describe how a signed language has developed in the deaf community of Nicaragua, as the natural product of language learning mechanisms.

In the late 1970s, when schools for educating deaf children in Nicaragua were first opened, the deaf community had no systematic gestural system for communication, other than "home signs" that varied greatly from person to person.
(A home sign is a sign or sign sequence made up by an individual.) Given the opportunity to interact more regularly with each other, deaf children began to develop a gestural system to communicate.
As a result of continued use (both in and out of school), that system eventually expanded into a rudimentary sign language with systematic properties. The language now has over 800 users, and Senghas and colleagues report that the youngest signers are also the most fluent and produce the language in its most developed form.

The process of language birth witnessed in the case of Nicaraguan Sign Language resembles the process through which **pidgins** turn into **creole languages**.
A pidgin is a communication system consisting of elements from more than one language.
A pidgin emerges in situations of language contact, when people who speak different languages come up with ways to communicate with each other. Pidgins have simplified structure and a lexicon consisting of words from the various languages of their speakers. Importantly, a pidgin has no native speakers: its users have learned the communication code as adults, and their ability to use it will be uneven.

When the pidgin becomes *nativized* – that is, when children begin to acquire it as their native language – the grammar stabilizes and becomes more complex, the lexicon grows, and the language is on its way to becoming a creole.

■ Children Everywhere Acquire Language on a Similar Developmental Schedule

There is a remarkable commonality to the milestones of language acquisition, no matter where in the world children acquire language.

Dan Slobin of the University of California at Berkeley has devoted his entire career to the cross-linguistic study of language acquisition and wrote a seminal essay entitled "Children and language: They learn the same all around the world" (Slobin 1972). Like the milestones of motor development (infants roll over, sit up, crawl, and walk at similar ages everywhere), the milestones of language acquisition are also very similar.

Babies coo in the first half of their first year and begin to babble in the second half. The first word comes in the first half of the second year for just about everyone. In all societies, babies go through a one-word stage, followed by a period of early sentences of increasing length; finally, complex sentences begin. By the age of 5 the basic structures of the language are in place, although fine-tuning goes on until late childhood.

Children all over the world are sensitive to the same kinds of language properties, such as word order and inflection. They make remarkably few errors, but their errors are of a similar type. While there is much individual variation in the age at which children acquire aspects of language, that variation is conditioned by individual characteristics of the child rather than by the language being acquired or the culture in which the language is used. One would never expect to hear, for instance, that Spanish-speaking children do not use their first word until they are 3, or that acquisition of Spanish syntax is not completed until adolescence. Nor would one expect to hear that infants in Zimbabwe typically begin speaking at the age of 6 months and are using complex sentences by their first birthday.

There is clearly a developmental sequence to language acquisition that is independent of the language being acquired – although, as we will see in some detail in Chapter 4, some features of language are acquired more easily and earlier than others. In fact, those aspects of language that are easier and those that are more difficult are similar for all children.

All children learn regular patterns better than irregular ones, and they actually impose regularities where they do not exist. For instance, children learning English will regularize irregular past tenses and plurals, producing things like *eated* and *sheeps*.

All children make similar kinds of "errors" – no matter what language they are acquiring. Not only is the sequence of development similar for all children, the process

of acquisition is similar as well. This is exactly what one would expect if the acquisition of a mental system is being developed according to a genetically organized, species specific and species-universal program.

Lenneberg's fourth criterion, claiming that certain aspects of behavior emerge only during infancy, points to an important property of language acquisition: for children everywhere there seems to be a **critical period** in the acquisition of their first language.

Although the details of this concept are controversial, most researchers agree that the optimal period for first language acquisition is before the early teen years, after which a fully complex linguistic system will not develop. The evidence for this comes from reports of so-called "wild children," particularly from the case of Genie, a California girl who was locked in a closet by an abusive father for the first 13 years of her life (Curtiss et al. 1974; Curtiss 1977, 1988). During that time, Genie was deprived of any linguistic input of any kind. After she was rescued, in November 1970, researchers from the University of California at Los Angeles worked for years with her to help her acquire English, but to no avail. She acquired words and the ability to communicate verbally, but she never acquired the full morphological and syntactic system of English.

Examples of her utterances in (2) illustrate the level of her language ability:
(2) a. Genie full stomach.
b. Applesauce buy store.
c. Want Curtiss play piano.
d. Genie have mama have baby grow up.

Genie's hearing was normal, as was her articulation ability. There is some question as to whether her intelligence was completely normal, but even if it was not, this alone could not account for her inability to acquire language. Clearly, Genie had been terribly traumatized for many years, and her emotional development could not have been normal; however, the degree to which she was psychologically impaired could not account for her lack of language.
Actually, Genie was quite friendly and used language well socially. Her problems were solely in morphology and syntax, the formal aspects of language structure that researchers suspect are subject to critical period effects.

Stories like Genie's or those of other "wild children" attempting to learn their first language beyond the early teen years illustrate the claim that after a certain critical period the brain can no longer incorporate the formal properties of language, but they are riddled with hard-to-answer questions related to the unusual life circumstances for these children. Less controversial evidence comes from studies of congenitally deaf adults who learned American Sign Language (ASL) at different ages. Elissa Newport and colleagues (Newport 1990) examined the linguistic competence of users of ASL who acquired the language from birth ("native"), around ages 4–6 ("early learners"), or after age 12 ("late

learners"). The three groups of participants did not differ on tests tapping sensitivity to basic word order, but they differed greatly in tests tapping syntax and morphology.

Native learners outscored early learners, who in turn outscored late learners.
While the existence of a critical period for first language learning is fairly well accepted, its relationship to second language learning is complicated.

Lenneberg himself noted that people's ability to learn a second language in adulthood is puzzling: it is difficult to overcome an accent if you learn a language after early adolescence, yet "a person *can* learn to communicate in a foreign language at the age of forty" (Lenneberg 1967: 176).
Research on age effects in second language acquisition confirms the intuition that older second language learners achieve lower levels of proficiency in their second language.

Existing evidence taps different levels of linguistic competence, including: judgments of degree of foreign accent in speech (Flege, Munro, and MacKay 1995), performance on tests tapping competence in morphology and syntax (Birdsong and Molis 2001), and self-reported oral proficiency (Bialystok and Hakuta 1999).
Studies like this have something important in common (Birdsong 2005): as the learner gets older, the achieved level of competence *gradually* gets lower. Importantly, studies like this suggest that aging makes some aspect or aspects of acquisition harder, but they do not demonstrate that there is a critical period for second language acquisition. Learning a new language later in life will be difficult, but it will not be impossible.

In Chapter 4 we will consider the extent to which an adult learner acquires a second language by processes similar to that of a child acquiring a first language. A related issue, which we will address a little later in this chapter, is whether a second language is represented in the brain in a similar way as the first language. Notice that the very act of posing questions such as these assumes a biological basis for both first and
second language acquisition.

Language Development Is Triggered by the Environment

Lenneberg's final criterion is about the necessity of stimulation from and interaction with the environment. Certain biological systems will not develop without environmental stimuli to trigger them.

Children will not develop language if language is not accessible in their environment or nobody is there to interact with them. Earlier we described the example of how a sign language developed in the deaf community of Nicaragua, in the absence of language in the environment. Yet Nicaraguan signers had an important environmental stimulus: each

other. For a biological system, the environmental input is a stimulus that triggers internal development. We will come back to this in more detail in Chapter 4, when we discuss what characteristics of the language in the environment are necessary for language development.

Anatomical and Physiological Correlates for Language

The most fundamental biological fact about language is that it is stored in the brain, and, more importantly, that language function is localized in particular areas of the brain. This is hardly a new idea, going back at least to Franz Joseph Gall, the eighteenth-century neuroanatomist who developed the field of phrenology.

Gall believed that various abilities, such as wisdom, musical ability, morality, and language, were located in different areas of the brain and could be discovered by feeling bumps on a person's skull.

Gall was, of course, wrong about the bumps, but it seems to be true that some neurally based abilities, such as language, have specific locations in the brain. The first conclusive demonstration that language was localized in the brain took place in 1861 when a French neurologist named Paul Broca presented to the Paris Anthropological Society the first case of **aphasia** (Dingwall 1993).

Aphasia is a language impairment linked to a brain lesion. Broca had a patient who had received a blow to the head with the result that he could not speak beyond uttering *Tan, Tan*, and, thus, Broca called him Tan-Tan. Upon autopsy, he was found to have a lesion in the frontal lobe of the left hemisphere of his brain. Ten years later a German neurologist named Carl Wernicke reported a different kind of aphasia, one characterized by fluent but incomprehensible speech (Dingwall 1993).

Wernicke's patient was found to also have a left hemisphere lesion, farther back in the temporal lobe. **Neurolinguistics** is the study of the representation of language in the brain, and the discovery of aphasias led to the birth of this interdisciplinary field.

The two predominant kinds of aphasia are still called by the names of the men who first described them, as are the areas of the brain associated with each. **Broca's aphasia**, also known as *non-fluent aphasia*, is characterized by halting, effortful speech; it is associated with damage involving Broca's area in the frontal lobe of the left hemisphere.

Wernicke's aphasia, also called *fluent aphasia*, is characterized by fluent meaningless strings; it is caused by damage involving Wernicke's area in the temporal lobe of the left hemisphere. These two kinds of aphasias, among others, differ markedly in terms of the grammatical organization of the patient's speech.

The speech associated with Broca's aphasia has been characterized as **agrammatic**; it consists of primarily content words, lacking syntactic and morphological structure. In contrast, the speech of people with Wernicke's aphasia has stretches of grammatically

organized clauses and phrases, but it tends to be incoherent and meaningless. In conversation, it appears that people with Broca's aphasia comprehend what is said to them, while people with Wernicke's aphasia do not. Thus, a general clinical characterization has been that people with Broca's aphasia have more of a problem with speech production than with auditory comprehension, whereas people with Wernicke's aphasia produce fluent and well-articulated but meaningless speech, and have problems with auditory comprehension.
Psycholinguists studying the comprehension abilities of people with Broca's aphasia discovered something very interesting.

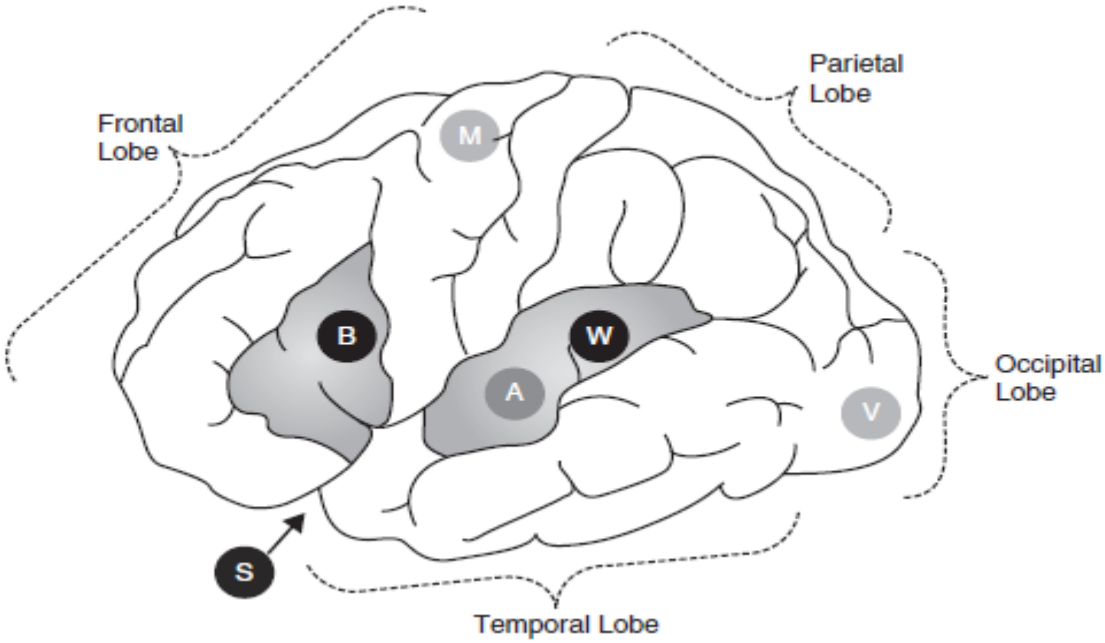
People with Broca's aphasia had no difficulty in understanding sentences like (3a),
but had difficulty with sentences like (3b) (Caramazza and Zurif 1976):

(3) a. The apple the boy is eating is red.
b. The girl the boy is chasing is tall.

Both sentences are constructed of common words; both sentences also have identical structures, including a relative clause modifying the subject noun. There is, however, a profound difference between them: realworld knowledge allows a person to successfully guess the meaning of (3a), but not (3b). Comprehension of (3b) requires an intact syntactic processing system. Caramazza and Zurif's result suggests an explanation as to why people with Broca's aphasia seem to have little trouble with comprehension in conversational contexts. People with aphasia compensate for their impaired grammatical processing system by using realworld knowledge to figure out the meanings of sentences in discourse.

In ordinary conversation with people one knows well and with whom one shares a great deal of real-world knowledge, one can understand much of what is said without having to do a full analysis of sentence structure.

The question remains, of course, as to whether the grammatical problems of people with aphasia are a result of an impaired linguistic competence or are the result of difficulty in using that competence to produce and understand speech. It is very difficult to answer this question experimentally, but some researchers have found people with agrammatic aphasia whose metalinguistic skills with respect to syntax are better than their



Figure

3.1 Diagram of the left hemisphere of the human cerebral cortex (side view). The diagram indicates the location of the primary language areas (Broca's and Wernicke's areas, 'B' and 'W', and the Sylvian fissure 'S'), as well as the approximate areas recruited for motor (M), auditory (A), and visual (V) processing.

ability to produce syntactically complex sentences (Linebarger, Schwartz, and Saffran 1983). This would suggest that the performance system is more impaired than the underlying grammar.

Figure 3.1 provides a sketch of the **left hemisphere** of the cortex of the brain, with Broca's and Wernicke's areas indicated. Broca's area is located near the motor area of the cortex, while Wernicke's is near the auditory area. Importantly, despite the proximity of these areas to motor and auditory areas, aphasias are purely linked to language, and not to motor abilities or audition. Users of signed languages can also become aphasic if they experience damage to the relevant areas in the left hemisphere.

Their signs are non-fluent, halting, and agrammatic. This is true, despite the fact that they have no motor disability in their hands and can use them in everyday tasks with no difficulty (Poizner, Klima, and Bellugi 1987). The fact that signers become aphasic is dramatic confirmation of the fact that signed languages not only have all the formal properties of spoken language, but are similarly represented in the brain. It also demonstrates that the neurological damage that produces aphasia impairs language systems, rather than motor systems.

87& 86 &85 & 84 **THE BIOLOGICAL BASIS OF LANGUAGE**

Aphasia is not a simple or clear-cut disorder. There are many different kinds of aphasia in addition to those classified as fluent and non-fluent, and many different behaviors that characterize the various clinical types of aphasia. Furthermore, much more of the left hemisphere is involved with language than just Broca's and Wernicke's areas; the area all along the Sylvian fissure, deep into the cortex, is associated with language function. Consequently, the localization of the damage for Broca's or Wernicke's patients does not always neatly correspond with the classical description (De Bleser 1988; Willmes and Poeck 1993).

People with aphasia differ greatly in the severity of their symptoms, ranging from mild impairment to a global aphasia where all four language modalities – auditory and reading comprehension, and oral and written expression – are severely impaired.

Language lateralization

To say that language is lateralized means that the language function is located in one of the two hemispheres of the cerebral cortex. For the vast majority of people, language is lateralized in the left hemisphere.

However, in some people language is lateralized in the right hemisphere, and in a small percentage of people language is not lateralized at all, but seems to be represented in both hemispheres.

The hemisphere of localization is related to handedness, left-handed people being more likely than right-handed people to have language lateralized in the right hemisphere. Exactly why this should be the case is unclear, but, as illustrated in Figure 3.2, control of the body is **contralateral**: the right side of the body is controlled by the left motor and sensory areas, while the left side of the body is controlled by the right motor and sensory areas. Thus, left-handed people have rightdominant motor areas, while right-handed people have left-dominantmotor areas.

Many investigations of hemispheric lateralization for language are based on studies of patients about to undergo brain surgery. In these cases, surgeons must be certain where their patients' language functions are localized so these areas can be avoided and an

aphasic outcome prevented. Some procedures used to determine the localization of language in the brain are rather invasive.

One common procedure for determining the hemispheric location of language functions in preoperative patients is the **Wada test**. In this procedure, sodium amytol is injected into one of the two hemispheres of a patient's brain. The patient is asked to count or name pictures presented on an overhead

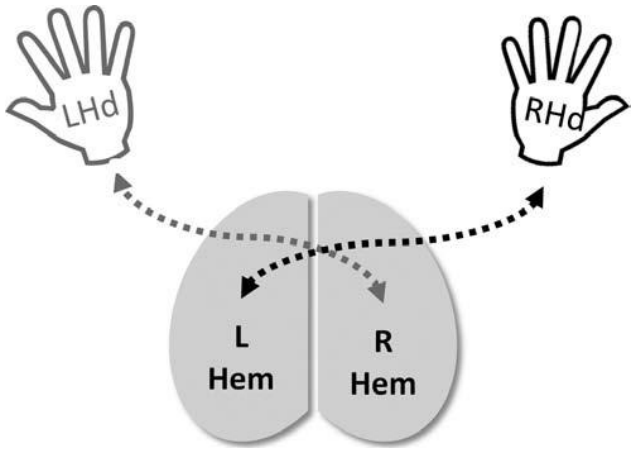


Figure 3.2 Schematic diagram of contralateral control. The shaded lobes represent the two hemispheres of the human brain, looked at from above. The dashed gray lines represent the direct paths from the right hemisphere to the left hand; the dotted black lines, paths from the left hemisphere to the right hand.

screen. Because each hemisphere controls the functioning of the opposite side of the body, the injection produces paralysis on the side of the body opposite from the affected hemisphere. The injection also disrupts verbal behavior, only briefly if the non-dominant hemisphere has been injected, but for several minutes if it has been the dominant hemisphere.

A study of 262 people who were administered the Wada test (Rasmussen and Milner 1977) found that 96 percent of right-handers had language lateralized in the left hemisphere, and only 4 percen in the right. In contrast, only 70 percent of left-handers in the sample were left-lateralized, 15 percent were right lateralized, and 15 percent had language function located in both hemispheres. It is evident that the majority of left-handers are left lateralized, but there is a slightly higher probability that they will have language located in either the right hemisphere or in both.

Another procedure, called **brain mapping**, was originally developed by Penfield and Roberts in the 1950s (Penfield and Roberts 1959), and is still widely used to localize

language function in preoperative patients; it is described extensively by Ojemann (1983). Patients are given a spinal anesthetic so they will be able to communicate with the clinician.

The skull is opened and the brain is exposed, but because the brain itself has no nerve endings, this is not a painful procedure. Various areas are marked along the surface of the brain, and a brief electric current is administered at the same time the patient is performing a verbal task. For example, the patient is shown a picture of a ball and instructed to say, *This is a ball*.
At the moment the word *ball* is about to be produced, a mild electric current is applied to a small area of the exposed brain. If that is a language area, the patient will experience a temporary aphasic-like episode, and will not be able to say *ball*.
If the electric current is applied to a non-language area, there will be no interruption in speech. Surgeons do not cut within 2 centimeters of the areas identified in this manner.

Ojemann (1983) found that his patients had language areas located in Broca's area in the frontal lobe, in Wernicke's area in the temporal lobe, and all around the Sylvian fissure in the left hemisphere, but nowhere in the right hemisphere. Further, there seemed to be some areas that were specialized for word naming and others that were specialized for syntax (although most areas included both abilities).

Ojemann's sample included seven Greek–English bilinguals, for whom there were a few areas in which Greek, but not English, was located, and other areas where English, but not Greek, was located. Importantly, in many areas, both languages overlapped. Ojemann's findings help explain some of the different recovery patterns reported for bilingual aphasics. A brain lesion could affect the two languages of a bilingual in parallel, or differentially (one language will be more affected than the other), or even selectively (one language will not be affected at all).

These and other recovery patterns can be accounted for neuroanatomically (Green 2005): recovery will vary, depending on the area of the brain affected by a lesion.
A particularly fascinating demonstration of the lateralization of language function comes from patients who have had a surgical procedure called *commissurotomy*, in which the two hemispheres of the cortex are separated by cutting the *corpus callosum*, a thick sheaf of nerve fibers joining the two hemispheres. This procedure is performed in cases of severe epilepsy in order to prevent the electrical impulses that cause seizures from surging from one hemisphere to the other. Roger Sperry (1968) received the Nobel Prize for work with people who have had this surgery (Gazzaniga 1970). Bear in mind that the right side of the body is controlled by the left side of the brain, and vice versa. For a person who has had a commissurotomy, the neural pathways controlling the motor and sensory activities of the body are below the area severed by the commissurotomy, so the right motor areas still control the left hand and the right sensory areas receive information from the left side of the body. However, the right hemisphere of the brain cannot transfer

information to the left hemisphere, nor can it receive information from the left hemisphere.
Suppose that a commissurotomy patient has language lateralized in his left hemisphere. If his eyes are closed and a ball is placed in his left hand, he will not be able to say what it is. However, he would be able to select from an array of objects the object that he had held in his hand.
The right hemisphere has knowledge of the identity of the ball, but it lacks the ability to name it. If the ball is placed in his right hand, he is able to name it, just as any person would be able to do with either hand. If a person with an intact corpus callosum were to close her eyes and have someone put a ball in her left hand, the information that it is a ball would register in her right hemisphere, then her right hemisphere would send the information to the left hemisphere, which would name it.
If a person with a split brain is presented with a picture of a spoon in the left visual field (which we come back to below), he will not be able to name it, but he will be able to select a spoon from an array of objects with his left hand. This shows that the right hemisphere recognized the spoon, although it could not name it. The step that is missing for the person who has a split brain is the information transfer from the
right to the left hemisphere.

Obviously, few people have split brains, so psycholinguists have developed a number of experimental techniques for studying the effects of lateralization in intact brains. These include visual field studies, dichotic listening studies, and studies involving neuroimaging.
All demonstrate the language lateralization of the human brain. Visual field studies rest on the fact that it is possible to present information to either the left visual field, which sends information to the right hemisphere, or to the right visual field, which sends information to the left hemisphere.
The left visual field is not the same thing as the left eye; it is a bit more complicated than that. Information in the left visual field comes from both eyes (as does information in the right visual field), but what is of interest here is that the information from the left visual field goes only to the right hemisphere, and information from the right visual field goes only to the left hemisphere.
The fact that visual information can be presented to one or the othe hemisphere has allowed psycholinguists to study in some detail the kinds of linguistic tasks each of the hemispheres can perform. While the right hemisphere is mute, it can recognize simple words, suggesting that there is some sort of lexical representation in the right hemisphere. However, there seems to be no representation of formal aspects of language.
The right hemisphere cannot rhyme, suggesting that it does not have access to the internal phonological structure of lexical items.
Neither does the right hemisphere have access to even simple syntax.

89& 88 **THE BIOLOGICAL BASIS OF LANGUAGE**

In an experiment that tested whether participants could match simple sentences presented to the right hemisphere with pictures they had been shown, participants could not distinguish between the (a) and (b) versions of sentences like the following (Gazzaniga and Hillyard 1971):

(4) a. The boy kisses the girl.
b. The girl kisses the boy.

(5) a. The girl is drinking.
b. The girl will drink.

(6) a. The dog jumps over the fence.
b. The dogs jump over the fence.

Thus, while the right hemisphere may possess some rudimentary lexical information, it is mute and does not represent the phonological, morphological, and syntactic form of language.

Further evidence of the dominance of the left hemisphere for language comes from studies of **dichotic listening**. In this kind of experiment, participants are presented auditory stimuli over headphones, with different inputs to each ear. For instance, the syllable *ba* might be played into the right ear, while at the same exact time *da* is played to the left ear.

The participant's task is to report what was heard. On average, stimuli presented to the right ear are reported with greater accuracy than the stimuli presented to the left ear.

This is known as the **right-ear advantage for language**. It occurs because a linguistic signal presented to the right ear arrives in the left hemisphere for decoding by a more direct route than does a signal presented to the left ear. From the left ear, the signal must travel first to the right hemisphere, then across the corpus callosum to the left hemisphere (Kimura 1961, 1973). Thus, information presented to the right ear is decoded by the left hemisphere earlier than the information presented to the left ear.

The right-ear advantage exists only for linguistic stimuli. Non-speech signals produce no ear advantage, and musical stimuli demonstrate a left-ear advantage (Kimura 1964).

Lateralization apparently begins quite early in life. Evidence suggests that the left hemisphere is larger than the right before birth, and infants are better able to distinguish

speech from non-speech when the stimuli are presented to the left hemisphere (Molfese 1973; Entus 1975).

Early language, however, appears not to be lateralized until the age of about 2. If the left hemisphere is damaged in infancy, the right hemisphere can take over its function. This ability of parts of the young brain to assume functions usually associated with other areas is called *plasticity*.
An infant or young child who suffers left hemisphere damage is far more likely to recover without suffering aphasia than an adult, whose brain is far less plastic.
Even children who have undergone surgery in which the left hemisphere is removed can develop quite good language functions. However, studies have shown that such children are deficient in the formal aspects of language morphology and syntax.

Thus, the right hemisphere may be limited in its plasticity in that it cannot incorporate the structural analytical aspects of language associated with the left hemisphere (Dennis and Whitaker 1976).

Neuroanatomical correlates of language processing

Our understanding of how the brain represents and processes language has broadened dramatically with the development of neuroimaging techniques like event-related potentials and functional magnetic resonance imaging. Neuroimaging research focuses on identifying neuroanatomical correlates for the competence repositories and performance mechanisms for language.

While the brain is at work, active neurons emit electrical activity.
This voltage can be measured by attaching electrodes to the scalp at different locations; the technical term for this is electroencephalography (EEG).
Event-related potentials (or **ERPs**, for short) are changes in the electrical patterns of the brain that are associated with the processing of various kinds of linguistic stimuli. In ERP experiments, sentences are presented either visually (one word at a time) or auditorily, while measurements are collected that provide information about the timing, the direction (positive or negative), and the amplitude of the voltage.

The brain has different electrical responses to different types of linguistic anomalies. This is strong support for the proposition that different brain mechanisms are employed in processing semantic-pragmatic information on the one hand and morphosyntactic information on the other.

One of the best-known ERP effects is the **N400 component**, so called because its signature is a negative (N) voltage peak at about 400 milliseconds following a particular stimulus.

This component is sensitive to semantic anomalies, such as the ones in (7a) and (7b), compared to (7c) (Kutas and Van Petten 1988):

(7) a. *The pizza was too hot to cry.
b. *The pizza was too hot to drink.
c. The pizza was too hot to eat.

Studies investigating morphological and syntactic anomalies have discovered ERP components associated with structural processing (Friederici 2002). Morphosyntactic errors, like subject–verb agreement violations, elicit a **left anterior negativity (LAN)**, which occurs between 300 and 500 milliseconds. Another ERP component linked to syntactic structure building is a very **early left anterior negativity (ELAN)**.

At around 150–200 milliseconds, the ELAN is even earlier than the LAN, and is characterized by electrical activity that is more negative when building syntactic structure is not possible, as in (8a), compared to (8b) (Neville et al. 1991):

(8) a. *Max's of proof
b. Max's proof

A late centro-parietal positivity, the **P600 component** (for positive voltage between 600 and 1000 milliseconds, also called the **Syntactic Positive Shift**, or **SPS**), is elicited with syntactic violations (Osterhout and Holcomb 1992), with sentences that require reanalysis (we will come back to these in Chapter 7), and with sentences that are syntactically complex (Friederici 2002).

Figure 3.3 summarizes some of the results from a study by Osterhout and Nicol (1999), which compared ERP responses to grammatical sentences and sentences with semantic anomalies (top panel), syntactic anomalies (middle panel), or both semantic and syntactic anomalies (bottom panel). The semantic anomalies elicited N400 effects, while the syntactic anomalies elicited P600 effects.

There are ERP components that have been associated with other aspects of language processing. For example, the **Closure Positive Shift (CPS)** is an ERP component linked to the processing of prosodic phrasing: intonational boundaries inside sentences elicit positivity (Steinhauer, Alter, and Friederici 1999). A different ERP component, the **P800**, is elicited when the intonation of a sentence does not match its form, for example, when a question has the intonation of a declarative, or vice versa (Artésano, Besson, and Alter 2004).

Studies using **functional magnetic resonance imaging (fMRI)** and **positron emission tomography (PET)** provide detailed information about the areas of the brain implicated

in language processing. These technologies measure blood flow levels, capitalizing on the fact that increased neuronal activity in a particular area of the brain is supported by increased blood flow. fMRI data provide topographical information about what regions of the brain are specialized for different aspects of language representation and processing tasks. In addition to Broca's and Wernicke's areas,

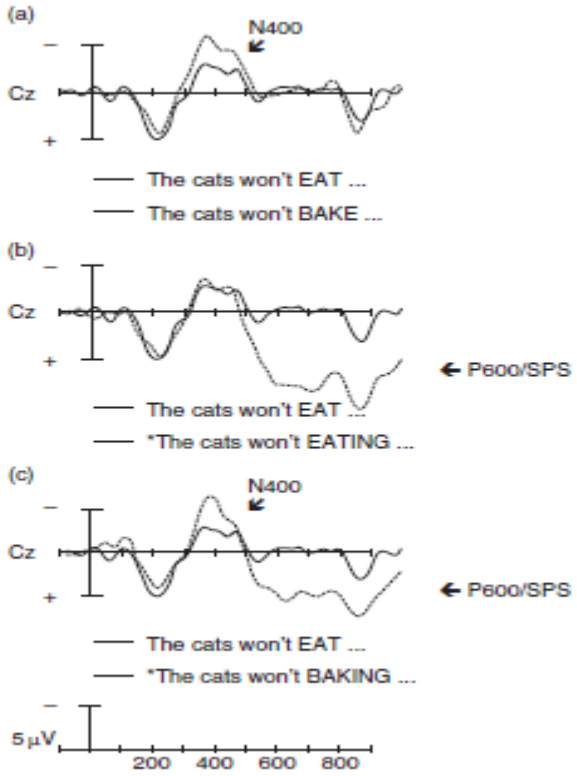


Figure 3.3 ERPs recorded at one electrode (labeled *Cz*), for contrasts between grammatical sentences (solid lines) and anomalous sentences (dotted lines).

The anomaly was semantic (a)
, syntactic (b)
, or both semantic and syntactic(c).

The graphs indicate voltage on the *y*-axis (negativity is up, positivity down), and time (in milliseconds) on the *x*-axis. Sentences with semantic anomalies elicited greater negativity at around 400 milliseconds (the signature of the N400 effect), while sentences with syntactic anomalies elicited greater positivity around 600 milliseconds (characteristic of the P600 effect). The figure is a composite of data reported by Osterhout and Nicol (1999). We thank Lee Osterhout for his permission to reproduce this figure.

other areas of the brain have been found to be involved in processing language, with specific neuroanatomical correlates for different types of processing (Bornkessel-Schlesewsky and Friederici 2007).
ERPs are useful to study the time course of processing, while fMRI is better at detecting the areas of the brain that are involved in processing tasks.

The search for a genetic basis for language

The ultimate indicator of the biological nature of language would be the discovery of the genetic basis of language, as all aspects of human biology are directly encoded in our DNA.
Researchers began genetic investigations by conducting pedigree studies. These are studies that examine the heritability of a particular trait (or disorder) in several generations of a family. Gopnik (1990, 1997) showed that members of over three generations of one family had suffered from specific language impairment (SLI), dyslexia, and other language disorders, indicating that genetic anomalies associated with language development can be inherited.

A major breakthrough came with the discovery by Lai and colleagues (Lai et al. 2001) of a specific gene, **FOXP2**, that was implicated in the language disorders of an extended family.

Members of the family exhibited symptoms like those of agrammatic aphasics:
effortful and non-fluent speech,
lacking in syntactic organization.
Their grammar appeared to be broadly impaired; they had difficulty manipulating phonemes and morphemes and understanding complex sentences (Watkins, Dronkers, and Vargha-Khadem 2002).
The disorder was attributable to a mutation of the FOXP2 gene, which was transmitted by heredity. If a mutated version of a gene is responsible for language disorders, it is reasonable to infer that an intact version of that gene is implicated in normal language development and representation. It was suggested that a "gene for language" had been discovered.

The FOXP2 gene is associated with the development of other parts of human anatomy unrelated to language, including the lung, the gut, and the heart. It is also a gene that is not confined to *Homo sapiens*; it is also found in other mammals, including mice (Marcus and Fisher 2003).
While the relationship of FOXP2 to heritable language disorders is an exciting breakthrough, it is important to remember that it cannot be *the* gene for language. All complex behaviors are attributable to the interaction of many genes and their schedules

of expression. So FOXP2 is almost certainly only one gene in a network of multiple genes involved in the language abilities of humans.

5 The Speaker: Producing Speech

The processes that underlie the production and comprehension of speech are information processing activities. The speaker's job is to encode an idea into an utterance. The utterance carries information the hearer will use to decode the speech signal, by building the linguistic representations that will lead to recovering the intended message.

Encoding and decoding are essentially mirror images of one another.
The speaker, on the one hand, knows what she intends to say; her task is to formulate the message into a set of words with a structural organization appropriate to convey that meaning, then to transform the structured message into intelligible speech. The hearer, on the other hand, must reconstruct the intended meaning from the speech produced by the speaker, starting with the information available in the signal.

In this and the next three chapters, we will describe the information processing operations performed rapidly and unconsciously by the speaker and the hearer, as well as the mental representations constructed by those operations. It is worth emphasizing that a hearer's successful recovery of a speaker's intention when uttering a sentence involves shared knowledge that goes well beyond knowledge of language and well beyond the basic

meaning of a sentence – a topic we will explore in Chapter 8. But before we can examine contextualized language use, we describe the operations that use knowledge of language in encoding and decoding linguistic signals.
This chapter focuses on production.

Since the mid-1970s, production has gradually become a central concern in the study of language performance (Bock 1991), alongside the study of perception. The sections that follow provide an introduction to some of that research. We will first discuss the components of a general model for language production. We will then describe the mental mechanisms that constrain how speakers encode ideas into mental representations of sentences, which are eventually uttered, written, or signed.
The chapter concludes with details on how those mental representations are transformed into an acoustic speech signal.

A Model for Language Production

The production of a sentence begins with the speaker's intention to communicate an idea or some item of information. This has been referred to by Levelt (1989) as a **preverbal message**, because at this point the idea has not yet been cast into a linguistic form. Turning an idea into a linguistic representation involves mental operations that require consulting both the lexicon and the grammar shared by the speaker and hearer.

Eventually, the mental representation must be transformed into a speech signal that will be produced fluently, at an appropriate rate, with a suitable prosody. There are a number of steps to this process, each associated with a distinct type of linguistic analysis and each carrying its own particular type of information. Figure 5.1 summarizes, from left to right, the processing operations performed by the speaker.

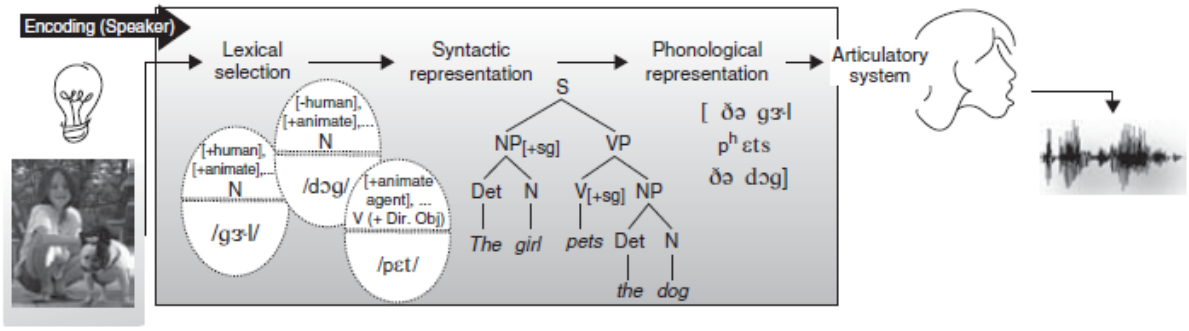


Figure 5.1 Diagram of some processing operations, ordered left to right, performed by the speaker when producing the sentence *The girl pets the dog*.

(This figure expands on parts of Figure 1.3, Chapter 1.) Production begins with an idea for a message (the light bulb on the far left) triggering a process of lexical selection. The capsule-like figures represent lexical items for the words *girl*, *dog*, and *pet*, activated based on the intended meaning for the message; these include basic lexical semantic and morphosyntactic information (top half) and phonological form information (bottom half).

The tree diagram in the center represents the sentence's syntactic form. The phonetic transcription to the right represents the sentence's eventual phonological form, sent on to the articulatory system, which produces the corresponding speech signal.
The different representations are accessed and built very rapidly and with some degree of overlap.
The first step is to create a representation of a sentence meaning that will convey the speaker's intended message. This semantic representation triggers a lexical search for the words that can convey this meaning.

(In Figure 5.1, the words *girl*, *dog*, and *pet* are activated.) The meaning of a sentence is a function of both its words and their structural organization (*The girl pets the dog* does not mean the same thing as *The dog pets the girl*), so another encoding stage involves assigning syntactic structure to the words retrieved from the lexicon. This process places the words into hierarchically organized constituents.

Morphosyntactic rules add morphemes to satisfy grammatical requirements – for example, the requirement that a verb and its subject must agree in number.

A phonological representation can then be created, "spelling out" the words as phonemes. Phonological and morphophonological rules then apply to produce a final string of phonological elements.

This phonological representation will specify the way the sentence is to be uttered, including its prosodic characteristics.

The final representation incorporates all the phonetic detail necessary for the actual production of the sentence.

In this representation phonological segments are arranged in a linear sequence, one after the other, as if they were waiting in the wings of a theater preparing to enter the stage.

This representation is translated into instructions to the vocal apparatus from the motor control areas of the brain, and then neural signals are sent out to the muscles of the lips, tongue, larynx, mandible, and respiratory system to produce the actual speech signal.

Recall that Chapter 1 drew a number of important distinctions between language and speech, between language and other aspects of cognition (like general intelligence), and between language and communication.

Language is a distinct, autonomous system that can be described without reference to other systems (as we did in Chapter 2).

Yet the interaction of linguistic and non-linguistic systems, a recurring theme in this book, is key to understanding psycholinguistic processes.

Psycholinguists disagree on some details regarding the nature and the degree of interaction between linguistic and non-linguistic systems, but that they do interact is uncontroversial. Sentence planning offers an excellent example of this phenomenon: the processes inside the gray box of Figure 5.1 use the speaker's knowledge of language to link ideas to signals, two non-linguistic and quite distinct representations.

An idea is a product of the speaker's general cognition and intellect. Speech is a complex motor activity engaging the vocal tract and respiratory physiology. The elements outside the gray box – ideas, articulatory processes, and acoustic signals – are not part of language and do not have abstract linguistic representations (though they certainly have abstract non-linguistic properties).
All of the representations inside the gray box of Figure 5.1 are abstract linguistic representations. Linguistic theory provides a vocabulary and a framework to represent syntactic structure, morphemes, and phonological segments.

The general model presented in Figure 5.1 can account for two aspects of language production not covered in this chapter: writing and signing.

If a sentence is to be written rather than uttered, its phonological representation will be sent to the motor system responsible for engaging the hands either in handwriting or typing. Very little is known from a psycholinguistic perspective about the writing process, though occasionally psycholinguists employ writing as the medium for eliciting production. For signed languages, the phonological representation of a sentence will be very different than for spoken languages, and the articulation of that representation will be handled by a motor system that engages the hands and face to create gestures. Other than those differences, writing and signing involve the same stages of sentence planning as we will describe in detail below about speaking.

The model in Figure 5.1 – a much simplified version of models like Levelt's (1989) or Garrett's (1988), among others – has been refined empirically over the years.
A great deal of what is known about the

levels of production planning comes from analyses of **speech errors** (also called **slips of the tongue**) by Garrett (1980a, 1980b, 1988), Fromkin (1971, 1980, 1988), and others. This research draws on speech error corpora, collected by the investigators, by noting the occasions when they or their interlocutor produced a speech error. An **interlocutor** is a participant in a conversation. Other evidence comes from studies using a range of techniques to elicit speech production under controlled laboratory conditions; the objective of such work is to examine how fluent speech is produced, and what conditions cause fluent speech to break down.

Production in bilinguals and second language learners

Few adjustments need to be made to the working model in Figure 5.1 to account for production by people who speak two or more languages.

We need to assume that a bilingual has two language-specific grammars, and a lexicon with language-specific entries, and we need to specify how these language-specific knowledge repositories are activated (or deactivated) – but that is all. When a bilingual is speaking in a **unilingual mode** (only one language), only one of the grammars is consulted to build structural representations, and the active language's lexical entries are activated. When in a **bilingual mode** (when the bilingual's two languages are being used in the same conversation), access to both grammars and lexical items from both languages
must be possible (Grosjean 2001). Models of bilingual language production, like de Bot's (2004) or Green's (1986), incorporate mechanisms to control activation of the language or languages of the conversation (or inhibition of the language or languages not being used).

Choosing what language (or languages) to activate during a conversation is guided by the speaker's communicative intent and other nonlinguistic variables like conversation participants, topic, and context.

(For more discussion of language choice, see Chapter 8.) The process resembles how a monolingual chooses between speaking formally or informally.

Importantly, the steps for production continue to be the same in both the unilingual and the bilingual mode of production, and for monolingual and bilingual speakers:
lexical items are selected;
a syntactic structure is built;
a phonological representation is generated.

However, knowledge of two languages has at least two important consequences for language production: it permits intentional switching from one language to the other, and it triggers occasional unintentional slips into a language not active in the conversation.

One type of alternation between languages in bilingual speech is **code-switching**. Code-switching is switching between two *codes* (two languages, or two distinct dialects of the same language) within the same discourse. A switch can take place between sentences (*intersentential code-switching*). A switch can also occur within the same sentence (*intrasentential code-switching*), at clause boundaries, or at smaller phrasal boundaries. A third category, *tag-switching*, involves the insertion of frequently used discourse markers, like *so*, *you know*, *I mean*, etc. (Lipski 2005). The example in (1), produced by a Spanish–English bilingual (cited by Romaine (1995: 164)), illustrates all three types of code-switching; the underlined phrases are translated below the example:

(1) … they tell me 'How'd you quit, Mary?' I don't quit I … I just stopped. I mean it wasn't an effort that I made
a- que voy a dejar de fumar por que me hace daño o this or that uh-uh. It's just that I used to pull butts out of the waste paper basket yeah. I used to go look in the …
b- se me acaban los cigarros en la noche. I'd get desperate
c- y ahí voy al basurero a buscar, a sacar, you know.

a 'that I'm going to quit smoking because it's harmful to me or'
b 'I run out of cigarettes at night'
c 'and so I go to the trash to look for, to get some out'

Code-switching is a discourse style that is most typical in bilinguals who are highly proficient speakers of both languages (Poplack 1980), which is not surprising: producing utterances that alternate between two languages requires sustained activation of the grammars and lexicons of each language, and of the rules that govern grammatical switching.

Code-switching generally serves a communicative function (Myers- Scotton 1988). A bilingual may switch to the other language to emphasize something just said, to quote something or someone, or to modify a statement further; code-switching can also be used to include or exclude an interlocutor, or to signal power relations between interlocutors. In the example in (1), the speaker switches into Spanish for the more personal parts of her message. In some bilingual speech communities, the default communication style when in a bilingual mode involves frequent alternation between two languages (Myers-Scotton 1988).

Code-switching is guided by the same production mechanisms involved in unilingual production. Research examining large code-switching corpora has demonstrated that

naturally occurring code-switching is highly principled behavior (Myers-Scotton 1993). As such, code-switching offers insights about the cognitive architecture that supports bilingualism.

140 THE SPEAKER: PRODUCING SPEECH

In Chapter 2 we used the phenomenon of **borrowing**, in which a word from one language is incorporated into the lexicon of another, to illustrate how a borrowed word might be transformed to conform with the phonotactic constraints of the incorporating language. Borrowing is also a feature of bilingual language use, and it is sometimes difficult to distinguish from code-switching. One difference between the two is the degree of integration of the guest word in the host language. A borrowed word (also called a *loan*) typically undergoes both orthographic and phonological adaptation into the host language; the example in (2a) illustrates orthographic adaptation (the loan in English is not capitalized and loses the umlaut over the third vowel). Loans are sometimes translated into an equivalent word in the host language, and are then called *loan translations* or *calques*; an example is in (2b). Bilinguals often borrow to fill lexical gaps in one of their languages. Loanwords sometimes become established in the language, and even monolinguals will begin to use them.

(2) a. doppelganger
'ghostly counterpart of a person'
(from German *Doppelgänger*)

b . thought experiment
(from German *Gedankenexperiment*)

It is important to distinguish between deliberate alternations, like code-switching or borrowing, and unintentional non-native-like elements in the speech of a second language learner. A second language grammar may differ – slightly or dramatically – from the grammar of a native monolingual speaker. No doubt, you have heard second language learners speak with an accent, use words in ways that do not match native speakers' intuitions, and even produce sentences with unusual syntax. Non-native-like production by second language learners can be the result of rules from the first language being incorporated into the second language, a phenomenon called **transfer**. Non-native-like production can also be linked to the use of acquisition strategies like overgeneralization (see Chapter 4).

■ Planning Speech Before It Is Produced

Producing a sentence involves a series of distinct operations and representations:

lexical, syntactic, morphological, and phonological. The following sections discuss some of the evidence that has led researchers to posit these different levels of production planning.

Accessing the lexicon

As mentioned above, the process of language production begins with an idea that is encoded into a semantic representation. This sets in motion a process called **lexical retrieval**. Remember that the lexicon is a dictionary of all the words a speaker knows. A lexical entry carries information about the meaning of the word, its grammatical class, the syntactic structures into which it can enter, and the sounds it contains (its phonemic representation). A word can be retrieved using two different kinds of information: meaning or sound. The speaker retrieves words based on the meaning to be communicated and has the task of selecting a word that will be appropriate for the desired message. The word must also be of the appropriate grammatical class (noun, verb, etc.) and must be compatible with the structure that is being constructed. It is most certainly not the case that the structure is constructed before the words are selected, nor are all the words selected before the structure is constructed. In fact, the words and the structure are so closely related that the two processes take place practically simultaneously. Ultimately, the speaker must retrieve a lexical item that will convey the correct meaning and fit the intended structure.

This means that a speaker must enter the lexicon via information about meaning, grammatical class, and structure, only later to retrieve the phonological form of the required word. The hearer's task, which will be discussed in detail in the next chapter, is the mirror image of the speaker's. The hearer must process information about the sound of the word and enter his lexicon to discover its form class, structural requirements, and meaning.

Important psycholinguistic questions concern the organization of the lexicon and how it is accessed for both production and comprehension.

The speed of conversational speech varies by many factors, including age (younger people speak faster than older people), sex (men speak faster than women), nativeness (native speakers are faster than second language speakers), topic (familiar topics are talked about faster than unfamiliar ones), and utterance length (longer utterances have shorter segment durations than shorter ones); on average, though, people produce 100 to 300 words per minute (Yuan, Liberman, and Cieri 2006), which, at the slower end, is between 1 and 5 words (or 10 to 15 phonetic elements) per second. (Notice that this includes the time it takes to build syntactic and phonological representations and to move the articulators, not just time actually spent in lexical retrieval.) Clearly, the process of accessing words is extremely rapid.

According to Miller and Gildea (1987), adults with a high school education know around 40,000 words. All the different versions of a single word count as one word. For example, *write*, *writer*, *writes*, *written*, and *writing* together count as one word. If one adds to that total another 40,000 proper names of people and places, the adult lexicon is estimated to contain around 80,000 words. If each word a person uses must be retrieved from a bank of 80,000 in less than half a second, it is obvious that the processes employed in lexical retrieval must be extremely efficient, and these processes are affected by the way the lexicon is organized.

One way the lexicon is organized is by frequency of use, a topic we will explore in more detail in Chapter 6. During production, more common words are retrieved more rapidly: for example, it is easier and faster to retrieve the word *knife* than the word *dagger*. Studies of pauses and hesitations in speech have shown that hesitations often occur before low-frequency words (Levelt 1983).

Words are also organized by their meaning, so close associates are stored near one another. Speech errors can give some insight into this meaning-based organization. It is extremely common for a word retrieval error to result in the selection of a semantically and structurally similar word. Consider the following examples:

(3) a. I just feel like whipped cream and mushrooms.
{I just feel like whipped cream and strawberries.}

b. All I want is something for my elbows.
{All I want is something for my shoulders.}

c. Put the oven on at a very low speed.
{Put the oven on at a very low temperature.}

d. I hate … I mean, I *love* dancing with you!

(In all examples of speech errors in this chapter, the intended utterance is in curly brackets, beneath the actual utterance containing the slip.) In each of the examples in (3), the speaker has erroneously selected a word that is of the same grammatical class (nouns) and that shares many aspects of meaning with the intended word. This kind of error is very common and is probably responsible for many of the so-called *Freudian slips* that people make, such as the one in (3d). However, rather than representing a repressed desire for mushrooms or a secret loathing of one's dance partner, the errors in (3) are more likely driven by the fact that words sharing semantic features are stored together.

Antonyms – words that are the opposite of one another, like *love* and *hate* – actually

share a great many aspects of meaning. *Love* and *hate* are both verbs that refer to internalized feelings one person can have about another; the only difference between them is that they refer to distinct (and opposite) feelings. Speech errors often involve the production of *forget* instead of *remember*, *give* instead of *take*, and so on.
Sometimes words that sound alike are implicated in speech errors, like the following:

(4) a. If you can find a gargle around the house …
{If you can find a garlic around the house …}

b. We need a few laughs to break up the mahogany.
{We need a few laughs to break up the monotony.}

c. Passengers needing special assistance, please remain comfortably seated until all passengers have complained … uh, deplaned.

In these errors, the grammatical class of the intended and the intruding word is the same, even though the meaning is completely different.
Errors like these suggest that words are organized by phonological structure, forming "neighborhoods" of words that sound similar.
Semantically based and phonologically based errors, like those in (3) and (4), respectively, provide evidence for the distinction between two components of lexical representations discussed in Chapter 2: meaningbased and form-based.

A phenomenon in lexical retrieval that has fascinated psycholinguists for decades is the **tip-of-the-tongue phenomenon** (Brown and McNeill 1966; Aitchison 2003).

A tip-of-the-tongue state occurs when the speaker knows the word needed but cannot quite retrieve it. It is a very uncomfortable mental state, and when people experience it, they might say "I've got that word right on the tip of my tongue!" What people experience during a tip-of-the-tongue state offers a glimpse into the steps involved in lexical retrieval. Typically, people have access to the meaning- based part of the lexical representation, but experience a tip-of-thetongue state when they fail to find a fully specified form-based representation (Bock and Levelt 1994).

However, people typically know something about the word they are unsuccessfully searching for. They can often think of the initial or final sounds or letters, how many syllables it has, where primary stress is located, and even words that sound similar. People experiencing a tip-of-the-tongue state will often also perform gestures that are suggestive of the meaning of the word, though it is not necessarily the case that gesturing helps retrieval (Beattie and Coughlan 1999).

While no one really understands tip-of-the-tongue states, it is a phenomenon that demonstrates that when people enter the lexicon through meaning, in order to produce a word, a great deal of information may be available even if the entire representation of the word is not retrieved. Tipof- the-tongue states, of course, are a rare occurrence, as are lexical retrieval errors like the ones in (3) and (4). Usually lexical retrieval produces an appropriate set of words required for the speaker's sentence.

Building simple sentence structure

Levelt (1989) refers to the creation of sentence structure during sentence planning as **grammatical encoding**. For this the speaker must consult the internalized grammar to construct structures that will convey the intended meaning. Again, speech errors provide information about some of the characteristics of the representations that are constructed.

We know, for instance, that words are represented as separate units.
Speech errors like the ones in (5) provide evidence for this:

(5) a. I left the briefcase in my cigar.
{I left the cigar in my briefcase.}

b. … rubber pipe and lead hose …
{… rubber hose and lead pipe …}

These examples illustrate a common type of error, **exchange errors**; the exchange units here are two words. **Word exchange errors** never occur between content words and function words and are usually limited to words of the same grammatical class, nouns in the case of the examples above.

An exchange error can involve units larger than individual words.
Such errors provide evidence that sentences are organized structurally during language production. Constituents that are larger than words, but which are units in the hierarchical organization of the sentence, can exchange with one another. Consider the following error:

(6) The Grand Canyon went to my sister.
{My sister went to the Grand Canyon.}

A noun phrase, *the Grand Canyon*, has changed places with another noun phrase, *my sister*. Thus, a constituent larger than an individual word has moved. Movement of two words that are not part of the same constituent is never observed. An error such as **The grand my sister to canyon went* would never be produced. In speech errors, syntactically defined constituents are moved, and the resulting sentences are always structurally well-formed sentences of English.

Exchange errors also demonstrate the existence of a level of representation where bound morphemes are represented separately from their stems, as the following examples illustrate:

(7) a. He had a lot of guns in that bullet.
{He had a lot of bullets in that gun.}

b. You ordered up ending.
{You ended up ordering.}

c. We roasted a cook.
{We cooked a roast.}

d. … gownless evening straps …
{… strapless evening gowns …}

In (7a), *gun* and *bullet* have been exchanged, but the plural morpheme *–s* appears in the intended structural position. In (7b), the words *end* and *order* have been exchanged, but the morphemes *–ed* and *–ing* appear in their intended structural positions. The same type of analysis applies to (7c), in which *roast* and *cook* have been exchanged, but the morpheme *–ed* has not moved. These examples suggest that while speech errors may produce sentences with odd meanings, they rarely produce structurally bizarre sentences. The error in (7b), for instance, was not **You ordering up ended*, as it would have been if the bound morphemes and the stem had formed a unit at the time of exchange.

How are errors like those in (7) possible? Free morphemes, and the bound morphemes that attach to them, are separate units in the mental representations built during sentence production. Inflectional morphemes, like *–s*, *–ed*, and *–ing*, are added to specific structural positions, based on the syntax of the sentence, rather than based on the words they eventually attach to. The error in (7d) suggests that much the same applies to derivational morphemes, like *–less*. There is a level of representation at which free and bound morphemes are represented separately.

Errors like those in (7) also suggest that morphemes are added to the mental representation before morphophonological rules operate to specify the phonetic form by which the morpheme will be realized.

The example in (7c) is particularly relevant. (Notice that (7c) might initially appear to contradict the observation that only words of the same grammatical class are exchangeable, since *cook* is a verb and *roast* is a

147-146 THE SPEAKER: PRODUCING SPEECH

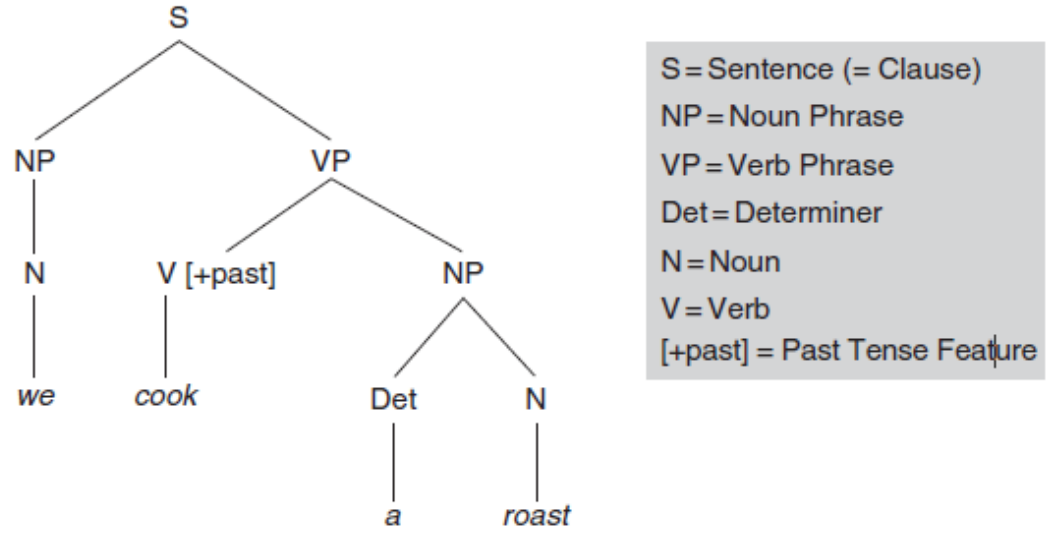


Figure 5.2 Representation of a past tense morpheme before the application of morphophonological rules.

noun. However, both words can be either a noun or a verb, so the example is not a contradiction.) The past tense morpheme *–ed* differs in the way it is pronounced depending upon the final segment of the verb to which it is attached. The past tense morpheme on *cook* surfaces as [t], while on *roast* it surfaces as [id]. In the speech error in (7c), the past tense feature is "spelled out" according to morphophonological rules attaching it to *roast*. Clearly, *roast* and *cook* were exchanged before morphophonological rules applied. The exchange error resulting in (7c) thus provides evidence for a level of representation as shown in Figure 5.2, where past tense is an abstract feature in the syntactic structure, but the morpheme that marks past has not yet been added to the word *cook*.

The words were exchanged at a processing level before morphophonological rules had applied. If the exchange error had occurred at a later processing stage, the sentence would have been uttered as **We roast a cooked*. Such a speech error would never occur.

The following speech error illustrates a similar interplay of morphology and phonology:

(8) If you give the nipple an infant …
{If you give the infant a nipple …}

In this example, *nipple* and *infant* have been exchanged before the morphophonological

rule specifying the pronunciation of the indefinite determiner has applied. The determiner would have been pronounced *a* before *nipple*, but instead became *an*, given the initial segment of *infant*. Had the exchange error occurred after the application of the morphophonological rule, the resulting sentence would have been **If you give the nipple a infant.*

Creating agreement relations

The errors we have described so far illustrate aspects of sentence planning related to placing lexical material in structural positions in a syntactic representation. There is another class of errors, which has been studied extensively in English and several other languages, involving subject–verb agreement. These errors are informative about the role of agreement features in production planning and execution.

Agreement is a requirement of the grammar, with some very language- specific properties. English requires that verbs and their subjects agree in number (and person). Since English has limited morphology, number agreement is only marked (by a bound morpheme) on verbs with third person singular subjects, like (9a), or on subjects when they are plural, like (9b):

(9) a. The bridge closes at seven.
b. The bridges close at seven.

Other languages have richer morphology for agreement, and require not only agreement of number and person features, but also of gender features. (Examples of some of these are in Chapter 2.)

Many languages require agreement between verbs and their subjects, and some languages also require agreement between verbs and their objects.

For an English speaker, producing sentences with grammatical number agreement is relatively straightforward, with one important exception. When a plural feature intervenes between a singular subject and its verb, the phenomenon of **plural attraction** can trigger an error, like the following:

(10) a. The time for fun and games are over.
b. The illiteracy level of our children are appalling.

In a landmark series of experiments, Bock and Miller (1991) presented English speakers with pre-recorded audio sentence preambles like the ones in (11); the participants' task was to complete the sentences as quickly as possible.

(11) a. The bridge to the islands …
b. The bridges to the island …
c. The bridge to the island …
d. The bridges to the islands …

Bock and Miller found that in the sentence completions produced by participants, agreement errors were about ten times more likely with preambles like (11a) than any of the other three. (Bear in mind that the overall proportion of errors in the experiments was always extremely low, typically below 2 percent.) Errors like **The bridge to the islands close at seven* are frequent, not only in speech but also in all sorts of writing – from unedited student essays to heavily edited periodicals and books.

The effect has also been replicated in dozens of studies, not only in English but also in a number of other languages, including French, Dutch, Italian, and Spanish (Vigliocco et al. 1996).

Evidently, there is something special about plural morphemes. When the structural path between a singular verb and its subject is interrupted by a plural feature, an error is more likely than when a singular feature interrupts the path between a plural verb and its subject.

Applying grammatical constraints in real time is something we are able to do automatically and without conscious effort, but certain configurations, structures like those in (10) and (11a), are more likely to trigger errors.
Plural attraction errors are yet another instance of the interplay between linguistic and non-linguistic information; the *marking and morphing* model developed by Kay Bock and colleagues (Bock et al. 2001) makes some explicit links between intended meanings and the linguistic representations created during sentence production. Plurality is assigned to nouns based on the intended meaning, a process called *number marking*. A separate process, called *number morphing*, adds number features to verbs, based on the subject they must agree with.

Attraction errors emerge during number morphing.

Building complex structure

A major goal of grammatical encoding is to create a syntactic structure that will convey the meaning the speaker intends. This requires accessing the speaker's grammar. In Chapter 2 we noted that one of the tasks of the grammar is to combine simple sentences into complex, multiclausal sentences. It turns out that this function of the grammar has a number of important psycholinguistic ramifications. Ferreira (1991) compared speech initiation times associated with sentences with a simple subject NP, such as (12a), to sentences with complex subjects, such as (12b) (which contains a relative clause), and found that speech initiation times for sentences with complex subjects were significantly longer than for sentences with simple subjects.

(12) a. The large and raging river …
b. The river that stopped flooding …

This finding, replicated by Tsiamtsiouris and Cairns (2009), indicates that planning complex sentence structure recruits more computational resources than does planning simple structures.

In the production of complex sentences, the clause appears to be the primary planning unit. Most speech errors that involve two elements – like the exchanges discussed above, and some other error types discussed below – take place within a single clause. This suggests that sentences are organized in clause-sized bundles before they are produced.

Not surprisingly, clause boundaries have been identified as loci for sentence planning. Numerous studies report more pauses at the beginnings of clauses than within them (Boomer 1965; Ford 1978; Beattie 1980; Butterworth 1980), indicating the presence of planning processes.

McDaniel, McKee, and Garrett (2010) elicited sentences containing relative clauses from children and adults, and found that pauses clustered at the clause boundaries.

Evidence for increased production planning cost associated with subject–object relative clauses (described in Chapter 4, sentence (7a)) comes from a study by Tsiamtsiouris, Cairns, and Frank (2007), who report longer speech initiation times for sentences with subject–object relatives than for sentences with object–subject relative clauses (like (7b) in Chapter 4).
Tsiamtsiouris and colleagues (2007) also observed longer speech initiation times for passive sentences than active sentences, suggesting that producing sentences that are out of canonical word order increases planning cost.

The phenomenon of **syntactic priming** provides further insight into the mechanisms of production planning. Bock (1986) and Bock and Griffin (2000) described an effect they referred to as *syntactic persistence*, by which a particular sentence form has a higher probability of occurrence if the speaker has recently heard a sentence of that form. For example, if you call your local supermarket and ask *What time do you close?*, the answer is likely to be something like *Seven*, but if you ask *At what time do you close?*, the response is likely to be *At seven* (Levelt and Kelter 1982). Speakers (and hearers) automatically adapt themselves to the language around them, and as a consequence align their utterances interactively to those produced by their interlocutors; this process of *interactive alignment* has the useful consequence of simplifying both production and comprehension (Pickering and Garrod 2004). Syntactic priming studies are designed to explore to what extent a structure just heard can affect the structure for an utterance being planned.
They exploit the fact that certain messages can be structured more than one way, as illustrated by the following examples (from Chapter 2):

(13) a. Robert gave a cashmere sweater to his girlfriend.
b. Robert gave his girlfriend a cashmere sweater.

(14) a. John hit the ball.
b. The ball was hit by John.

The example in (13) illustrates alternation between prepositional and double-object datives; the example in (14) illustrates the alternation between actives and passives. In syntactic priming experiments, participants are asked to describe images depicting scenes such as those described by (13) or (14). Prior to their descriptions, participants have just heard (either from a recording, or from an investigator or another participant in the experiment) a different sentence containing one of the structures of interest. For example, a person asked to describe a picture of John hitting a ball might have just heard a completely different sentence structured as active (e.g., *Mary is eating the cherries*) or passive (e.g., *The cherries are being eaten by Mary*). The sentence just heard will prime the structure of the sentence being produced; that is, the participant's description of the target picture will be more likely to match the structure of the prime sentence just heard.
Syntactic priming has been used to study a number of aspects of production.
One such aspect is production complexity. Smith and Wheeldon (2001) demonstrated that production is facilitated for a structure that has just been heard; speech initiation times were shorter for sentences with primed structures than for those with unprimed structures.
Tsiamtsiouris and Cairns (2009) replicated those findings. Another question pursued by this line of research is what psychological mechanisms underlie syntactic priming. A common view is that once a particular structure has been constructed, it remains for some time as a memory trace and facilitates the construction of a similar structure.

Syntactic priming is a robust effect, which has even been documented across languages, when the two languages involved have comparable alternative structures. Studies that have examined priming between languages, with bilinguals or second language learners, have confirmed that the structure of an utterance heard in one language can affect the structure of an utterance produced in another language (Loebell and Bock 2003; Hartsuiker, Pickering, and Veltkamp 2004). The study of syntactic priming between languages contributes to current models of the type of cognitive architecture that supports some of the linguistic behaviors bilinguals can engage in: code-switching, borrowing, and transfer (Loebell and Bock 2003). If structures in one language can prime structures in another language, the two languages of a bilingual are not impermeable and fully separate; instead, the same language production mechanism (susceptible to what the system has previously perceived) is recruited for language production, regardless of the language of the utterance.

Preparing a phonological representation

The mental representation of a sentence that serves as input to the systems responsible for articulation (speech, writing, or gestures) is phonological.
Some examples of slips of the tongue discussed earlier reflect the application of morphophonological rules, as a phonological representation for a sentence is prepared during production. There is an entire class of speech errors involving units of analysis that are smaller than phrases or words or morphemes, and these errors shed further light on the nature of the phonological representations built during language production. Consider the following:

(15) a. hass or grash
{hash or grass}

b. I can't cook worth a cam.
{I can't cook worth a damn.}

c. taddle tennis
{paddle tennis}

The example in (15a) is an example of a **segment exchange error**, in which the exchange is between two phonological elements: the final consonants in the two words.

In (15b), we have an example of a **perseveration error**, in which a segment (in this case the /k/ of *can't*) perseveres and intrudes in a later word (so the speaker utters *cam* rather than *damn*). In (15c), the example is of an **anticipation error**, in which a speech sound that has not yet been produced (the /t/ of *tennis*) intrudes in an earlier word.

Speech errors involving phonological segments never create phonemes that are not part of the phonemic inventory of the speaker's language, nor do they create words that violate the phonotactic or phonological rules of the speaker's language. A speaker might slip and say *tips of the slung*, but never **tlips of the sung*, because in the latter a sequence has been created that violates phonotactic constraints for English (Fromkin 1973). There are many other phonologically based regularities connected with speech errors. Consonants and vowels never substitute for one another, and substitutions and exchanges take place only between elements that are phonologically similar.
Errors like those in (15) demonstrate that there is a level of representation in which phonological elements are represented segmentally.
Such errors are revealing about the psychological reality of linguistic representations before sound is produced. Errors like these – anticipation errors in particular – demonstrate that there is a mental representation containing the phonological form of a sentence, some time before a sentence is actually produced. This representation is quite

abstract, as illustrated by the following exchanges, where what is exchanged is not a full phonological segment but only some of its phonological features:

(16) a. pig and vat
{big and fat}

b. spattergrain
{scatterbrain}

In (16a), what is exchanged is voicing: voiced /b/ is produced as voiceless [p], and voiceless /f/ is produced as voiced [v]. In (16b), place of articulation is exchanged: velar /k/ becomes bilabial [p], and bilabial /b/ becomes velar [g].
A final type of word exchange errors, in (17), illustrates that prosodic information is also supplied by the mental representation of a sentence, independently of the lexical items involved, but based on the syntactic structure of the sentence. This includes information about which words in the sentence will receive prosodic prominence (words with a focus accent are in capital letters):

(17) a. When the PAPER hits the story …
{When the STORY hits the paper …}

b. Stop beating your BRICK against a head wall.
{Stop beating your HEAD against a brick wall.}

In (17a), a focus accent occurred on *paper*, which landed in the same position in the sentence as *story* should have been: prosodic prominence applied based on the structure of the clause, rather than based on being associated with a particular lexical item. Put a different way, the focus accent – being associated with the structure of the sentence, rather than a particular word – did not move with the lexical item. If that had happened, the result would have been *When the paper hits the STORY*. The same phenomenon is illustrated in (17b), where again the prosodic prominence is associated not with a particular word but with a particular structural position in the sentence.

Summary: Sentence planning

Sentence planning is the link between the idea the speaker wishes to convey and the linguistic representation that expresses that idea. It must include words organized into an appropriate syntactic structure, as sentence meaning depends upon lexical items and their structural organization. From speech errors we have evidence for the psycholinguistic representation of words and their phonological forms, the representation of morphemes, and levels of sentence planning. Experiments that elicit various types of sentences offer evidence that clauses are planning units, and that multiple factors influence the resources recruited in sentence production.

The sentence planning process ends with a sentence represented phonologically (at both the segmental and suprasegmenta level), to which phonological and morphophonological rules have applied to create a detailed phonetic representation of the sentence, which now needs to be transformed into an actual signal, an utterance. This is the topic of the next section.

Producing Speech After It Is Planned

The abstract phonetic representation of the speaker's sentence is sent to the central motor areas of the brain, where it is converted into instructions to the vocal tract to produce the required sounds. Speaking is an incredibly complex motor activity, involving over 100 muscles moving in precise synchrony to produce speech at a rate of 10 to 15 phonetic units per second (Liberman et al. 1967). During silence, the amount of time needed for inhaling is about the same as for exhaling.

Respiration during speech is very different: the time for inhaling is drastically reduced, sometimes to less than half a second, and much more time is spent exhaling, sometimes up to several seconds. During speech, air from the lungs must be released with exactly the correct pressure. The respiratory system works with the muscles of the larynx to control the rate of vibration of the vocal folds, providing the necessary variations in pitch, loudness, and duration for the segmental (consonants and vowels) and suprasegmental (prosody) content of the utterance.

Muscles of the lips, the tongue, and other articulators must be carefully coordinated. Much precision of planning is required. For example, to make the vowel sound [u], different sets of nerves lower the larynx and round the lips. Impulses travel at different rates down those two sets of nerves, so timing must be carefully orchestrated: one impulse must be sent a fraction of a millisecond sooner than the other. This is an example of the level of planning carried out by the central planning system in the brain.

In this section, we examine how vowels and consonants are produced, with a focus on how the articulation of speech converts a sequence of discrete mental units (a phonological representation) into a continuous acoustic signal. The signal, as the end product for the speaker and the starting point for the hearer, must contain sufficient information for successful decoding. Our objective, then, is to identify some of the characteristics of the signal which carry information that will be used by the hearer.

The source-filter model of vowel production

Speech consists of sounds generated at the vocal folds being filtered as they travel through the vocal tract. (Figure 5.3 repeats a diagram used in Chapter 2, identifying the organs involved in producing speech, for your reference while reading this section.) The **source–filter model** of vowel production breaks down the process of producing vowels into two component parts: a source and a filter.

We will illustrate how the source–filter model works by considering the vowels [i], [a], and [u]. To articulate these vowels, you open your mouth and force air from your lungs through your larynx, where the vocal folds reside. This causes the vocal folds to vibrate – that is, to open and close in rapid sequence. The frequency of this vibration is called the **fundamental frequency** (or **F0**), and this is, in essence, the **source** in the source–filter model of speech production. Sounds with higher frequency are higher in **pitch**, pitch being the perceptual

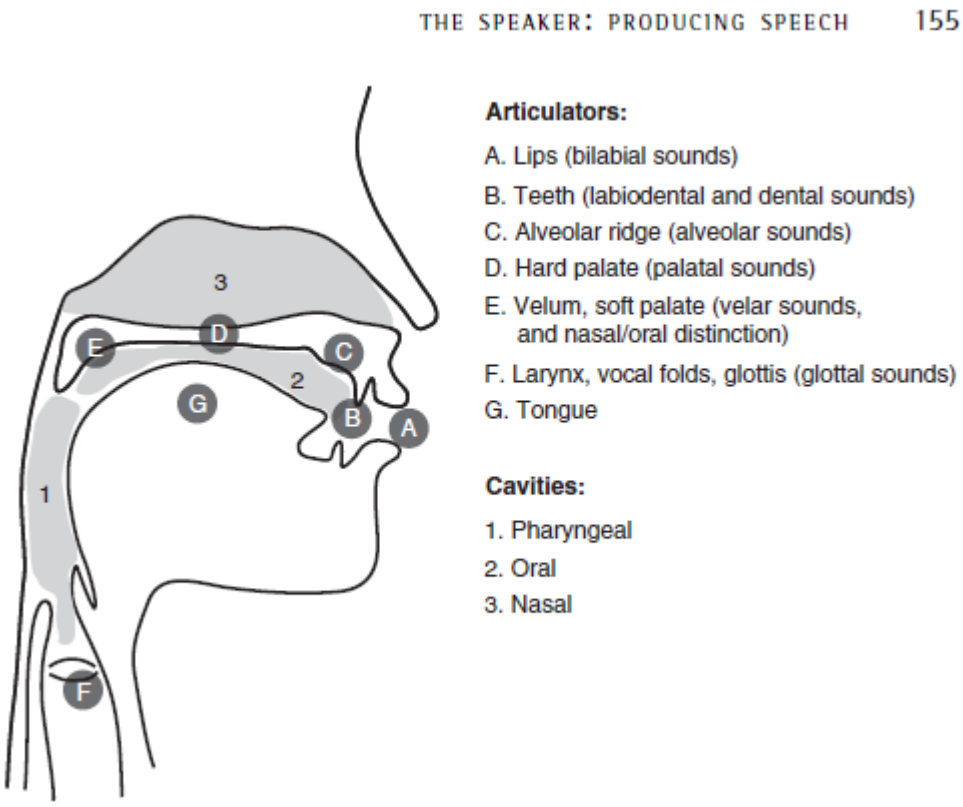


Figure 5.3 Diagram of the vocal tract, identifying the organs involved in producing speech (articulators) and the spaces in which speech sounds resonate (cavities). This figure repeats Figure 2.1 from Chapter 2.

correlate of fundamental frequency. Overall, men have lower pitch than women, who in turn have lower pitch than children (Katz and Assmann 2001). These differences are directly related to sex- and agebased differences in size of physique. In general, smaller vocal folds vibrate at a higher frequency, so people with small larynxes speak with higher overall pitch. Different vowels also vary in pitch: high vowels like [i] and [u] tend to have higher fundamental frequency than low vowels like [a] (Whalen and Levitt 1995). However, it is not F0 that serves to distinguish vowels from one another – after all, hearers distinguish between vowels uttered with high pitch just as well as between

vowels uttered with low pitch. Vowels are distinct from each other based on their acoustic form, or spectral properties, which we describe below.

A tuning fork creates a *simple* sound, with energy at a single frequency.
The left panel of Figure 5.4 shows some of the acoustic characteristics of a simple sound, a computer-generated pure tone: its waveform is evenly sinusoid (a sine wave) and its spectrogram has

THE SPEAKER: PRODUCING SPEECH 156-157

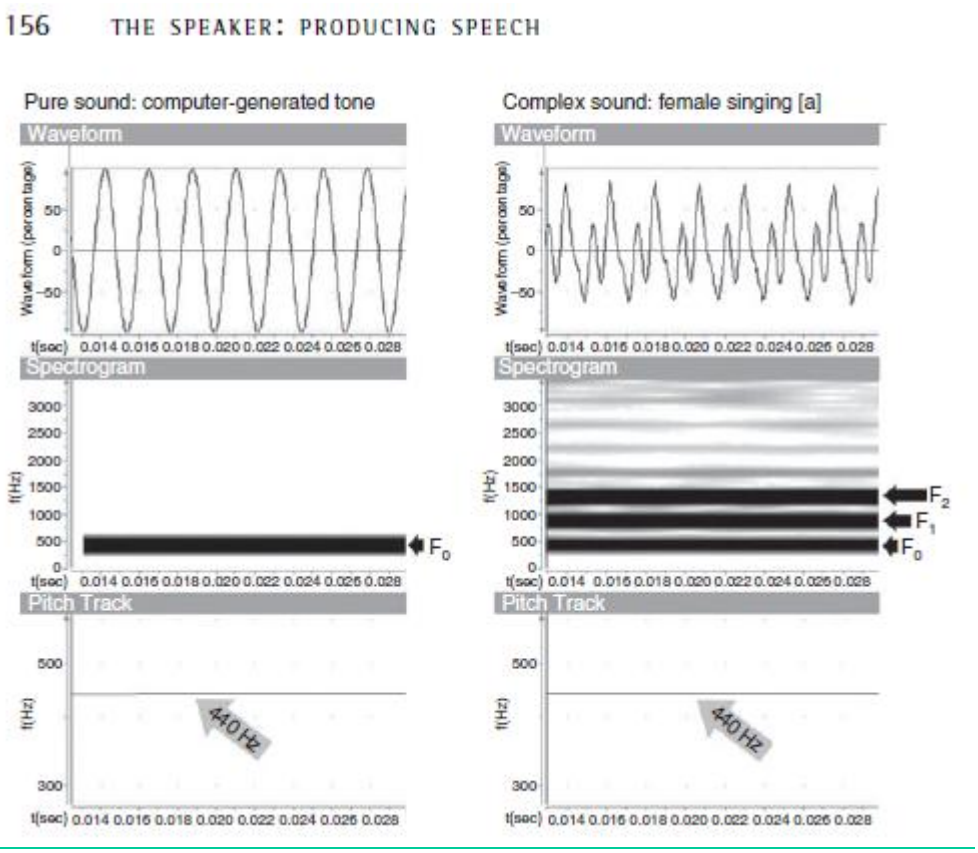


Figure 5.4 Waveform, spectrogram, and pitch track for a computer-generated pure tone (left panel) and a human-articulated complex tone (right panel). The two bottom graphs indicate that both tones have the same pitch, 440 Hz – which happens to be the pitch for the note called Concert A (the A above middle C).

Notice how the waveform of the computer-generated pure tone is perfectly sinusoid, unlike the waveform of the human-generated complex tone. Notice also, by comparing the spectrograms, that the pure tone has only one band of energy (the fundamental frequency, F0), while the complex sound has multiple bands of energy, the strongest being the fundamental frequency (F0), and the first two formants (F1 and F2).

only one band of energy, the one corresponding to the tone's F0. Yet most sounds people hear on a daily basis – speech, music, and so on – are *complex*. A **complex acoustic signal** is one that has energy at many frequencies in addition to the fundamental frequency. The graphs in the right panel of Figure 5.4 correspond to another tone, with the same fundamental frequency as the pure tone, only this one was produced by a female singing the vowel [a]. The complex sound wave generated by the vibrations of the human vocal folds is a complex sound, with acoustic energy at many frequencies. The frequencies carrying acoustic energy are multiples of the fundamental frequency of the voice. So a person who is speaking with a fundamental frequency of 150 cycles per second (cps) – also referred to as 150 Hertz (Hz) – will produce a complex sound wave with energy at 300 cps, 450 cps, 600 cps, and so forth. These bands of energy are called **harmonics**. For a given speech sound, the F0 and its **formants** – which we will define in a moment – are the sound's **spectral properties**.

How to read waveform graphs, spectrograms, and pitch track graphs

Some of the figures in this chapter (and elsewhere in the book) incorporate images generated by acoustic analysis software. The software we used is *Speech Analyzer* (SIL International 2007); other similar tools include *Computerized Speech Lab* (KayPENTAX 2008) and *Praat* (Boersma and Weenink 2009).

The figures contain waveforms, spectrograms, and pitch tracks for audio recordings illustrating various types of sounds. **Waveform** graphs, also called *oscillograms*, display time horizontally and air pressure variations vertically. **Spectrograms** display the spectral properties of the recorded sounds. A spectrogram plots information along three dimensions: time is displayed horizontally, frequency (in Hz) is displayed vertically, and energy intensity is indicated by shades of gray (the darker the gray, the more intense the energy). Finally, **pitch track** graphs give an estimate of pitch movements by plotting time on the horizontal axis and fundamental frequency (in Hz) on the vertical axis.

Key to understanding how the vocal tract acts as a **filter** is the concept of **resonance**. The vocal tract changes shape when different sounds are articulated. For example, when the vowels [i] and [u] are articulated, the tongue body is relatively high, compared to when [a] is articulated; the tongue body is farthest in the front of the mouth when articulating [i], slightly farther back for [a], and farthest in the back for [u]. For these three vowels, then, the oral and pharyngeal cavities are shaped slightly differently, relative to each other. Consequently, for a sound generated at the vocal folds traveling through these differently shaped cavities, some harmonics will be reinforced, and other harmonics will be cancelled.

In other words, energy at some frequencies will increase, and energy at other frequencies be eliminated. This is resonance.

Figure 5.5 shows plots of the average fundamental frequency and two bands of reinforced harmonics (**formants**) associated with the vowels [i], [a], and [u], as produced by four different speakers. The formant with the lowest frequency is called the **first formant (F1)**,

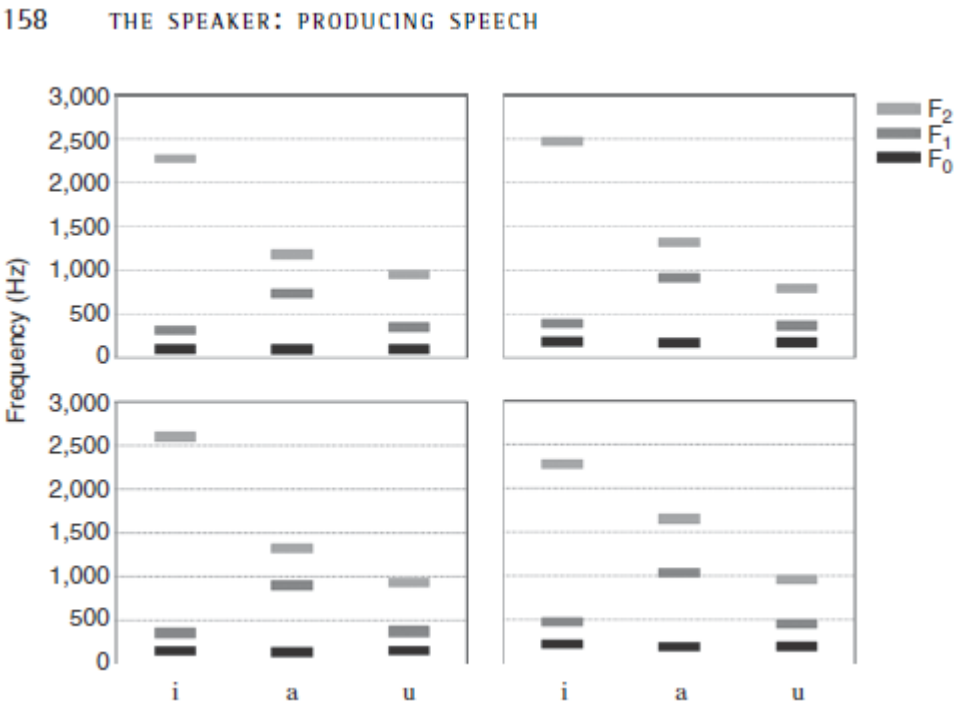


Figure 5.5 Average F0 (fundamental frequency, black), F1 (first formant, medium gray), and F2 (second formant, light gray), for the vowels [i], [a], and [u], as uttered by four speakers of American English: an adult male (top left), an adult female (top right), a young male (12 years old, bottom left), and a young female (11 years old, bottom right).

the second lowest is the **second formant (F2)**. While all four speakers have different F0 averages (the adult male has the lowest, 180 Hz on average, the young female the highest, 221 Hz on average), the pattern for F1 and F2 with respect to each other and with respect to F0 is remarkably similar. Figure 5.5 sketches only F1 and F2 because those two formants are sufficient to illustrate the distinctions between our example vowels. In Figure 5.5, it is clear that [a] has a much higher F1 than [i] or [u]; [i] is distinct from [u] because it has a very high F2.

The vowels [i], [u], and [a] are often called *point vowels* because they represent the maximal extent of F1 and F2 variation. A graph plotting these two dimensions relative to each other, also called a *vowel triangle*, is presented in Figure 5.6. All of the world's languages have, at minimum, these three vowels, but many languages have several others, representing other combinations of F1 and F2 within the vowel space.

Resonance depends on the size and shape of the filter – the cavities a sound travels through. F1 and F2 thus vary based on the size and shape of the pharyngeal and oral cavities, the size of which is determined by the position of the tongue. F1 correlates with the width of the pharyngeal

159-160- 161 **THE SPEAKER: PRODUCING SPEECH**

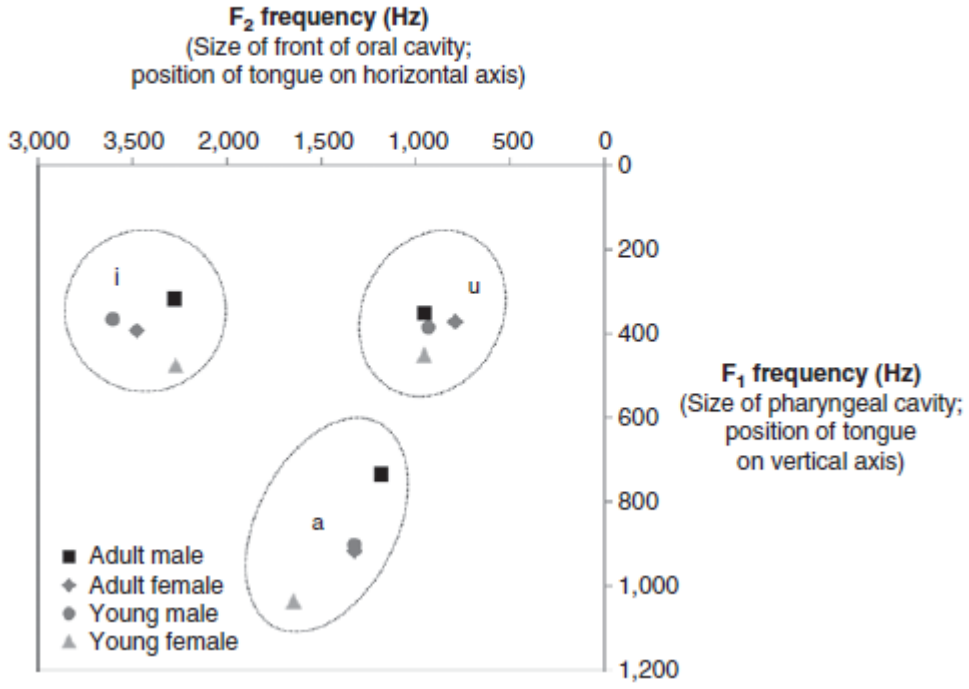


Figure 5.6 F1 and F2 data from Figure 5.5, plotted together to represent the vowel triangle. If you connect the dots from [i] to [u] to [a] and back to [i], for the data points corresponding to each of the speakers, you will come up with an upside-down triangle for each. The horizontal and vertical axes in the graph are plotted in reverse, so that the high front vowel is plotted on the top left, as in conventional vowel charts.

cavity and with the position of the tongue on a vertical axis: F1 frequency is higher when the pharynx is narrow and the tongue is low, as with the low vowel [a]. F2, in contrast, correlates with the length of the front of the oral cavity and the position of the tongue on a horizontal axis: F2 frequency is higher when the oral cavity is short and the tongue is forward in the mouth, as in the production of [i]. The vowel [u] has a low F2 because the oral cavity is elongated, as the lips are rounded and the larynx is lowered.

The exact frequencies of the formants will differ from speaker to speaker, generally being higher for women than for men. For instance, Kent (1997) reports – based on data from Hillenbrand et al. (1995) – an average F1 and F2 for the vowel [i] for male speakers of 342 Hz and 2,322 Hz, respectively. For female speakers, the formants average 437 Hz and 2,761 Hz. Exact frequencies will also differ in rapid speech, because the articulators might not have time to reach their target position. However, the relationship between F1 and F2 will be the same for vowels in every speech situation.

Acoustic characteristics of consonants

A complete description of the **acoustic characteristics** of speech sounds is beyond the scope of this book, but there are some general properties of certain classes of consonants that are worth pointing out. Table 5.1 has additional details and examples. In Chapter 2 we distinguished between obstruent and sonorant consonants. We use this distinction here again. Obstruents are characterized by an obstruction in the vocal tract during articulation. Full closure followed by release is the characteristic feature of **stops**, like [p] and [t]. The acoustic indicator of closure is silence.

A feature distinguishing between many consonants is voicing. For voiced sounds, like [z], the vocal folds are engaged during the articulation of the consonant. For voiceless sounds, like [s], voicing will not begin until the vowel that follows is articulated. The acoustic indicator of voicing is fundamental frequency. For stops, like [b] and [p], the key acoustic indicator of voicing is *voice onset time* (frequently abbreviated as *VOT*). VOT is the time between the release of closure of a stop and the onset of voicing. Voiced stops have very short VOTs, while voiceless stops have relatively longer VOTs. In Chapter 6, we will discuss how the continuous variation in VOT is perceived categorically by speakers of different languages.

Fricatives, like [s] and [ʃ] are produced by creating turbulence as air is forced through two articulators, a sound much like the hiss of white noise; the acoustic indicator of such turbulence is high-frequency noise.
Articulating the third class of obstruent consonants, **affricates**, like [tʃ] and [ʤ], involves combining a stop and a fricative. Affricates, therefore, have acoustic properties of both stops and fricatives: silence followed by sustained high-frequency noise.

Sonorants, being close to vowels in their articulation, are close to vowels in their acoustic form, and therefore have the characteristic formant configurations of vowels. In articulating **nasals**, like [n], [m], and [ŋ], the velum – the flap that opens and closes the opening between the nasal cavity and the oral cavity – is lowered; as a result, the resonance of the air in the nasal cavity combines with the resonance of the oral cavity.

The nasal cavity causes resonances to decrease in energy, resulting in an overall attenuation of the signal. (You might have noticed that humming is never as loud as regular singing. Humming involves resonance in the nasal cavity.) **Approximants** –

which include **liquids** (e.g., [l] and [r]) and **glides** (e.g., [w] and [y]) – are very vowel-like and have clear formant structure. The two liquids in English, [l] and [r], have similar articulation, but differ in terms of tongue placement, as described in Table 5.1. The acoustic consequence of this articulation difference is reflected acoustically in the shape of the third formant (F3), as shown in the spectrograms for these sounds in Table 5.1.

Coarticulation

Probably the most important psycholinguistic aspect of speech production is the phenomenon of **coarticulation**. Coarticulation simply means that the articulators are always performing motions for more than one speech sound at a time. The articulators do not perform all the work for one speech sound, then another, then another. The genius of speech production is that phonological segments overlap, so the articulators work at maximum efficiency, in order to be able to produce 10 to 15 phonetic segments per second – more in rapid speech. This transmission speed would be close to impossible to achieve if each phonological unit were produced individually. As it is, speech is produced more slowly than necessary for the speech perception system. People can actually understand speech that has been sped up (compressed) at several times the normal rate (Foulke and Sticht 1969). But coarticulation is not just a matter of convenience for the speaker: if speech were not coarticulated – that is, if phonological units did not overlap – speech would actually be too slow and disconnected for the hearer to process it efficiently.

A simple example of coarticulation is the articulation of [k] in *key* and *coo*. When uttering *key*, while the back of the tongue is making closure with the top of the mouth for the [k], the lips – not ordinarily involved in articulating [k] – begin to spread in anticipation of the following vowel [i]. Similarly, when uttering *coo*, the lips round during the articulation of [k], in anticipation of the upcoming [u]. One aspect of coarticulation, then, is that the actual articulation of a phonological segment can be influenced by upcoming sounds. This is sometimes referred to as *regressive assimilation*.

Coarticulation can also be influenced by a phonological segment that has just been produced, a phenomenon sometimes called *progressive assimilation*. The [t] in *seat* is pronounced slightly more forward in the mouth than the [t] in *suit*. This is because the tongue position for the [t] is influenced by the preceding vowel ([i] is a front vowel and [u] is a back vowel).

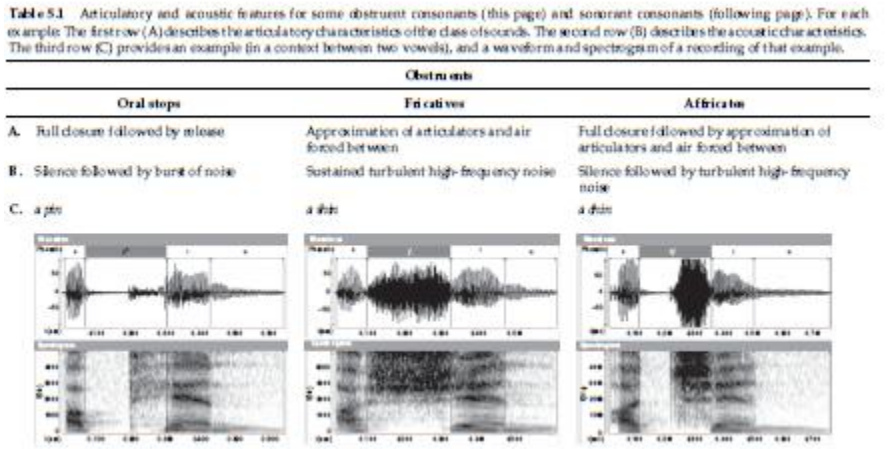


Table 5.1 Articulatory and acoustic features for some obstruent consonants (this page) and sonorant consonants (following page). For each example: The first row (A) describes the articulatory characteristics of the class of sounds. The second row (B) describes the acoustic characteristics. The third row (C) provides an example (in a context between two vowels), and a waveform and spectrogram of a recording of that example.

	Obstruents		
	Oral stops	**Fricatives**	**Affricates**
A.	Full closure followed by release	Approximation of articulators and air forced between	Full closure followed by approximation of articulators and air forced between
B.	Silence followed by burst of noise	Sustained turbulent high-frequency noise	Silence followed by turbulent high-frequency noise
C.	*a pin*	*a thin*	*a thin*

Table 5.1 Articulatory and acoustic features for some obstruent consonants (this page) and sonorant consonants (following page). For each example: The first row (A) describes the articulatory characteristics of the class of sounds. The second row (B) describes the acoustic characteristics.
The third row (C) provides an example (in a context between two vowels), and a waveform and spectrogram of a recording of that example.

Obstruents Oral stops Fricatives Affricates

A. Full closure followed by release Approximation of articulators and air forced between Full closure followed by approximation of articulators and air forced between

B. Silence followed by burst of noise Sustained turbulent high-frequency noise Silence followed by turbulent high-frequency noise

C. *a pin*

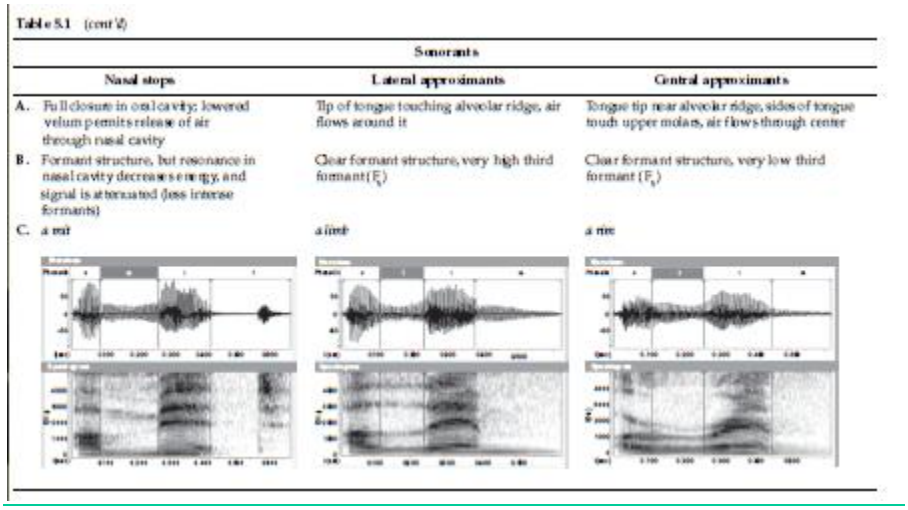


Table 5.1 (cont'd)

	Sonorants		
	Nasal stops	Lateral approximants	Central approximants
A.	Full closure in oral cavity; lowered velum permits release of air through nasal cavity	Tip of tongue touching alveolar ridge, air flows around it	Tongue tip near alveolar ridge, sides of tongue touch upper molars, air flows through center
B.	Formant structure, but resonance in nasal cavity decreases energy, and signal is attenuated (less intense formants)	Clear formant structure, very high third formant (F_3)	Clear formant structure, very low third formant (F_3)
C.	a mit	a lit	a rit

Table 5.1 (*cont'd*)
Sonorants Nasal stops Lateral approximants Central approximants

A. Full closure in oral cavity; lowered velum permits release of air through nasal cavity Tip of tongue touching alveolar ridge, air flows around it Tongue tip near alveolar ridge, sides of tongue touch upper molars, air flows through center

B. Formant structure, but resonance in nasal cavity decreases energy, and signal is attenuated (less intense formants) Clear formant structure, very high third formant (F3) Clear formant structure, very low third formant (F3)

C. *a mit*

Coarticulatory effects can span several segments. For example, the [b] in *bag* will be articulated slightly differently than the [b] in *bat*, as a consequence of the differences in the syllable-final phonemes [g] and [t]. What is most important for the present discussion, however, is that sounds produced by speakers are not discrete (separate) units, but rather form part of a continuous speech signal. The mental representation of the phonological form of an utterance is definitely segmental, phonemes lined up one after the other; however, in the process of speaking, phonemes overlap and blur together.

The linguist Charles Hockett offered an apt metaphor for coarticulation (1955: 210):
Imagine a row of Easter eggs carried along a moving belt; the eggs are of various sizes, and variously colored, but not boiled. At a certain point, the belt carries the row of eggs between the two rollers of a wringer, which quite effectively smash them and rub them more or less into each other. The flow of eggs before the wringer represents the series of

impulses from the phoneme source; the mess that emerges from the wringer represents the output of the speech transmitter. At a subsequent point, we have an inspector whose task it is to examine the passing mess and decide, on the basis of the broken and unbroken yolks, the variously spread-out albumen, and the variously colored bits of shell, the nature of the flow of eggs which previously arrived at the wringer.

Hockett's words foreshadow discussion, in Chapter 6, of the effect of coarticulation on the perception of speech.

One final aspect of coarticulation is central to understanding the production (and perception) of stop consonants. Stops involve producing a complete closure somewhere in the vocal tract: [p] and [b] involve closure at the lips, [t] and [d] closure at the alveolar ridge, and [k] and [g] closure at the velum.

The effects of coarticulation can be seen in Figure 5.7, which provides the waveform and spectrogram for the same vowel preceded by three stop consonants with different place of articulation:

bilabial [ba], alveolar [da], and velar [ga]. F1 is very similar in all three syllables, curving upwards from the onset of voicing to the steady formant of the vowel. F2, in contrast, is slightly different for each syllable:

it starts low and curves upwards for [ba], but it starts high and curves downwards for [da] and [ga].

The spectrogram reflects the changing shape of the oral cavity from the moment the stop is released (when voicing begins) and the tongue moves into position for the vowel. The movement of the tongue is tracked as **formant transitions**, visible in Figure 5.7 as lines of frequency

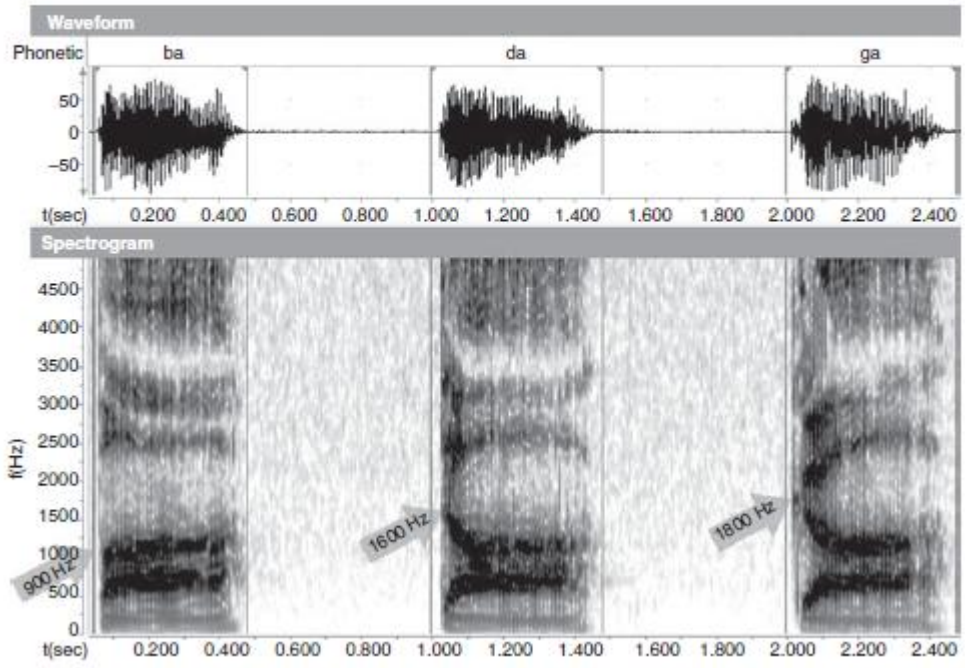


Figure 5.7 Waveform and spectrogram for three syllables produced by a male speaker of American English: [ba], [da], and [ga]. F1 is similar for all three syllables, but F2 differs. The F2 for the vowel is about 1,100 Hz in all three cases, but F2 begins at 900 Hz and rises to 1,100 Hz for [ba] on the left; F2 begins at 1,600 Hz and falls for [da], in the middle; and F2 begins at 1800 Hz and falls for [ga], on the right.

that change rapidly, as the shape of the oral cavity rapidly changes before assuming the final position for the vowel. The segmental nature of the representation before production has been transformed into a continuous signal, with information about the two segments combined and spread over less than 100 milliseconds of sound. Most remarkable about this is the fact that place of articulation information for stop consonants is actually carried by the vowel rather than the consonant itself.

■ Words in Speech

The speech rate of 10 to 15 phonetic units every second works out to about 125 to 180 words per minute; conversational speech can be much faster, reaching up to 300 words per minute. When people talk, they do

THE SPEAKER: PRODUCING SPEECH 166-167

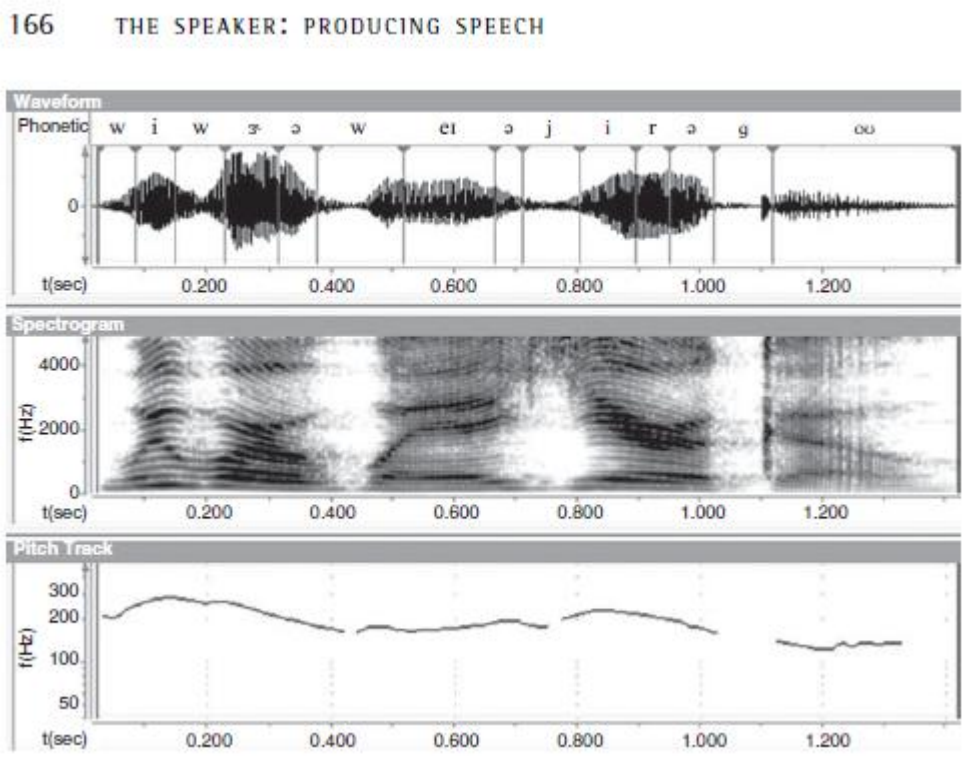


Figure 5.8 Waveform (top panel), spectrogram (middle panel), and pitch track (bottom panel) for the sentence *We were away a year ago*, spoken by a female speaker of American English. The vertical lines in the waveform indicate the approximate boundaries between segments. As can be appreciated by inspecting all three graphs, the signal is continuous.

not pause between words; words are run together just as the phonetic units are. In a continuous speech signal, neither the phones nor the words are segmented. Figure 5.8 provides an acoustic snapshot of the sentence *We were away a year ago*, produced by a female speaker.

The figure shows how words flow continuously, without spaces or discontinuities, from one to the next. In fact, the only period of silence is associated with the stop consonant [g] in the word *ago*. The word boundaries are completely obliterated by the continuous movement of the articulators as the sentence is produced.

■ Summing Up

The production of even a fairly simple sentence requires a complex coordination of preproduction planning of structure, lexicon, and phonology, followed by a series of movements that are highly organized and precisely coordinated. Underlying the actual production of a continuous, coarticulated speech signal is an abstract representation of

individual words made up of segmented phones. Psycholinguists know that this representation precedes the production of speech because speech errors demonstrate that individual phones and words move as units. Like all psycholinguistic processes, the planning and execution of sentence production is effortless and unconscious, even though it is extremely complex. The complexity is related to the fact that language production recruits vast amounts of information (lexical and grammatical, as well as real-world knowledge), and it is sensitive to both the context of the conversation and the speaker's and hearer's communicative intents. The speaker transmits the speech signal, which is the outcome of this process, to the hearer, whose job it is to recover the speaker's idea by making sense of those sound waves, by recreating an abstract representation of discrete linguistic units, using the information carried by the continuous speech signal. Chapter 6 focuses on the hearer's task.

New Concepts

- acoustic characteristics of:
  - affricates
  - approximants
  - fricatives
  - glides
  - liquids
  - nasals
  - stops
- anticipation errors
- bilingual mode
- borrowing
- coarticulation
- code-switching
- complex acoustic signal
- exchange errors
- F_0, fundamental frequency
- F_1, first formant
- F_2, second formant
- filter
- formant transitions
- formants
- grammatical encoding
- harmonics
- interlocutor
- lexical retrieval
- perseveration errors
- pitch
- pitch track
- plural attraction
- preverbal message
- resonance
- segment exchange errors
- source
- source–filter model
- spectral properties
- spectrogram
- speech errors (slips of the tongue)
- syntactic priming
- tip-of-the-tongue phenomenon
- transfer
- unilingual mode
- waveform
- word exchange errors

Study Questions

1. How can the study of speech errors demonstrate that speech consists of segmented words and phonemes before it is produced? Why is this interesting?

2. What are some of the similarities and differences between monolingual and bilingual models of production?

3. How does the study of speech errors demonstrate that speech is rep resented at various processing levels before it is actually produced?

4. What characteristics of speech errors demonstrate that they are not random, but honor linguistic classifications and constraints?

5. At some point before an utterance is produced it is represented in a form to which phonological and morphophonological rules have not yet applied. What characteristics of speech errors support this claim?

6. Freud suggested that word retrieval errors were a result of repressed feelings. Consider the following spoonerism: Work is the curse of the drinking classes. What is the psycholinguistic view of this error?

7. How do studies of syntactic priming demonstrate that speakers and hearers align their utterances interactively in conversations?

8. In the source–filter model of vowel production, what is the source? What is the filter? How do the source and filter operate together?

9. What are the primary acoustic characteristics of different classes of consonants? How is the acoustic signal related to articulation?

10. What is coarticulation? Why is it such an important feature of speech production?

11. Why do psycholinguists say that the speech signal is continuous? Are mental representations of sentences, before they are produced, also continuous? What is the evidence for this?

6 The Hearer: Speech Perception and Lexical Access

Chapter 5 dealt with the operations the speaker performs, using knowledge of language, when encoding a mental message into a physical signal accessible to the hearer. The hearer's task is almost the mirror image of the speaker's task. First, using information from the acoustic signal, the hearer reconstructs a phonological representation. The hearer enters the lexicon using that phonological representation to retrieve the lexical items that match. This permits the hearer to recover the semantic and structural details of the words in the message. The next

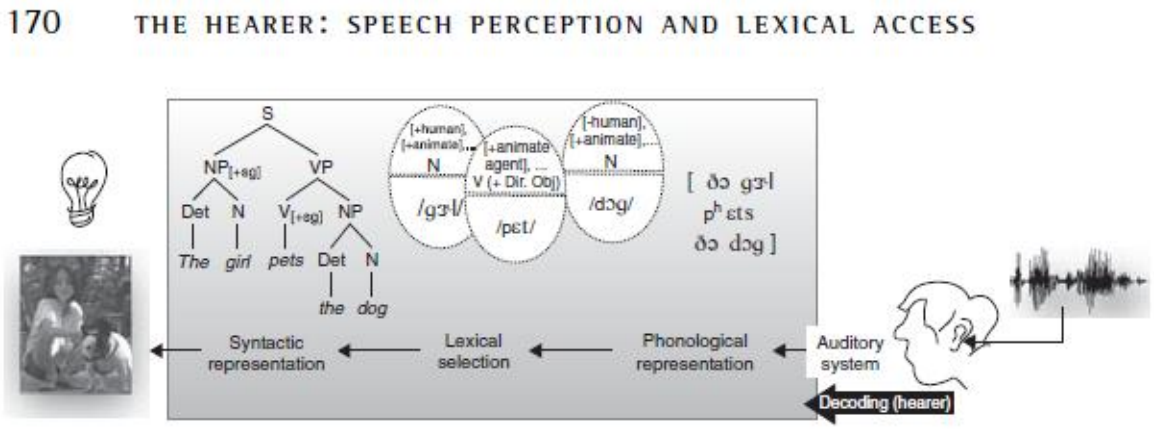


Figure 6.1 Diagram of some processing operations, ordered right to left, performed by the hearer when decoding the sentence *The girl pets the dog*. (This figure expands on parts of Figure 1.3, Chapter 1, and is parallel to Figure 5.1, Chapter 5.) The speech signal on the far right, perceived by the auditory system, serves to recover the phonological form for the sentence, indicated by the phonetic transcription. The capsule-like figures in the middle represent lexical
items, activated by their phonological form (bottom half), but whose morphosyntactic

features (included in the top half) help the processor recover the intended syntactic structure. The tree diagram on the left represents the sentence's syntactic form, used to decode the meaning of the sentence. The light bulb indicates that the hearer has successfully recovered the idea the speaker intended to convey.

step is to reconstruct the structural organization of the words, to create a syntactic representation – necessary for recovering the meaning of the sentence. This chapter and Chapter 7 describe these operations, represented graphically (from right to left) in Figure 6.1.

Chapter 7 focuses on syntactic processing (parsing). In this chapter, we examine the two steps in perception that precede parsing: **speech perception** and **lexical access**. We address these two aspects of perception together because they interact in interesting ways. Both phonetic elements and words must be extracted from a continuous, unsegmented, highly coarticulated signal. There are no spaces between phonetic units, and there are no spaces between words. Thus, some of the same problems exist for both speech perception and lexical access.

Perceiving Speech

The hearer plays the role of the inspector in the metaphor by Charles Hockett cited in Chapter 5 (Hockett 1955: 210). The phonetic "eggs" have been mangled and mixed together by articulatory processes; it is the hearer's task to identify from the resulting mess of the **speech signal** what the original phonetic elements were. There are three features of

171-172 THE HEARER: SPEECH PERCEPTION AND LEXICAL ACCESS

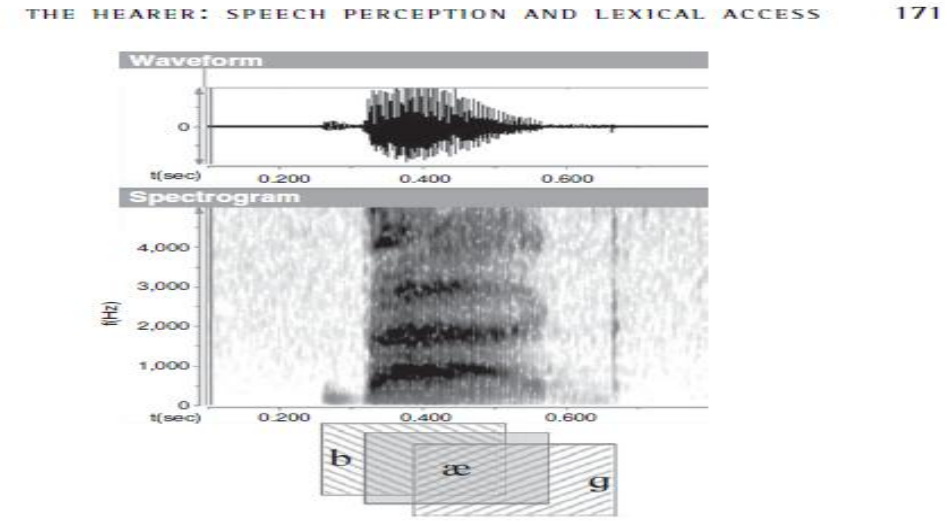


Figure 6.2 Illustration of parallel transmission of phonetic information. The figure is an adaptation of Figure 5 in Liberman (1970: 309).

the speech signal that the speech perception system must deal with: the signal is continuous, it transmits information in parallel, and it is highly variable.

We have pointed out elsewhere (briefly in Chapter 1 and in more detail in Chapter 5) that the speech signal is **continuous**: there are no spaces between consonants and vowels, or even between words. A central objective for the mechanisms involved in speech perception is to segment a continuous signal into discrete units: phonemes, syllables, and, ultimately, words. Because of coarticulation, the speech signal is characterized by the **parallel transmission** of information about phonetic segments (Liberman et al. 1967).

Figure 6.2 illustrates how information about the three phonological units in the word *bag* is distributed across the word. In the recording whose waveform and spectrogram appear in the figure, the vowel has a duration of approximately 250 milliseconds, of which approximately 50 to 75 milliseconds carry information about all three phonological units. Properties of the word-initial /b/ spill into the vowel and persist through the beginning of the word-final /ɡ/. Properties of /ɡ/ begin at the offset of /b/ and continue through the second half of the vowel. The vowel /æ/ influences the pronunciation of the entire word, and carries acoustic information about both of the consonants in the word. This is an example of parallel transmission, of how the speech signal transmits information about more than one phonological unit simultaneously. The speech perception system must sort out all that information and figure out what the units are.

A third feature of the speech signal is its **variability**, or **lack of invariance**.
The abstract mental representation of a phonological element does not vary. However, a speech sound may vary greatly each time it is actually produced. Many factors contribute to the fact that the same consonant or vowel, the same syllable, and even the same word are never pronounced exactly the same.

First, there is variability among speakers. Human anatomy is broadly similar, but there is individual variation in every aspect of our physique, which includes the organs involved in speech production. As a consequence, many aspects of the signal are intrinsically different for different speakers, including fundamental frequency and the spectral properties of consonants and vowels. In fact, a person's voice is as unique an identifier as are the person's fingerprints or retinas.

Second, there is variability within speakers. People sometimes speak fast, and other times slowly; they sometimes speak with chewing gum in their mouths; they mumble; they shout; they speak while being overcome with feelings of sadness or joy. All these variables affect the speech signal, and can make the acoustic signal associated with a single word very different each time it is uttered, even by the same speaker.

A third factor that makes the signal variable is ambient noise. Rarely do we speak to each other in noise-free environments. Other voices and other sounds (like music or traffic) can alter the speech signal dramatically.

The same utterance will sound different in a small quiet room, in a large loud room, or coming from the room next door. The same voice could sound very different in person and on the telephone, and telephone transmissions will vary further depending on the connection and the equipment being used.

A fourth factor affecting variability in the signal is the context. The articulation of phonemes is affected by the phonemes around them, as illustrated in Chapter 5, and as just described with respect to parallel transmission. In addition to effects caused by coarticulation of phonological units, sentence context and neighboring words can also affect the pronunciation of individual lexical items.

THE HEARER: SPEECH PERCEPTION AND LEXICAL ACCESS 173 - 174

The existence of all of these factors affecting the signal suggests that the accurate recognition of phonemes from the speech signal is nothing short of miraculous. In actuality, accurate decoding of speech is the norm, rather than the exception, because the speech perception mechanism operates in ways that overcome the variability of the signal. How does the speech perception mechanism overcome variability, to identify the phonological units that the signal carries? Speech perception relies on the relationships among acoustic elements, such as the fact that the F1 for /a/ is high relative to the F1 for /i/ and /u/, no matter who is speaking. Speech perception also exploits the reliability of certain acoustic cues as associated with distinct phonological units, some of which we discussed in Chapter 5. Stops are associated with a few milliseconds of silence, followed by a burst and a formant transition into the following vowel, glide, or liquid. Fricatives have high-frequency noise. Nasals have an attenuated signal. Glides, liquids, and vowels have clear formant structure. Though these are hardly invariable cues, they provide a great deal of guidance during speech perception.

Since the hearer is also a speaker, he can compensate for much of the variability produced by speaker characteristics, like speech rate and shouting. People use knowledge of their speech production in their perception of the speech of others: "to perceive an utterance is to perceive a specific pattern of intended gestures" (Liberman and Mattingly 1985). Some interesting evidence of how production influences speech perception comes from a speech shadowing study cited by Raphael, Borden, and Harris (2006: 264). In this experiment (Chistovich et al. 1963), the investigators asked participants to shadow speech (that is, to repeat words presented auditorily as fast as they could). Shadowers were able to produce consonants before all of the relevant acoustic cues for those consonants had been heard, suggesting that they were being guided by their production routines. More recently, Shockley and colleagues (Shockley, Sabadini, and Fowler 2004) have demonstrated that shadowers imitate with high fidelity phonetic details of the words

they have just heard, and argue that this perception-driven response is due to a more general tendency of speakers to accommodate their speech (accent, rate, loudness, etc.) to that of the speech they are hearing from their interlocutors. These adjustments have important social consequences: they are ways that help speakers and hearers to get on the same "wavelength." The hearer also adapts rapidly to abnormal situations. For example, speech with a non-native-like accent, and sometimes speech produced by young children, can be difficult to decode, but this difficulty is overcome relatively quickly. Clarke and Garrett (2004) demonstrate that processing is slowed down for English native speakers listening to speech in English produced with a non-native (Spanish or Chinese) accent, compared to speech produced with a native accent. This slowdown, however, is reduced within one minute of exposure to the accented speech.

How much a hearer can adapt to abnormal situations can be affected by many variables. Speech presented in noise, for example, is understood more easily (and more efficiently) by native than non-native speakers. One study compared monolingual speakers of English, bilinguals who were native speakers of Spanish and acquired English early (as infants or toddlers), and people who were native speakers of Spanish and acquired English as a second language after adolescence (Mayo, Florentine, and Buus 1997). Participants were asked to listen to prerecorded English sentences, presented with or without noise, and to indicate what they thought the last word of each sentence was. The preceding context made the target words, at the end of each sentence, either predictable (*The boat sailed across the bay*) or unpredictable (*John was thinking about the bay*).

Monolingual participants were best at tolerating high levels of noise and using the context to predict what they heard, followed by early bilinguals. The participants with the worst performance were those who acquired English after adolescence. These findings suggest that age of acquisition is an important determinant in how accessible high-level information (top-down information, which we will discuss later in the chapter) is for a hearer. The study also demonstrates that perceiving speech involves much more than just experiencing the acoustic signal.

The phonemic inventory and speech perception

Accurate speech perception is efficient and effortless because hearers rely on what they know about the language they are processing. One of the primary sources of information is knowledge of the phonemic inventory – as described in Chapter 2, this is the set of phonemes that are contrastive for the language.

As an example, let us consider how the dramatic variability in the acoustic signal for the consonant /d/, depending on the vowel that follows, is overcome by what an English speaker knows about /d/ as a phoneme. In Chapter 5, we noted that formant transitions depend on both the place of articulation of the consonant and the articulation of the vowel that follows it; formant transitions are essentially the acoustic

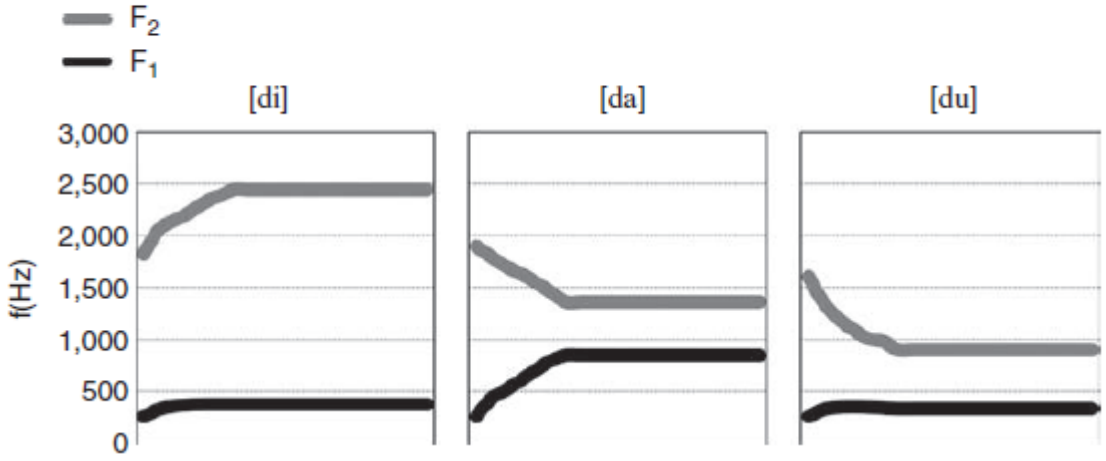


Figure 6.3 F1 and F2 measurements for three syllables, [di], [da], and [du], uttered by a female speaker of American English. The measurements show the transitions and steady-state formants for those three syllables. Notice that, in all three syllables, F2 moves away from the same frequency.
traces of movement of the articulators, from the consonant to the vowel.

Figure 6.3 shows the first and second formants (including their transitions) for three syllables, [di], [da], and [du].
Clearly, the formant transitions do not look anything alike, from one to the next syllable – nor do the consonants sound anything alike, if we eliminate the vowel. If we take a recording of [di], [da], and [du] and delete the three vowels (easily done with the help of a computer), we do not hear [d] three times. Instead, we hear little chirps with different tones for each of the three trimmed syllables. So where is the [d]? It is in the hearer's mind, and not in the physical signal. The speech signal containing the three syllables carries information allowing the hearer to reconstruct the three consonants that the speaker articulated, but the physical signals associated with the three syllables beginning with [d] do not contain three identical acoustic events corresponding to the phoneme we hear as [d]. Put another way, the hearer perceives different acoustic events as belonging to the same category.

The phenomenon of **categorical perception** (Liberman et al. 1957) helps explain the powerful effect that knowledge of the phonemic inventory has on speech perception. We will illustrate this phenomenon using the voicing contrast in stop consonants, since this has been extensively studied (and is very well understood). Recall from Chapter 5 that the primary acoustic difference between a voiceless and a voiced stop ([p] versus [b], for example) is **voice onset time (VOT)**. VOT is the time

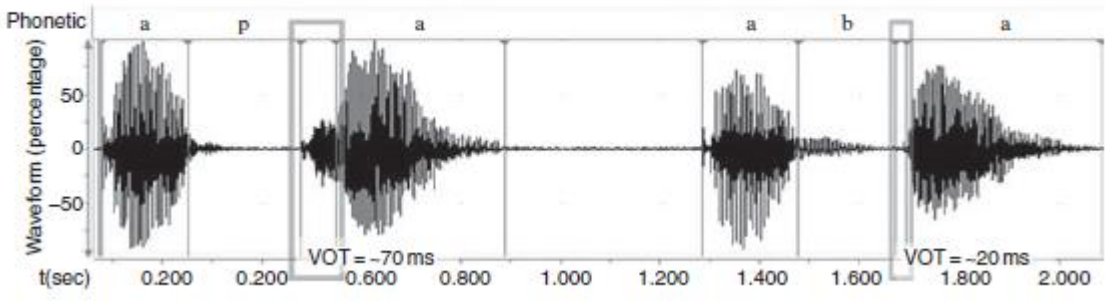


Figure 6.4 Waveform for [apa] and [aba], as produced by a female speaker of American English. The regions between the release of the stop and the onset of voicing are marked by a gray rectangle; approximate measurements for that region (VOT region) are indicated. The visible noise inside the VOT for [p] is aspiration.

that elapses between the release of the closure and the onset of voicing for the following vowel. For [b], voicing begins either the moment the closure is released (for a VOT of 0 milliseconds) or within the first 30 milliseconds after release of the closure. In contrast, [p] has a VOT between 40 and 100 milliseconds. Figure 6.4 illustrates how VOT is measured.
VOT offers an excellent example of variability in the speech signal.
Unlike the phonological feature of voicing, which is binary (a sound is categorized as either voiced or voiceless), VOT is a *continuous variable*:
stops have VOTs that can vary between 0 and 100 milliseconds – there are literally hundreds of different possible VOTs for stops. Yet, in the mind of the average English speaker, stops are either voiced or voiceless.

People perceive the VOT continuum categorically, ignoring differences between sounds drawn from each of the perceived categories. How these categories are set depends crucially on the phonemic inventory of the language; more on this in a moment.

One way to study categorical perception is to synthesize a series of sounds that vary along a continuum of interest (like VOT), play those sounds for people, and ask what they think they have heard. Speech synthesis software can be used to generate a vowel preceded by a stop consonant with a VOT of 0 milliseconds, another with a VOT of 10 milliseconds, and so on. Everything about the signal will be identical, except the VOT. Different techniques exist to assess what participants think they heard. In some experiments, participants hear pairs of sounds and judge whether the pair is the same sound repeated or two different sounds; in other experiments, participants hear three sounds, and judge whether the third sound is the same as the first or the second. Some

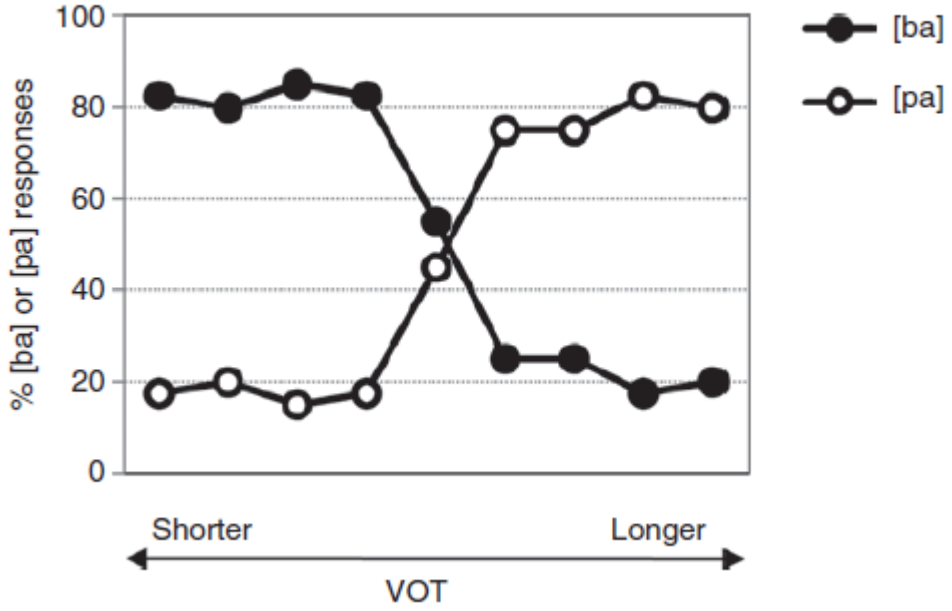


Figure 6.5 Hypothetical results of a categorical perception experiment, for participants listening to nine syllables in a VOT continuum, and asked to indicate whether they have heard [ba] or [pa]. The horizontal *x*-axis plots responses for each of the 9 syllables, varying from short VOT (left) to long VOT (right). The vertical *y*-axis indicates the percent of [ba] or [pa] responses for each signal.

experiments measure judgment responses and reaction times. Other experiments additionally collect information about brain activity while participants are listening to sequences of sounds and making judgments about them, or track participants' eyes as they listen to sequences of sounds and use a mouse to click on a display where they can indicate their choice.

The graph in Figure 6.5 illustrates what the results might look like for a very simple categorical perception experiment. In this hypothetical
experiment, people are asked to listen to a single sound and make a binary choice about what they think they heard. The graph displays the percent of participants who heard [ba], and the percent who heard [pa], for each of 9 signals that varied in VOT in increments of 10 milliseconds. Shorter VOTs are on the left, longer VOTs are on the right.

For the four signals on the left, with shorter VOTs, the signal is heard as [ba] about 80 percent of the time; in contrast, for the four signals on the right, with longer VOTs, the signal is heard as [pa] about 80 percent of the time. Notice that only one of the signals – the fifth signal, in the middle of the chart – is responded to at chance level (both [ba] and [pa] responses are close to 50 percent); this is referred to as the *cross-over*

point in the acoustic continuum.

178 – 179 – 180 – 181- 182 – 183 – 184 - 185 – 186

THE HEARER: SPEECH PERCEPTION AND LEXICAL ACCESS

Our hypothetical experiment illustrates a crucial aspect of categorical perception: physically different acoustic signals are categorized by the perceptual system as belonging to the same phonemic category. The difference in VOT between the first and fourth signals is greater (30 milliseconds) than between the fourth and sixth signals (20 milliseconds), yet the first and fourth signals are perceived as the same sound, while the fourth and sixth are perceived as different sounds.

Does this mean that the speech perception system cannot reliably distinguish between small differences in VOTs? An experiment in which participants' judgments about sameness were timed sheds some light on this question. Tash and Pisoni (1973) report that participants took slightly longer to say that two dissimilar members of the same category were the same (e.g., signals with VOTs of 40 versus 60 milliseconds) than to say that two identical members were the same (e.g., two signals with VOTs of 40 milliseconds, or two with VOTs of 60 milliseconds). The auditory system is evidently sensitive to small differences in VOT, but the speech perception mechanism conflates the acoustically different signals and perceives them all as the same phoneme.

Such decisions made by the speech perception mechanism are guided by knowledge of the language being heard – specifically, knowledge about the phonemes of the language. English makes a two-way phonemic distinction in stop consonants, as illustrated by the minimal pair *bat–pat*. (Recall from Chapter 2 that English has both aspirated and unaspirated voiceless stops in its repertoire of sounds, but [pʰ] and [p] are in complementary distribution: the former occurs initially at the beginning of stressed syllables, like *pat*, the latter in consonant clusters beginning with [s], like *spat*.

Acoustically, an aspirated voiceless stop has a longer VOT than an unaspirated voiceless stop, as shown in Figure 6.4.) Differently from English, Thai makes a three-way distinction in its stop consonant phonemic inventory, and has near-minimal triplets like the following: *bâ*, pronounced [ba], means 'crazy'; pâ [pa] means 'aunt', and *phâa* [pʰa] means 'cloth' (Ladefoged 2005: 138).

Because of these differences in the phonemic inventory for the two languages, English and Thai hearers perceive signals from the same VOT continuum very differently (Lisker and Abramson 1964). Englishspeaking participants divide the VOT continuum into two categories; Thai-speaking participants divide the same VOT continuum into three categories.

The biases that knowledge of a language confer on a hearer – making speech perception effortless, despite the variability in the signal – have consequences in the context of second language acquisition. Simply put, it is unavoidable, especially in early stages of second language acquisition, to listen to a new language with the "ears" of one's first language. As a result, new phonemic contrasts will be difficult to perceive. For example, English speakers have a hard time "hearing" the three-way contrast among Thai stops – particularly the unaspirated–aspirated contrast – even after instruction about it (Curtin, Goad, and Pater 1998). Catherine Best and colleagues propose that discriminating speech sounds in a second language crucially depends on how well they can be perceptually "assimilated" into the existing phonemic categories of the first language (Best, McRoberts, and Goodell 2001). In a similar vein, James Flege has argued that non-native characteristics in the speech of bilinguals are usually linked to interference from prior learning of phonetic categories in the first language (Flege 2003).

VOT is not the only acoustic continuum that is perceived categorically.
Another similar phenomenon is observed with place of articulation. In Chapter 5 (see Figure 5.7), we discussed how F2 transitions differ, based on place of articulation of a stop consonant. The F2 transition of a naturally produced or computer-generated consonant can be manipulated using speech synthesis tools. With place of articulation continua, categorical perception effects are also observed: most signals in the continuum are grouped together into one or another category, and only a few signals (those at the cross-over points) are responded to at chance level. In a later section, we will describe a study that uses stimuli drawn from a place of articulation continuum.

Constructive speech perception and phonological illusions

Another important property of the speech perception system is that it is **constructive**. This means that the speech perception system takes information anywhere it can find it to construct a linguistic percept of the acoustic signal. As mentioned earlier, different phonemes have unique acoustic properties. The hearer also actively uses knowledge of the phonemic inventory, along with internalized information about how speech is produced.

Some interesting facts about the constructive nature of the speech perception system come from the study of phonological illusions, much as the study of optical illusions provides insights about visual perception. One such illusion – the **McGurk effect** (McGurk and MacDonald 1978) – illustrates how visual and auditory information together affect the construction of a phonological percept. If you watch a video of a person mouthing [ga ga ga …], together with the audio track of a person saying [ba ba ba …], you will hear neither [ba] nor [ga] – but [da]. Depending on the combinations used, the visual will override the audio, the audio will override the visual, or – as in our example – the audio and the visual will combine into a new "sound." Since it is a true illusion, you will perceive it the same way even if you know that the audio and the video

do not match. Most stunning about the version of the illusion described here is that if you close your eyes, you will clearly hear [ba], and if you turn down the volume you will clearly "see" [ga], so it is not the case that the individual signals are inadequate.

The McGurk effect is compelling, but it is not really all that surprising.
We all perceive speech better if the speaker is in view. If people are asked to report speech that has been made difficult to understand by embedding it in noise, comprehension is improved if participants can see the speakers (Macleod and Summerfield 1987). Also, the lip-reading abilities of many deaf people are quite remarkable. This is another example of a point made earlier, that knowledge of the way speech is produced is one type of information available to our speech perception system.

Another kind of illusion that illustrates the constructive nature of speech perception, **phoneme restoration**, was discovered by Warren (1970). Warren took a recording of the sentence *The state governors met with their respective legislatures convening in the capital city*, removed the [s] from the word *legislatures*, and replaced it with a recording of a cough of exactly the same duration as the excised [s]. Listening to this sentence, people reported that the [s] was present in the signal, and that the cough was in the background. Moreover, listeners tended to hear the cough either before or after the word *legislatures*, and not in the middle of it. This is a phenomenon known as *perceptual displacement*, which will come up again in Chapter 7. If a stimulus arrives while a perceptual unit is being processed – here, the word *legislatures* – the stimulus will be perceived as occurring either before or after the perceptual unit.

Another example of phoneme restoration involves inserting silence into words. With a recording of a word like *slice*, we can add silence between the [s] and the [l]. When the interval of silence is just the right duration – about 30 or 40 milliseconds – English speakers systematically hear *splice*. (As the interval of silence gets longer, the illusion disappears.) Remember that silence is a key acoustic indicator of the presence of a voiceless stop – any voiceless stop. How come, then, we don't hear *sklice* or *stlice*? Speech perception is constructive; the hearer uses knowledge of language to rule out *stlice* (on phonotactic grounds, since [stl-] is an impossible syllable onset for English) and *sklice* (if not on phonotactic grounds, then surely because [skl-] is such an infrequent onset – occurring in rare and oddly spelled words like *sclerosis* – that it is dispreferred relative to the more frequent [spl-]).

The phoneme restoration illusion is stronger when the sound being replaced and the sound used to fill in the gap are close acoustically (Samuel 1981); for example, replacing an [s] with a cough – a sound with lots of high-frequency noise – is more effective than replacing an [s] with a tone, and it doesn't work if the [s] is replaced by silence. The illusion is also stronger with obstruent consonants than it is with vowels. The effectiveness of phoneme restoration depends as well on whether the word carrying the missing sound is presented in isolation or inside a sentence. The phenomenon of

phoneme restoration demonstrates the perceptual system's ability to "fill in" missing information, while actively trying to recover meaning from an acoustic signal: what we hear is sometimes not what we perceive.

The explanation for the effectiveness of phonological illusions lies in the operation of the lexical retrieval system. It locates words using as much acoustic information as is available. After a word has been retrieved, its full phonological representation is checked against what has been heard. This is called **post-access matching**. If the match is good enough, the word is accepted as correct and the full phonological representation from the lexicon becomes the percept. This process allows even a degraded acoustic signal to provide enough information to allow retrieval to take place; the phonological details are then filled in by the phonological information associated with that lexical item.

Taking this view into account, plenty of acoustic information was available in the above examples for the words *legislatures* and *splice* to be accessed and to survive the post-access check. Thereafter, the [s] in *legi_ latures* and the [p] in *s_lice* were "heard" based on the invariant phonemic information obtained from the lexicon, rather than from the initial acoustic signal. The fact that people can perceive the phonetic structure of nonsense words (e.g., *plice*) demonstrates that speech perception based solely on the acoustic signal is indeed possible, with no assistance from the lexicon (by definition, nonsense words are not stored in the lexicon, so they cannot engage post-access matching). However, the existence of phonological illusions, like phoneme restoration, demonstrates how the perceptual system can cope when it encounters inadequate acoustic information. In fact, all of these illusions demonstrate the constructive nature of all speech perception, not just perception in the laboratory or perception in the absence of adequate acoustic information. Consider the phenomenon of categorical perception. If we did not perceive divergent acoustic signals categorically, we could not communicate by speech. Categorical perception is the speech perception system's way to convert a variable acoustic signal into a phonological representation.

Slips of the ear (Bond 2005) bear some resemblance to phoneme restoration effects. Consider the person who "heard" *She had on a French suit*, from a signal produced by a speaker who intended to say *She had on a trench suit*. Slips of the ear are also called *mondegreens*, after a famous mishearing of a line in a ballad:

(1) They hae slain the Earl Amurray,
And Lady Mondegreen.

(In the original song, the second line is *And laid him on the green*.) An important difference between slips of the ear and phoneme restoration effects is that the former are often the result of inattentiveness to the signal, while the latter can be truly illusory.
Certain types of phoneme restorations are provoked even when the hearer is paying close attention and knows that the signal has been altered. Slips of the ear, in contrast, are

frequently the result of the hearer being distracted. Slips of the ear are more likely when the signal is noisy (which explains why song lyrics are so susceptible to being misheard) or when the signal is ambiguous (e.g., hearing *traitor* instead of *trader*, since the two words are identical when pronounced with a flap between the vowels, or hearing *fine me* instead of *find me*, since the /d/ in *find* is likely to be elided due to coarticulation).

Hearers can be very tolerant of the sometimes rather bizarre meanings that result from slips of the ear (Bond 2005). Consider, for instance, the strange but funny mishearing of a Beatles' song lyric: *the girl with colitis goes by* (the original lyric is *the girl with kaleidoscope eyes*). Bizarre meanings aside, slips of the ear, similarly to slips of the tongue, tend to result in "heard" sentences that conform to the grammatical properties of the language. For example, Bond cites the following mishearing of *A fancy seductive letter*:

(2) A fancy structive letter

The signal very likely contained sufficient information for the hearer to recover a voiced [d] (in the word *seductive*), though the vowel in the first syllable was perhaps reduced enough so as to be inaudible. Yet the hearer did not "hear" *[sd vktiv], a form that violates the phonotactic constraints of English.

Bottom-up and top-down information

An influential concept in psycholinguistics (and in psychology in general) is the distinction between bottom-up and top-down processing.

Psycholinguistic processes are, at their core, information processing routines; we can ask to what extent these processes are triggered automatically based only on the acoustic signal (bottom-up) or are aided by contextual information, either in the communication situation or within the sentence being processed (top-down).

Let us illustrate with an example. Suppose a friend walks up to you and says "Cat food," clearly and distinctly. You will, effortlessly, be able to decode the acoustic signal and retrieve the uttered words from your lexicon. In this situation, **bottom-up information** guides your processing:

details of the acoustic signal help you build a phonological representation.
Once you have retrieved the words, you might think that your friend saying *Cat food* out of the blue is a bit odd – or not.

Consider a different scenario: you and your roommate have a cat, and you are headed to the supermarket. Your roommate hollers from the kitchen (where the dishwasher is running noisily), "Fluffy's bowl is empty! Be sure to buy some cat food!" The acoustic information that reaches your ear is highly degraded; maybe you catch *Fluffy*, *bowl*, *buy*.

You guess that *cat food* is somewhere in the sentence. You have understood this version of *cat food* (which you didn't even really hear) by using **top-down information**. This is information that is not part of the acoustic signal – contextual information that helps you understand what your roommate said absent a clear acoustic signal. In this case, part of the information guiding your processing was carried by the signal – the words you did catch, especially your cat's name. But other information well beyond the signal helped you too: usually you're the one who buys Fluffy's food and your roommate knew that you were going shopping. All of this conspires to allow you to understand *cat food* as a likely candidate for what your roommate might have been saying.

When bottom-up information inadequately specifies a word or phrase, top-down information can allow the hearer to select among a range of possibilities. If bottom-up information is adequate, however, top-down information will not be necessary. Recall from Chapter 3 that people with Broca's aphasia are good at understanding conversational speech but poor at understanding sentences for which they have to do a detailed analysis. The suggestion is that they are using contextual (top-down) information to understand what is said to them.

An experiment by Pollack and Pickett (1964) provides additional evidence for the interaction of top-down and bottom-up information.

Pollack and Pickett asked participants to listen to single words excised from the sentences they had been produced in. Participants did not do very well understanding the words presented in isolation, but when the same words were presented inside their corresponding sentences, participants understood the words without difficulty. Evidently, the words alone provided inadequate information for bottom-up processing to proceed successfully; the surrounding context, though, provided just the right amount of top-down information.

Some experiments have focused on how specific aspects of the context – in this case, semantic information – affect speech perception (Garnes and Bond 1976; Borsky, Tuller, and Shapiro 1998), offering yet another illustration of how bottom-up and top-down processing depend on both the signal and the available context. In the experiment by Garnes and Bond, the investigators created a series of stimuli along a place of articulation continuum ranging from [beɪt] to [deɪt] to [geɪt]. (Remember that the place of articulation continuum involves differences in formant transitions of the stop consonants.) There were "perfect" versions of the stimuli for each of [b], [d], and [g], but there were also versions of the stimuli that were right in between [b] and [d], and [d] and [g] – "indefinite" signals, like the stimulus at the cross-over point in the VOT continuum in Figure 6.5, which is ambiguous between [ba] and [pa].

The stimuli were embedded in sentence contexts like the following:

(3) a. Here's the fishing gear and the …
b. Check the time and the …
c. Paint the fence and the …

Clearly, the most plausible continuation for the first carrier phrase is *bait*, for the second, *date*, and for the third, *gate*. Participants were simply asked to report which word they thought they had heard in the final position of the sentence.

Not surprisingly, when the signal was a "perfect" [beɪt], participants reported hearing *bait*, regardless of sentence context, even though in some cases the resulting sentence was somewhat senseless (e.g., *Check the time and the bait*). The same happened with "perfect" instances of [deɪt] and [ɡeɪt]. However, when the signal was "indefinite" – for instance, a stimulus beginning with a consonant at the cross-over point for [b] and [d], a word acoustically ambiguous between *bait* and *date* – it was reported as the word that fit the context: it was heard as *bait* if carried by the phrase in (3a), but as *date* if in (3b). Experiments like this offer important insights about speech perception: clear and unambiguous segmental information in the signal is decoded precisely as it is presented, with bottom-up processing, even if this leads to implausible meanings; but indefinite or ambiguous information is processed by using whatever contextual information might be available, thus recruiting top-down processing.

Suprasegmental information in the signal

In the preceding sections we have focused on how segments (consonants and vowels) are recovered from the speech signal. We now turn to how suprasegmental information is recovered, and how that contributes to lexical retrieval. (In Chapter 7 we will discuss how suprasegmental information in the signal can influence syntactic parsing.)
Suprasegmental information is signaled in speech with variations in duration, pitch, and amplitude (loudness). Information like this helps the hearer segment the signal into words, and can even affect lexical searches directly.

In English, lexical stress serves to distinguish words from each other (as we discussed in Chapter 2); for example, compare *trusty* and *trustee*.
Not surprisingly, English speakers are attentive to stress patterns during lexical access (Cutler and Clifton 1984). In Mandarin Chinese, tone is lexically specified (another point illustrated in Chapter 2), and so to recover words from the speech signal, Mandarin speakers are attentive not only to segmental information but also to suprasegmental information (Fox and Unkefer 1985).

Suprasegmental information can be used to identify the location of word boundaries also. In languages like English or Dutch, monosyllabic words are durationally very different

than polysyllabic words. For example, the [hæm] in *ham* has longer duration than it does in *hamster*.

An investigation by Salverda, Dahan, and McQueen (2003) demonstrates that this durational information is actively used by the hearer. In this study, Dutch speakers were asked to listen to sentences (for example, the Dutch equivalent of *She thought that that hamster had disappeared*) while looking at displays (for this sentence, a display containing a picture of a ham and a picture of a hamster, along with other distracter pictures). Their task was to look for a picture that matched a word in the sentence they were listening to, and manipulate that picture using a computer mouse. Eye movements were tracked during the entire procedure.

Participants were sensitive to whether the [hæm] they heard was monosyllabic or embedded in the bisyllabic word, even before the rest of the sentence was heard. For example, for a sentence containing *hamster*, there were more looks to the picture of a hamster than to the picture of a ham, very soon after the onset of the ambiguous segments [hæm], and before the disambiguating segments [stə-]; participants evidently used suprasegmental information (syllable duration) to process the lexical material. Salverda and colleagues use evidence like this to argue that durational cues can signal to the hearer that there is no word boundary after [hæm] in a sentence containing *hamster* but that there is a word boundary in a sentence containing *ham*.

The durational variation described for Dutch and English has to do with the basic rhythm of these languages: Dutch and English are stresstimed languages. Recall from Chapter 2 that stress-timed languages emphasize syllables that are stressed (those syllables are longer, louder, and higher in pitch, relative to other syllables). Some of the characteristics of a stress-timed language are that they permit syllables with consonant clusters in coda position and that they reduce vowels in unstressed syllables (Ramus, Nespor, and Mehler 1999). Speakers of stress-timed languages are very sensitive to stress patterns. One such pattern for English is that content words (especially nouns, as described in Chapter 2) tend to start with a stressed syllable. In an experiment asking people to identify words in a stream of speech, a procedure called *word spotting*, English speakers found it more difficult to find a word like *mint* in a sequence with two stressed syllables, like *mintayve* – pronounced [ˈmɪnˈtʰeɪv] – than in a sequence with a stressed syllable followed by an unstressed syllable, like *mintesh* – pronounced [ˈmɪntəʃ] (Cutler and Norris 1988).

Differently from English and Dutch, languages like Spanish, French, or Italian have regular durations, syllable to syllable, regardless of whether the syllables are stressed, and so they are classified as syllabletimed.

Speakers of syllable-timed languages use syllable information in segmenting speech (Cutler et al. 1986). As mentioned in Chapter 2, languages like Japanese are mora-timed, and so speakers of Japanese are sensitive to moras when segmenting the speech signal (McQueen, Otake, and Cutler 2001).

THE HEARER: SPEECH PERCEPTION AND LEXICAL ACCESS 187- 188

The Role of Orthography

As we move into the next major topic for this chapter, lexical retrieval, we should address an important question that has probably crossed your mind: what about reading? People living in literate societies spend much of their time decoding language in written form. How different is decoding words in writing from decoding words in speech? Researchers concerned with how written language is decoded have found that phonology plays a crucial role in decoding words while reading, but so does orthography (Frost and Ziegler 2007). The **orthography** of a language is its writing system, including the characters (graphemes) it uses and the set of conventions for spelling and punctuation.

The basis of reading is the ability to decode individual words; this involves matching each orthographic symbol (each grapheme) with a phoneme. Programs for literacy and reading readiness that focus on training in phoneme-to-grapheme correspondences have been very successful. This fact provides evidence of how closely linked reading is to phonology. The form-priming experiments described later in this chapter offer more evidence of the fact that phonological forms are recovered for words, even when we are reading them. The involvement of phonology in reading has been confirmed even for languages with writing systems that represent morphemes rather than sounds, like Chinese (Perfetti, Liu, and Tan 2005). Thus, retrieving words presented in writing involves reconstructing their phonological representations.

There is also some evidence that people's knowledge of orthography can mediate how they access their lexicon. For example, one study found that speakers of French were less likely to be able to identify the phoneme /p/ in words like *absurd* than *lapsus*, because in the former, pronounced [apsyrd], the /p/ is spelled with the letter *b* (Halle, Chereau, and Seguí 2000). Another study measured how well Hebrew–English bilinguals performed in a phoneme deletion task, involving monosyllabic words that sound exactly alike in the two languages, like [gʌn] (*gun* in English, "garden" in Hebrew) or [bʌt] (*but* in English, "daughter" in Hebrew) (Ben-Dror, Frost, and Bentin 1995).

Importantly, English uses three letters (each corresponding to one of the phonemes) to represent these words, but Hebrew only represents the consonants: גן (*gn*) for "garden" and בת (*bt*) for "daughter".
Participants were asked to listen to the words and delete the first sound; words in each language were presented separately. Native English speakers (for whom Hebrew was a second language) performed very well; however, native Hebrew speakers (for whom English was a second language) frequently committed an interesting error, related to the way Hebrew is written: rather than just deleting the initial consonant, they deleted the initial consonant plus the following vowel. Studies like this show that how one's language is written can affect phonological awareness; indeed, it has been shown that literacy itself has a strong effect on a person's ability to consciously manipulate phonemes (Morais et al. 1979).

As you continue reading this chapter, bear in mind that both orthography and phonology mediate access to the lexicon; the two systems interact bidirectionally (Frost and Ziegler 2007). The sections that follow will describe experiments performed predominantly using written stimuli, but generally assume that a phonological representation is built from those stimuli.

Accessing the Lexicon

The speaker enters the lexicon using information about meaning so she can retrieve the phonological structure of the appropriate words to convey the meaning she is constructing for a sentence. The hearer's (or reader's) task is the opposite. He uses a phonological representation (decoded using information from the acoustic signal) to retrieve information about meaning. The hearer looks for a lexical entry whose phonological representation matches the one he has heard. When there is a match, a word is retrieved, and information about the word's meaning and structural requirements is then available. As pointed out in Chapter 5, the speed of lexical retrieval is remarkable – it takes a mere fraction of a second to find a word in a lexicon consisting of some 80,000 items.

The lexicon is searched by meanings in production and by phonological forms in perception. Evidence about both the process of retrieval and the way the lexicon is organized is provided by studies that examine how lexical access is affected by meaning and form relations among words, as well by variables such as phonotactics, word frequency, and lexical ambiguity.

A technique widely used to investigate lexical access is the **lexical decision task**. Participants are briefly shown a string of letters and asked to push one button if the letters constitute a word in their language, and a different button if they do not. Responses in a lexical decision task tend to be very rapid, ranging between 400 and 600 milliseconds. In a lexical decision experiment, participants will see equal amounts of words and non-words, and within the many words

Table 6.1 Word list for simulated lexical decision task. For each string, write Y if it is a word of English, N if it is not.

CLOCK ☐	DOCTOR ☐	ZNER ☐	FLOOP ☐
SKERN ☐	NURSE ☐	TABLE ☐	FABLE ☐
BANK ☐	TLAT ☐	URN ☐	MROCK ☐
MOTHER ☐	PLIM ☐	HUT ☐	BAT ☐

they will see throughout the experiment, a subset of those is of interest to the investigator: those words contain a contrast being investigated in the experiment.

To simulate how a lexical decision task works, consider the 16 letter strings in Table 6.1, and write Y or N next to each one, to indicate for each whether it is a word of English. Try to write your responses as quickly as possible.

You probably wrote N next to six of the letter strings, and might have even noticed that you responded to three of them very quickly – TLAT, ZNER, and MROCK – and to the other three somewhat more slowly – SKERN, PLIM, and FLOOP. All six strings are non-words in English, but the first three violate the phonotactic constraints of the language. **Impossible non-words**, like TLAT, ZNER, and MROCK, are rejected very rapidly in a lexical decision task. It is as if the lexical retrieval system were carrying out a phonological screening of sorts, not bothering to look in the lexicon when the string is not a possible word in the language. In contrast, **possible non-words**, like SKERN, PLIM, and FLOOP, take longer to reject, as if the retrieval system conducted an exhaustive, ultimately unsuccessful, search for their entries in the lexicon.

Experimental evidence for the distinction in lexical access between possible and impossible non-words is abundant; one interesting example is a brain imaging study that used positron emission tomography (PET) to measure blood flow changes in the brain while people were presented with real words (BOARD), possible non-words (TWEAL), impossible strings of characters (NLPFZ), and strings of letter-like forms – "false fonts" (Petersen et al. 1990). Petersen and colleagues found that the same areas of the brain are activated in response to real words and possible non-words, and that these areas are different from those activated in response to impossible nonwords and "false fonts" strings.

Of the real words in Table 6.1, you probably responded faster to the more frequent ones (like CLOCK and BANK) than to the less frequent ones (like HUT and URN). The **lexical frequency** of a word can be measured by counting how many times a particular word occurs in a large corpus for that language. Lexical frequency is correlated with lexical decision times and with responses to other types of lexical access tasks: more frequent words are responded to faster (Forster and Chambers 1973; Forster 1981). Words that are used often are evidently more available to the lexical retrieval system.

Another property of words that has been used to study lexical retrieval is lexical ambiguity. Lexically ambiguous words are words that have more than one meaning.

Some research has examined whether such words have more than one lexical entry, and whether having more than one lexical entry can lead to retrieval advantages.

Consider the word *bank*, which as a noun can be a money bank, a river bank, or a snow bank; *bank* can also be a verb. Some lexically ambiguous words have multiple meanings that are completely unrelated (e.g., the noun *punch* can refer to a type of drink, or to a blow with the fist, or to a piercing instrument); such ambiguous words are called *homonyms*. Other ambiguous words have meanings that appear to have a systematic relationship to each other (e.g., the noun *eye* refers to an organ used for vision, or to the opening in a needle, or the aperture of a camera); these words are *polysemous*. Rodd, Gaskell, and Marslen-Wilson (2002) compared these two types of ambiguity in a series of lexical decision experiments, and found that ambiguous words with related senses (polysemous words like *eye*) are retrieved faster than ambiguous words with unrelated senses (homonyms like *punch*). Homonyms have multiple meanings that compete against each other, resulting in delayed recognition. In contrast, the semantic relationships between the multiple senses of polysemous words facilitate their retrieval.

One final variable we will discuss affecting lexical access routines is **priming** (Meyer and Schvaneveldt 1971). Priming is actually a very general property of human cognition: a stimulus you just experienced will affect how you respond to a later stimulus – and this associative response is true not just with linguistic stimuli, but with stimuli of any type (pictures, smells, non-linguistic sounds, etc.). In the list in Table 6.1, the words DOCTOR and NURSE are related semantically, and the words TABLE and FABLE related phonologically. Reading the words in each pair consecutively might have influenced how quickly you responded to the second member of the pair.

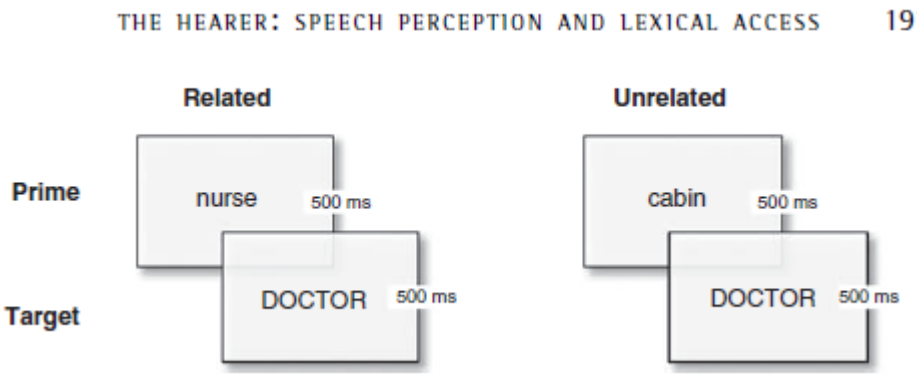


Figure 6.6 Example of two prime–target pairs in a lexical decision experiment.
The primes are in small letters, the targets in capital letters. The figure simulates the display sequence: the prime appears by itself and remains on the screen for a few hundred milliseconds; then the target appears. On the left, the prime and target are

semantically related; on the right, they are unrelated. Notice that the primes, *nurse* and *cabin*, are matched in length (both are five characters long); primes are also usually matched by frequency and other variables.

How does priming work? When you encounter a stimulus of a given type, you activate its mental representation, but as you search for the unique mental representation for the stimulus, you activate associates for that stimulus, as well. Priming, then, is residual activation from previously experienced stimuli.

In a lexical decision experiment concerned with measuring priming effects, a **prime** word is presented for a brief amount of time; it then disappears and a **target** word takes its place; Figure 6.6 illustrates this graphically. (In many priming experiments, primes are presented in small letters, while targets are presented in capitals, and participants are asked to make lexical decisions only on words presented in capital letters.) The experiment includes primes that are related to the target (e.g., for a target like DOCTOR, a related prime would be "nurse"), as well as primes that are unrelated to the target. Responses to the target will be faster when it is preceded by a related than by an unrelated prime.

Many studies have used semantic priming techniques to study to what extent semantic representations are shared between translationequivalent words in bilingual lexicons. This research has confirmed that when a prime and a target are in different languages, a semantic relation between them facilitates retrieval of the target word; for a French–English bilingual, access to *cat* is facilitated by both *dog* and *chien* (Kroll and Sunderman 2003). The strength of the priming can be asymmetric: priming is typically stronger from the bilingual's dominant language (which is usually, though not always, the bilingual's first language) than from the non-dominant language. The idea is that the dominant (or first language) lexicon is bigger, since it was learned first, making the links to (non-linguistic) conceptual representations stronger for words in the dominant language than in the non-dominant language (Kroll and Dijkstra 2002).

Sometimes words that have the same (or very similar) form between two languages are not related semantically at all. A pair of languages can have *interlingual homographs* (words that are written the same way between the two languages), like *coin* in English and *coin* ('corner') in French, as well as *interlingual homophones* (words that sound the same in the two languages), like *aid* in English and *eed* ('oath') in Dutch. Notice that these two examples are pairs of words that are not translation equivalent, but rather interlingual "false friends." False friends have become useful in research that examines to what extent bilinguals are able to inhibit one language while retrieving lexical entries in a unilingual mode. In research like this, bilingual participants perform lexical decisions in one language only, and the experiment compares reaction times to interlingual homographs and frequency-matched controls, for example, on the assumption that false

friends will result in processing cost if the other language is not inhibited. Generally, studies such as this have found that interlingual homographs take as long to process as control words, suggesting that the bilinguals' other language is inhibited during unilingual processing; however, other experiments using priming techniques (discussed in more detail in the following section) have demonstrated that words with orthographic and semantic overlap in the two languages can affect processing time (Dijkstra 2005). One such study (Beauvillain and Grainger 1987) asked participants to make a lexical decision on word pairs consisting of one word in French (like *coin*) followed by a word in English (like *money*). Beauvillain and Grainger found that when the prime and target were semantically related, in English, reaction times on the target word (*money*) were faster, suggesting that even though the prime had been accessed in French, the corresponding English lexical representation had been activated as well.

Types of priming

The examples we have given above are of **semantic** or **associative priming**.
In this type of priming there is a meaning relationship between the prime and target word. Other aspects of words also produce priming

193-194 **THE HEARER: SPEECH PERCEPTION AND LEXICAL ACCESS**

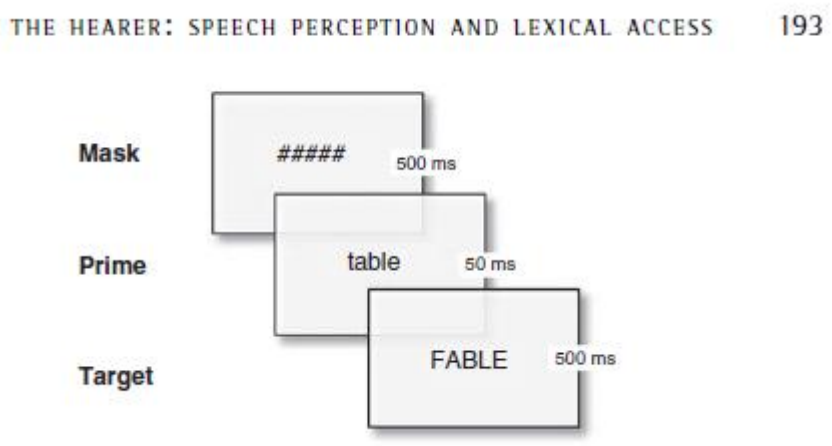


Figure 6.7 Example of a prime–target pair in a masked priming experiment.

effects. There is, for instance, **form priming**, in which the prime and the target are not related semantically, but are related in their phonological form: for instance, *table* will prime *fable*, and *able* will prime *axle*.

An experimental method called **masked priming** (Forster and C. W.
Davis 1984) demonstrates that the prime word may be presented so briefly that it is not consciously processed, but will still result in the priming effect. In this technique, the prime is "sandwiched" between a mask ("#####", for example) and the target, as illustrated in Figure 6.7.
The mask and the target each remain on the screen for 500 milliseconds, typically, but the prime only flashes on the screen for the impossibly brief time of 50 milliseconds. This is not enough time for the word to register consciously – people generally report seeing a flicker between the mask and the target, but they have no memory of having seen an actual word. However, it is apparently enough time for the stimulus to prime the following target. Masked priming technique has been used to study both form priming (C. J. Davis 2003) and semantic priming (Carr and Dagenbach 1990).

Masked priming can be very useful to study the relationship between words in two languages, in bilinguals, because a task involving some aspect of lexical access can be presented to participants as being only in one language, but masked primes can be presented in the other language.

One such experiment, by Sánchez-Casas, C. W. Davis, and García-Albea (1992), examined the relationship between a special type of translation-equivalent words: cognates. **Cognates** are word pairs, like *rico–rich* in Spanish–English, which share not just semantic representations but also have a common stem and are phonologically very similar. In their study, Sánchez-Casas and colleagues presented Spanish–English bilinguals with a task in which targets in Spanish were preceded by masked primes in either Spanish or English. The prime was either identical to the target (*rico–rico*, *pato–pato*), an English cognate of the target (*rich–rico*), an English non-cognate (*duck–pato*), or a control non-word (*rict–rico*, *wuck–pato*). Sánchez-Casas and colleagues found that responses to target words were as fast with cognate primes as with identical primes, and both of these were faster than either non-cognate primes or non-word primes. These findings suggest that cognate words for bilinguals have a special kind of morphological relationship that is represented in the lexicon differently than are words that are translationequivalents (Sánchez-Casas and García-Albea 2005).

Bound morphemes

In Chapter 2 we discussed bound morphemes, which are affixes attached to word stems to form new words. There are *inflectional morphemes*, like the *–s* attached to nouns to make them plural (*car*/*cars*) or the *–ed* attached to verbs to make them past tense (*kiss*/*kissed*). There are also *derivational morphemes*, which can change the meaning of a

word, and sometimes the grammatical class of the word as well. For instance, the suffix –*er* can be attached to a verb, changing it into a noun meaning a person who performs the activity of that verb (*play*/ *player*). An important question about bound morphemes is whether words created by adding morphemes to them are stored separately in the lexicon or whether derived forms are created when they are produced.

With respect to lexical retrieval, if a derived form is stored as a whole, it will be retrieved as a single word. If the derived form is created by adding a morpheme to a stem, the morpheme must be removed before the stem is accessed. This is called **morpheme stripping** (Taft 1981).
There is general agreement among psycholinguists that inflectional morphemes are stripped before their stems are accessed (Marslen- Wilson 2005). Of course, the meaning of the stripped morpheme contributes to the meaning of the sentence that it appears in. Derivational morphemes, however, differ in their productivity. The agentive suffix –*er* can be attached to almost any verb. In contrast, morphemes like –*tion* in words like *derivation* are not only less productive, they also change the pronunciation of the stem. Bradley (1980) reports a series of experiments demonstrating that words derived by affixing productive morphemes are not stored lexically, but are subject to morpheme stripping; in contrast, words containing less productive morphemes are stored in the lexicon, so they do not require morpheme stripping.

THE HEARER: SPEECH PERCEPTION AND LEXICAL ACCESS 195_196 – 197- 198 -199 – 200

The cohort model of lexical access

Models about lexical access help us understand more about the rapid and unconscious retrieval of words from the lexicon. One such model, the **cohort model of lexical access** (Marslen-Wilson and Tyler 1980; Marslen-Wilson 1987) accounts for many facts about lexical retrieval and helps summarize a number of facts related to lexical access described in the preceding sections.

A word's **cohort** consists of all the lexical items that share an initial sequence of phonemes. According to the cohort model, acoustic information is rapidly transformed into phonological information, and lexical entries that match the stimulus phonologically are *activated*. After the first syllable of a word is received, all the lexical entries in its cohort will be activated; after the second syllable is received, a subset of those will remain activated (when an entry ceases to match, it deactivates).
Finally, at some point – before the end of the word, if the target word is unambiguous – a single lexical entry will be uniquely specified, and it will be retrieved. This is called the **recognition point** for the word, and on average it occurs within 200 to 250 milliseconds of the beginning of the word. Of course, if a word is ambiguous and has more than one lexical entry, there will be no recognition point before the end of the word, so all entries that are pronounced the same will be retrieved.

The fact that words can be retrieved before they are completed has been demonstrated by Holcomb and Neville (1991) in an event-related potentials (ERP) experiment. Recall from Chapter 3 that there is a brain response, the N400, associated with the presence of semantic anomaly in a sentence. Holcomb and Neville (1991) showed that the N400 begins long before the entirety of a semantically anomalous word has been heard.

According to the cohort model, an initial cohort of phonologically similar words is activated and, by the word's recognition point, one is selected and integrated into the representation of the sentence being constructed. If this results in a semantic anomaly, given the context, an N400 wave is the neurophysiological result.

The cohort model predicts that the initial part of a word will be more important for lexical access than its end, a prediction that has been confirmed by a number of different kinds of experiments.
Mispronunciations at the beginnings of words are detected more accurately than are mispronunciations at the ends of words (Cole, Jakimik, and Cooper 1978). The phoneme restoration effect is also more robust when the missing phoneme is in the middle or at the end of a word rather than at the beginning (Marslen-Wilson and Welsh 1978). Final consonants are also much more frequently involved in slips of the ear than are initial consonants (Bond 2005).

The cohort model (as well as other similar ones about lexical access) assumes that every word in the lexicon has some resting level of activation.
Stimulation by matching phonological information increases a word's level of activation. When activation reaches some threshold level, the word is retrieved and is then available for use for subsequent processing (be this making a lexical decision, or incorporating the word into an ongoing sentence). The notion of activation helps account for the observed frequency effects in lexical retrieval. High-frequency words have a higher resting level of activation than do low-frequency words.

Since retrieval depends on a lexical item reaching some threshold of activation, high-frequency words will reach that threshold faster than low-frequency words. The phenomenon of priming is also accounted for by the concept of activation. A prime increases the activation of words related by either form or meaning, enhancing their retrieval.

A factor that affects retrieval times for words is **neighborhood density**.
A word's **neighborhood** consists of all the lexical items that are phonologically similar.

Some words have larger cohorts than others:
the word *cot* has many words that are phonologically similar to it, so it is said to come from a dense neighborhood; in contrast, the neighborhood for a word like *crib* is more sparse. Words with larger phonological neighborhoods take longer to retrieve than those

from smaller neighborhoods (Connine 1994). The finding is reasonable: more phonological information is required to specify uniquely a word from a dense neighborhood than from a sparse neighborhood.

Another factor that has been found to affect retrieval is the similarity between the phonological information in the input and the phonological representation of the word in the lexicon. A priming experiment by Connine, Blasko, and Titone (1993) explored this factor, by using nonwords to prime actual words. Connine and colleagues created what they called minimal and maximal non-words, by replacing the initial phoneme of words by a phoneme that was minimally or maximally different from the original. For example, based on *doctor*, *toctor* is a minimal non-word (/t/ and /d/ are both alveolar stops), while *zoctor* is a maximal non-word (/z/ is a fricative, while /d/ is a stop). Both base words (*doctor*) and minimal non-words (*toctor*) facilitated retrieval of semantically related targets (like *nurse*), but maximal non-words (*zoctor*) did not have this priming effect. Recall that when we discussed phonemic restoration, we pointed out that the acoustic representation of the deleted phoneme must be similar to the actual phoneme for restoration to take place. This is because lexical retrieval (and post-access matching) will not be triggered if the acoustic signal is too different from the stored phonological representation of the word. Similarly, minimal non-words will trigger lexical retrieval, but maximal non-words will not.

Interlingual form relationships between words have been found to affect lexical access in bilinguals, evidence that both languages are active at all times. Phonotactic constraints of one language, for example, have been found to affect lexical decisions on non-words in another (Altenberg and Cairns 1983). A study by van Heuven, Dijkstra, and Grainger (1998) examined whether lexical decision times would be affected by neighborhood density, when the neighborhood included words from two languages. The participants were Dutch–English bilinguals, who performed lexical decision tasks on words either from only one language or from both languages. The experiment used words with large neighborhoods in one language but small neighborhoods in the other, as estimated by a corpus analysis; for example, *bird* has more neighbors in Dutch than in English, while *busy* has more neighbors in English than in Dutch. Van Heuven and colleagues found that making a lexical decision about a word in one language was affected by the number of neighbors that word had in the other language: response times to words in one language were systematically slowed down when the number of orthographic neighbors in the other language was large. This effect was absent in a monolingual control group making lexical decisions on English words with large or small neighborhoods in Dutch.

■ Lexical Access in Sentence Comprehension

Because one of the primary interests of this book lies in understanding how lexical retrieval relates to sentence comprehension, it is important to know to what extent these characteristics of lexical items operate as sentences are being processed. A typical approach to this question is to ask whether the presence of a lexical item with a particular

property facilitates or impedes the processing of a sentence. The effects on sentence processing of both word frequency and ambiguity have been studied.

Lexical frequency

Early research on the effects of word frequency in sentence processing used a procedure called the **phoneme monitoring task** (Foss 1969).
Participants listen to a pre-recorded sentence over headphones and are told to push a button when they hear a word beginning with a particular phoneme. The time is measured between the onset of the phoneme in the recording and the moment the participant pushes the button. This reaction time reflects people's ability to perceive and respond to the target phoneme, with an important added feature: the reaction time will vary depending on the cognitive effort involved in processing the sentence at the moment the phoneme was heard. Phoneme monitoring exploits a very general psychological principle known as **resource sharing**. If you are engaged in a complex cognitive activity, your motor responses will be delayed. For instance, if you are doing something difficult like multiplication problems in your head, it will take you slightly longer to push a button in response to a stimulus (like a light or a tone) than it would if you were not doing the multiplication problems.

In one of the experiments reported by Foss (1969), participants were told to monitor for words beginning with [b] (like *bassoon*), while listening to sentences like (4a) or (4b):

(4) a. The traveling bassoon player found himself without funds in a strange town.
b. The itinerant bassoon player found himself without funds in a strange town.

The difference between these sentences is the word preceding *bassoon*:
high frequency in (4a), low frequency in (4b). Foss reports that participants were slower to respond to *bassoon* following the low frequency *itinerant* than the high frequency *traveling*. Low-frequency words increase sentence processing complexity, a finding that fits well both with the lexical decision findings discussed above (more common words are retrieved more rapidly from the mental lexicon) and the observation in Chapter 5 that hesitations are more likely before low-frequency words.

Lexical ambiguity

As described in the preceding section, word frequency has an effect in sentence processing similar to its effect in lexical decision tasks. Now we turn to lexical ambiguity. How ambiguity is dealt with in sentence processing is of central concern in psycholinguistics, because ambiguity is rampant in human language. The majority of the 1,000 most common words in English are multiply ambiguous. Yet, people are rarely aware of making decisions about word meaning, and getting the correct meaning given a specific sentence context tends to be very easy.

The only exception to this is the **garden path sentence**, an example of which is in (5):

(5) The two masked men drew their guns and approached the bank, but the boat was already moving down the river.

Such misleading sentences are called "garden paths" because they lead the hearer "down the garden path," first to an incorrect representation, then to the realization that the sentence makes little sense, finally to a stage of reanalysis which may or may not lead to the correct interpretation.

When you read the sentence in (5), you probably interpreted *bank* as referring to a financial institution. When you got to *the river*, you might have realized that you were wrong about your initial assessment of *bank*, inferring that whoever wrote that sentence had probably meant *river bank*. Thus, you were "led down the garden path." The selection of the incorrect meaning of an ambiguous word can also lead to entertaining results. (In fact, it is the basis of all puns.) Newspaper headlines often contain amusing ambiguities:

(6) New vaccine may contain rabies.
(7) Prostitutes appeal to Pope.

Only in cases like these does one become aware of the presence of an ambiguous word in a sentence being processed, yet every sentence of any length likely has several ambiguities. People usually resolve these ambiguities correctly without creating either a garden path sentence or an amusing one. The existence of garden path sentences like (5) demonstrates that at some point following the ambiguous word, a single meaning has been selected. When and how is that single meaning selected, and why is it so often the correct one?
A phoneme monitoring experiment by Cairns and Kamerman (1976) compared sentences with ambiguous and unambiguous words.

Participants were asked to listen for [d] while listening to recordings of one of the following sentences:

(8) a. Frank took the pipe down from the rack in the store.
b. Frank took the cigar down from the rack in the store.

Both *pipe* and *cigar* are high-frequency words, but only *pipe* is ambiguous.
Cairns and Kamerman report that phoneme monitoring reaction times were longer following the ambiguous *pipe* than following the unambiguous *cigar*, indicating that the ambiguous word required additional processing resources. When processing sentences, all meanings of an ambiguous word need to be considered.

Cairns and Kamerman (1976) included another pair of sentences in their experiment; in these, the target phoneme was located a few syllables down from the ambiguous and unambiguous words:

(9) a. Frank took the pipe from the dollar rack in the store.
b. Frank took the cigar from the dollar rack in the store.

For the pair of sentences in (9), phoneme monitoring times did not differ. The additional complexity produced by the ambiguous word is over just a few syllables later. This suggests that when an ambiguous word is encountered while processing a sentence, all of its meanings are retrieved, but very quickly one of the meanings is selected. On what basis is one meaning selected over the other?

To answer this question, David Swinney developed an account of lexical ambiguity processing, using empirical evidence from a technique called **cross-modal priming**. In a cross-modal priming experiment, participants are asked to make lexical decisions on words presented visually while they are listening to sentences presented auditorily.

Sometimes the word that appears visually is a close associate of a word contained in the sentence presented auditorily. The logic of cross-modal priming is that ambiguous words will prime only those associates of the meaning or meanings that are currently active. For example, suppose you are listening to a sentence like the following:

(10) The man was not surprised when he found several bugs in the corner of his room.

The ambiguous word *bug* can mean either an insect or a covert listening device. If the insect meaning is active, then related words like *ant* should be primed. If the other meaning is active, then related words like *spy* should be primed. In one experiment, Swinney (1979) had participants listen to sentences like (10). At the offset of the word *bugs*, one of three words was displayed for lexical decision: *ant*, *spy*, or *sew* (this third word was unrelated to either of the meanings of *bug*). Participants responded faster to both *ant* and *spy* than to *sew*. This finding confirmed that all meanings of an ambiguous word are initially accessed while a sentence is being processed.

What about context? In sentences like (10), there is no prior context to disambiguate the lexical ambiguity. But consider a sentence like (11):

(11) The man was not surprised when he found several spiders, roaches, and other bugs in the corner of his room.

Will all meanings of *bug* still be retrieved? With sentences such as (11), Swinney (1979) found that both *ant* and *spy* were still primed, when presented at the offset of *bugs*.

A final manipulation in Swinney's (1979) investigation involved presenting the lexical decision targets *ant*, *spy*, and *sew* a few syllables after the offset of *bugs*. When the targets appeared between *the* and *corner*, the contextually related target *ant* was primed, but not the contextually unrelated target *spy* (or the fully unrelated target *sew*). Similar to Cairns and Kamerman's findings (with sentences like (9)), a single meaning for the ambiguous lexical item had been selected only a few words downstream, in this case, the contextually related meaning.

Accessing the lexicon while processing sentences, then, begins with phonological information activating all matching lexical entries, and is followed by selection among those entries of the one that best fits the current sentence. When the context offers a bias for one of the activated entries, the context-appropriate word is selected. When the context does not provide a bias, the most frequent meaning is selected.

Accordingly, the initial retrieval of all possible meanings is exclusively a bottom-up process. Information contained in the phonological representation of the word directs activation of all potential candidates for retrieval: every lexical entry matching that phonological structure is activated. Selection, however, involves top-down processing.

The hearer recruits any and all information available to direct the selection process: the context inside the sentence, context provided by the sentences preceding the current sentence, knowledge about the speaker, real-world knowledge, and so on. Thus, processing lexical ambiguity is another excellent example of the general observation that top-down processes are recruited when bottom-up processes prove to be insufficient.
Bottom-up processes do not uniquely specify a single lexical entry, so top-down processes take over.

■ Summing Up

We have described how hearers use information carried by an acoustic signal to determine the phonological form of an utterance and retrieve lexical items. The phonological representation is constructed from the signal, using multiple sources of information. Evidence that demonstrates how this works includes phonological illusions like the McGurk effect and phoneme restoration.

We also reviewed evidence about how the phonological representation being constructed guides lexical access, activating all potential matches. Research on how words are retrieved offers insights about both how the lexicon is accessed and how words in the lexicon are organized, both phonologically and semantically, with respect to each other, both within a single language (in monolinguals) and between languages (in bilinguals).

Finally, we explored how lexical access works while words are retrieved during sentence comprehension. Low frequency words increase processing cost, because they take longer to retrieve. Ambiguous words increase processing cost, because incorporating a word

into a sentence requires selecting a context-appropriate meaning.

Recovering a phonological representation and lexical retrieval are the two steps in sentence processing that are the precursors to syntactic processing, or parsing, the topic of the next chapter.

New Concepts

bottom-up information
categorical perception
cognates
cohort
cohort model of lexical access
constructive speech perception
cross-modal priming
form priming
garden path sentence
impossible non-words
lexical access
lexical decision task
lexical frequency
masked priming
McGurk effect
morpheme stripping
neighborhood
neighborhood density
orthography
phoneme monitoring task
phoneme restoration
possible non-words
post-access matching
prime
priming
recognition point
resource sharing
semantic (associative) priming
slips of the ear
speech perception
speech signal
  continuous
  parallel transmission
target
top-down information
variability, lack of invariance
voice onset time (VOT)

Study Questions

1. Why is coarticulation so important for speech perception?

2. When comparing the syllables [di], [da], and [du], what is meant by the statement that the initial consonant [d] exists in the speaker/hearer's mind but not in the physical speech signal?

3. What are the sources of variability in speech? How does speech perception overcome acoustic variability to create a mental percept?

4. Explain categorical perception, making reference to Figure 6.5.
   How does the hearer's linguistic competence influence his perceptual categories?

5. What does it mean to say the perceptual system is constructive?
   How do phonological illusions support this claim?

6. What are some ways that speech perception in a second language differs from speech perception in the native language of a monolingual?

7. What are some of the differences between languages in the way that suprasegmental information is used during speech perception?

8. What is the role of phonology during reading? What is the role of orthography? Do these two systems operate independently?

9. What is the difference between bottom-up and top-down processing? When do psycholinguists think that top-down processing is used by the hearer? Is this a conscious decision on the part of the hearer?

10. How does the frequency and ambiguity of lexical items affect subjects' performance on a lexical decision task? Do these variables have the same effect when words are processed in sentences?

11. What are "garden path" sentences? Why are they of interest to psycho linguists?

12. Lexical processing in sentence comprehension involves two operations: retrieval and selection. How do Swinney's crossmodal priming experiments demonstrate these processes with respect to ambiguous lexical items?

7 The Hearer: Structural Processing

In order to understand the message carried by a sentence, the hearer must reconstruct the structural units that convey the intended meaning.

Recall from Chapter 5 that the speaker creates mental representations of those elements: a set of words syntactically related to each other. As we discussed in Chapter 6, the hearer uses knowledge of language and information in the acoustic signal to reconstruct a phonological representation that is then used to retrieve a set of lexical items from the internalized lexicon. Identifying the syntactic relations between the perceived set of words is the essential next step, which eventually leads to recovering the basic meaning the speaker intended.

Reconstructing the structure of a sentence, the focus of this chapter, is a job undertaken by the **structural processor**, or **parser**.

A review of the basic operations of the syntax will assist in understanding the operation of the parser:

- it creates basic structures;
- it combines simple sentences into complex ones; and
- it moves elements of sentences from one structural position to another.

The parser needs to identify the basic components of sentences (elements like subjects and predicates, prepositional phrases, relative clauses, and so on). It can only do this if it is able to dismantle complex sentences into simple clauses. And it must also be able to identify elements that have been moved and link them up with the gaps they left behind in their original structural positions.

In the sections that follow, we explore what psycholinguists have discovered about the way the parser builds structure during sentence processing. We first take on the question of the psychological reality of sentence structure and provide evidence for the claim that the clauses that make up complex sentences are processed as individual units. We then discuss how studying structural ambiguities has shed light on how the parser operates, examining some of the basic strategies the parser follows when building syntactic structure. We then consider the different types of information that the parser can exploit to determine the syntactic relations among words.

The Psychological Reality of Syntactic Structure

Sentence processing involves recovering abstract mental structures based solely on the hearer's knowledge of language, since the signal itself carries no information about syntax. In writing, commas and periods help to indicate when clauses begin and end; in speech, prosody sometimes carries information about certain types of syntactic constituents (we will discuss this later on). But for the most part, syntactic units – from subject NPs to predicate VPs, and everything in between – are not labeled as such in the signal. Yet we think that hearers (and readers) systematically compute syntactic structure while processing sentences. How do we know this is so?

206-THE HEARER: STRUCTURAL PROCESSING

Early experiments studying sentence comprehension measured how processing sentences affected performance in other cognitive tasks, like memory and perception. In such experiments people would be asked to memorize lists of words, or to listen to lists of words presented in noise; the investigators would measure to what extent performance was impaired under different conditions. One experiment (Miller and Selfridge 1950) compared how well people memorized word lists like the following:

(1) a. hammer neatly unearned ill-treat earldom turkey that valve outpost broaden isolation solemnity lurk far-sighted Britain latitude task pub excessively chafe

competence doubtless tether backward query exponent prose resourcefulness intermittently auburn Hawaii uninhabit topsail nestle raisin liner communist Canada debauchery engulf appraise mirage loop referendum dowager absolutely towering aqueous lunatic problem

b. the old professor's seventieth birthday was made a great occasion for public honors and a gathering of his disciples and former pupils from all over Europe thereafter he lectured publicly less and less often and for ten years received a few of his students at his house near the university

The systematic finding in experiments like this was that unstructured sets of words, like the 50 words in (1a), were much harder to recall than structured sets, like the 50 words in (1b) (Miller and Selfridge 1950). A greater percentage of words was recalled from the structured than from the unstructured sets of words. A straightforward explanation of this effect proposes that syntactic structure is psychologically real. Recalling strings of words is easier if the words are related to each other syntactically.

The psychological reality of sentence structure is pervasive and profound, even though syntactic structure itself is abstract and not as consciously available as words are.

An alternative hypothesis is that recall in experiments like the one just described is facilitated, not by syntactic structure, but by the semantic relations among words: after all, the passage in (1b) means something, while the one in (1a) does not. Can people compute syntactic relations in the absence of meaning? Consider the opening verse of the poem "Jabberwocky," written by Lewis Carroll in 1872:

(2) 'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

THE HEARER: STRUCTURAL PROCESSING 207 – 208

When you read or hear Jabberwocky language – language consisting of pseudowords placed in grammatical syntactic frames – you cannot help but compute the syntactic relations, even though you may have no idea what the words actually mean. You (tacitly) know that *toves* is the head noun of the subject NP in the first clause, that *gyre* and *gimble* are verbs, and that *in the wabe* is a locative PP indicating where the toves gyred and gimbled. This has been demonstrated by a number of investigations of how people process Jabberwocky language. One experiment used functional magnetic resonance imaging (fMRI) to examine brain activity in people listening to speech input with or without meaning, with or without syntax (Friederici, Meyer, and Cramon 2000) – sentences like the following:

(3) a. The hungry cat chased the fast mouse.

b. The mumphy folofel fonged the apole trecon.
c. The cook silent cat velocity yet honor.
d. The norp burch orlont kinker deftey glaunch legery.

(The materials actually used were in German; these examples are the English translations provided by Friederici et al.) The study found that certain areas of the brain, in the left inferior frontal cortex, are exclusively recruited when processing input that contains syntactic relations – sentences like (3a) and (3b) – compared to simple word lists – like (3c) and (3d).

This and many other experiments examining brain activity during sentence processing indicate that syntactic processing has not only psychological reality but also specific physiological correlates. Investigations examining event-related potentials (ERP) have discovered components specifically related to processing syntax, some of which we discussed briefly in Chapter 3. Two such components are the very early left anterior negativity (ELAN) and the left anterior negativity (LAN). In both of these components, there is increased negativity with syntactically anomalous sentences. The ELAN is very early (around 150–200 milliseconds after the onset of the anomaly), and is a response to syntactic structure that cannot be computed,

like the example in (4a), compared to (4b) (Neville et al. 1991):

(4) a. *Max's of proof.
b. Max's proof.

The ELAN response is obtained both with regular sentences and sentences with Jabberwocky words (Hahne and Jescheniak 2001). The ELAN, thus, is the brain's response to **word category errors**, that is, when the category of a new word does not fit into the current structure being built by the parser.

The brain responds slightly differently to **morphosyntactic violations**:

(5) a. *The elected official hope to succeed.
b. The elected official hopes to succeed.

Subject–verb agreement violations, like the one in (5a) compared to (5b), elicit a LAN, involving negativity around 300–500 milliseconds after the onset of the anomaly (Osterhout and Mobley 1995).

Ungrammaticality, like word category errors and morphosyntactic violations, also elicits a P600 – an ERP component involving positivity at around 600 milliseconds (Osterhout and Holcomb 1993). We will see later in this chapter that the P600 is also a characteristic brain response to garden path sentences (introduced in Chapter 6), which are grammatical

but hard to process for structural reasons. All of these ERP components are different from the N400 component, which is elicited by semantic anomalies. That the brain should have such specific responses to different types of syntactic anomalies, which in turn differ from responses to semantic anomalies, is strong evidence of the psychological reality of syntactic structure building during sentence comprehension.

The clause as a processing unit

Recall from Chapter 2 that a clause consists of a verb and its arguments. (In the tree-diagramming notation introduced in Chapter 2, a clause is an S-node.) A given sentence can include an independent clause and one or more dependent clauses. Each clause corresponds to an integrated representation of meaning and an integrated representation of structure, so clauses are reasonable candidates for processing units. Clauses correspond to manageable units for storage in working memory during processing. In Chapter 5, we described research in sentence production suggesting that clause-sized units are used in planning. It is not surprising that clauses – units containing a verb plus its arguments – also play a role in perceptual processing.

Decades ago *click displacement studies* confirmed the idea that clauses constitute processing units (Fodor and Bever 1965; Garrett, Bever, and Fodor 1966). These studies worked on the principle of perceptual displacement that was briefly mentioned in Chapter 6.