Chapter 1
The Speaker as Information Processor

Speaking is one of man's most complex skills. It is a skill which is unique to our species. Each normal child starts acquiring it in infancy, clearly driven by a genetically given propensity for language. The mature skill takes all of childhood to develop. It requires extensive interaction between the child and its parents, peers, teachers and other members of the language community. There is, in fact, never a steady state. The mature language user keeps expanding his lexicon as new words are needed or arise in the language. There is also often a continuing growth of rhetorical and narrative abilities in the adult speaker.

The present book is about the organization of this skill. It will consider the speaker as a highly complex information processor who can, in some still rather mysterious way, transform intentions, thoughts, feelings into fluently articulated speech. The dissection of this skill is a scientific endeavor in its own right. It is, in particular, not enough to study the functions of speaking—the kinds of intentional acts a language user can perform through speech, such as referring, requesting, and explaining. Nor is it enough to study the patterns of spoken interaction between interlocutors—the ways they engage in conversation, take turns, signal misunderstanding, and so forth. These are, it is true, of crucial importance for the understanding of speakers as interlocutors. Indeed, these perspectives cannot be ignored with impunity when the skill of speaking is dissected. But they do not suffice. Developing a theory of any complex cognitive skill requires a reasoned dissection of the system into subsystems, or processing components. It also requires a characterization of the representations that are computed by these processors and of the manner in which they are computed, as well as specification of how these components cooperate in generating their joint end product. A theory of speaking will involve various such processing components, and the present chapter

will make a first go at partitioning the processing system that underlies the generation of speech.

By way of introduction, I will present a case study of a speaker's generation of a single utterance. This case study is phenomenological in nature, but it is not theory-free. Its purpose is to set the scene for conjecturing an architecture for the processing system that underlies speech production. Such an architecture will be proposed in section 1.2. It consists of various processing components which, together, translate the speaker's intentions into overt speech.

The nature of these processors is discussed further in sections 1.3 and 1.4. It will, in particular, be stressed that processing components are specialized and that they do their work in rather autonomous fashion. Most of the components underlying the production of speech, I will argue, function in a highly automatic, reflex-like way. This automaticity makes it possible for them to work in parallel, which is a main condition for the generation of uninterrupted fluent speech. The special way in which this cooperation between components is organized so as to result in "incremental production" is the subject of section 1.5.

The rest of the book is straightforward in structure. It will basically follow the components of the proposed architecture one by one, from the speaker's initial conception of something to express to his eventual articulation of an appropriate utterance. However, before venturing upon that voyage, I devote a second introductory chapter to the speaker as interlocutor. Many aspects of a speaker's information processing cannot be correctly evaluated if we lose sight of the canonical ecological context of talking: the speaker's participation in conversation.

1.1 A Case Study

The case to be analyzed is taken from page 868 of Svartvik and Quirk 1980. It appears in a tape-recorded exchange between two male academics, aged about 40, and a male about 18 years old who is applying for admission to college. The academics are apparently testing the student's knowledge of Shakespeare, and the following pair of turns emerges:

Academic 1: [e:m] ... would you say Othello was [e:] ... a tragedy of circumstance ... or a tragedy of character.

(lapse)

Student: I I don't know the way ... play WELL enough sir.

The target of analysis here will be the student's utterance. The academic's

utterance invited the student to provide certain information about the play *Othello*, presumably not because the academic lacked that knowledge but rather because he wanted to find out more about the student's informedness. And this, one might assume, was mutually known between the academics and the student. All three parties knew and accepted that the conversation was an interview, and that defined their roles.

The student started his utterance after a lapse, a long silence. Since academic 1 had addressed the question to the student (*would you say ...*), the situation obliged the cooperative student to take the floor. Hence, the lapse could not have been due to the student's expecting somebody else to take the floor. The student was, rather, involved in serious information processing. Of what sort? Was he retrieving whatever he knew about the play in order to infer a probable answer? This would mean that the student had conceived of the intention to assert the requested information, and that he was now engaged in inferring it. There is evidence in the interview that this was not what was going on.

The student was probably aware, but academic 1 apparently was not, that academic 2 had asked almost the same question five or ten minutes earlier (*Would you call Othello a tragedy of circumstance or of character?*) and that the student had then expressed his ignorance (*I don't know much about Othello, so I couldn't say*). It may or may not have been the case, moreover, that academic 2's subsequent turn in that sequence (*Well which others would you characterize as tragedies of circumstance?*) had given away the answer to the student. Although the student may have tried to remember that earlier discussion in order to come up with the correct answer, it is more likely that he was embarrassed by this repeated question and that he considered another move (namely, reminding academic 1 that academic 2 had preempted him on this issue, or some similar speech act). Under this interpretation, the lapse resulted from a conflict of intentions: What move should be made? The student's final decision was apparently to let politeness prevail, and to avoid embarrassing academic 1 by suggesting that he hadn't been very attentive. The student would, instead, express his ignorance again.

So far, the analysis suggests that, in planning an utterance, there is an initial phase in which the speaker decides on a purpose for his next move. This decision will depend on a variety of factors, and not in the last place on the speaker's needs, beliefs, and obligations. The speaker's choice of purpose relates in particular to what has been said before in the conversation, of which he must have kept some record. In the present example, the student took into account the previous turn (i.e., the academic's question,

the topic of the discourse—Shakespeare's plays) and, presumably, the earlier part of the discourse concerning *Othello*. This first step in planning an utterance is the conception of a communicative intention. In view of this end, appropriate means will have to be marshaled.

Let us return to the student's utterance. Having decided to politely reveal his ignorance, the student had to decide on the information he would have to express in order to convey that intention. The academic left the student with two alternatives: saying that *Othello* is a tragedy of circumstance and saying that it is a tragedy of character. Strictly speaking, the question left no other option open for the student. In particular, the interviewer did not explicitly allow for the possibility that the student did not know the answer. In that case, the question should have been phrased like this: *Do you know whether Othello was a tragedy of circumstance or a tragedy of character?* Neither of the two options given could be chosen to express the intention. What would have conveyed the intention would have been for the student to tell the academic straightaway that he couldn't give the answer. Because of the interview character of the conversation, that condition was on everybody's mind in any case. Still, the information the student selected for expression was slightly different. The student expressed less information than was required, because he did not say *I cannot answer your question*; at the same time, he expressed more than was required by saying that he didn't know the play well enough. Why did the student select the latter information as a means of conveying his intention?

There may have been two reasons. First, given the decision to answer politely, the student may have rejected the option of directly expressing information that would presuppose a third option, one not overtly given by the academic. It is, after all, slightly impolite for a questioner to ignore the listener's potential ignorance, and it would be equally impolite for the answerer to implicate that there had been a flaw in politeness. The student, rather, left it to academic 1 to *infer* his inability to answer the question (*well enough* for what?). That was the main implication of the information expressed, and the issue of impoliteness thus faded into the background. Second, the student may have wanted to reveal something else at the same time: that he did know *Othello*, contrary to what academic 1 might have inferred from a straight "I don't know" answer.

The content selected for expression was not an atom but a structured concept. It consisted of an experiencer ("me"), of whom it is predicated that his state of knowledge about subject matter X doesn't meet criterion Y, where X is Shakespeare's play *Othello* and Y is "sufficient for inferring the type of tragedy". This selection also reflected the speaker's decision not to

spell out criterion Y, so that the inference could be left to the interviewer. In addition, there was the decision to use a polite addressing form (which surfaced as *sir*).

The speaker's elaboration of a communicative intention by selecting the information whose expression may realize the communicative goals will be called *macroplanning* in this book.

In the example above, there were also other decisions taken with respect to the information to be expressed. Among them were (i) to refer to *Othello* in reduced but definite form because that referent had been introduced explicitly in the previous turn (surfacing as *play*), (ii) to acknowledge that *Othello* and the student's knowledge thereof was the topic the answer had to be about (resulting in sentence-initial placement), and (iii) to focus on the degree of knowledge of the play as the new information (surfacing as sentence-final and receiving tonic stress, *WELL enough*). All these decisions related in some way or another to the state of the student's record of the discourse so far. They determined the informational perspective of the utterance, its topic, its focus, and the way in which it would attract the addressee's attention. Conceptual planning activities of this kind—i.e., planning an informational perspective for an utterance—will be called *microplanning*.

So far, we have seen reasons to distinguish two phases in the planning of an utterance after a communicative intention has been conceived. During macroplanning the speaker selects and molds information in such a way that its expression will be an appropriate means for conveying the intention. In this phase the speaker spells out his communicative intention and marshals the appropriate information whose expression will reveal the intention to the addressee. This fixes the "speech act," i.e., the commitments the speaker is prepared to make by expressing a particular informational content as well as the chosen levels of directness and politeness. These bits of information are not independent. In the example, the degree of directness appeared to affect the content to be expressed. During the second phrase—microplanning—the speaker brings all this information into perspective, marking the information status of referents as "given" or "new" for the addressee, assigning topic and focus, and so on.

The student had to cast this highly structured package of information (which will be called the *message*) in an utterance of some sort—a phrase, or a rather elliptical sentence. He began with *I*, and there was still some hesitation. There may not have been a final decision on the information to be expressed—we will never know precisely—but the long silence made it important to do something. At any rate, *I* appeared again. It is the deictic

term referring to the experiencer "me" in the conceptual structure to be expressed. It is, moreover, in the nominative case (not *my*, *mine*, or *me*), which indicates that the speaker had selected it as the grammatical subject of the sentence. This choice does justice to treating the experiencer as the given topic of discourse, a reflection of the academic's *you*. It is the one about whom the comment is to be made. The choice also clearly restricts what the speaker can do next: He must select a verb that allows *I* to be its grammatical subject. If indeed the speaker had started to say *I* out of urgency, and before the necessary information had been made available, this restriction may explain the hesitation on *I*. The speaker selected as the main verb *know*, which does express the concept of "state of knowledge". Ignoring the *don't* for the moment, observe that the student realized the substance of that state of knowledge—Shakespeare's play *Othello*—as the grammatical object of *know*. In fact, he mapped that concept, to be expressed in reduced form, onto the noun *play*. The final part of the conceptualization, "not meeting criterion *Y*", and eliding *Y*, was mapped on an adverbial phrase: (not) *WELL enough*. To complete his utterance, the student accessed a polite address form for a male addressee: the conventional *sir*.

The way in which a speaker maps the package of information to be expressed onto spoken words involves, of course, the retrieval of lexical items from what I will call the *mental lexicon*—the store of information about the words in one's language. The speaker will use parts of the conceptual structure to retrieve the appropriate words (i.e., the lexical items that correctly express the intended meanings) from the lexicon. A lexical item is a complex entity. It is retrieved on the basis of its meaning, but in addition it contains syntactic, morphological, and phonological information.

There is evidence, to be discussed in chapters 6 and 7, that speakers construct the "framework" of an utterance without much regard for the phonology of words. Apart from the semantic information, they use the syntactic information (and sometimes aspects of the morphological information) contained in the retrieved items to build this framework. This nonphonological part of an item's lexical information will be called the item's *lemma information* (or, for short, the *lemma*). So, when we say that a speaker has retrieved a lemma, we mean that the speaker has acquired access to those aspects of a word's stored information that are relevant for the construction of the word's syntactic environment. Take, for instance, the word *know*, which our speaker used in his utterance. The lemma *know*

requires a subject that expresses the role of experiencer, and an object (or a complement) that expresses what is known, and there is a certain order in which these grammatical elements should appear. By some process, (which we will call *grammatical encoding*), the speaker retrieves the appropriate lemmas for the concepts to be expressed and puts the lemmas in the right order. It is presumably as part of this process that the negative element (in "not meeting criterion Y") is mapped onto an auxiliary verb, which eventually yields *don't*. In addition, certain features are assigned to lemmas during grammatical encoding, such as that they are definite (for *play*), that they should receive pitch accent (as for *WELL*), or that they should have a certain case (e.g., nominative for *I*). This initial move in mapping the information to be expressed onto words creates what will be called a *surface structure*.

But how then could the speech error *way* appear? In order for the lemma *way* to become active, the speaker should have been thinking of its meaning. Maybe the speaker thought of something like "I don't know the way". If the above phenomenology is correct, however, the meaning of *way* was not part of the message, and its lemma therefore did not appear in the surface structure. The error presumably arose when the phonological forms of the words were accessed. It is not far-fetched to suppose that *way* is the result of blending the sound realizations of *WELL* and *play*. At the critical moment in time, the student had both lemmas available in his surface structure, and a slight mistiming in the activation of their phonological patterns created the blend. Note that *way* was not accented; rather, it carried the level prosody intended for *play*, not the raised pitch that *WELL* should receive. Such speech errors are an important argument for distinguishing an independent level of *phonological encoding*. After retrieving the phonological forms for the lemmas in the surface structure, the speaker can build a *phonetic* or *articulatory plan* for the utterance.

The transcription of the above conversation doesn't tell us how the utterance really sounded. It will have been delivered with some specific pitch and loudness contour, it will have displayed the student's characteristic timbre, and there will have been some degree of blending or "coarticulation" between successive speech sounds. All these and many other features of the utterance are aspects of the speaker's articulation—the execution of the phonetic plan by the delicately tuned musculature of the vocal apparatus.

It is, finally, not trivial that the student *noticed* that he had said *way* instead of *play*. In fact, he noticed it right after he said it. He stopped the

flow of speech, there was a short moment of silence, and he replaced the error by an edited version. How did the student know that something had gone wrong? Had he listened to himself speaking and noticed that *way* was not what he had intended to say? Or would he have discovered the error even without listening to himself? And why did he replace *way* with *play*, and not with *the play* or *know the play*? There is, apparently, some way for the speaker to monitor his own speech and to adapt things correspondingly. In conversation, moreover, interlocutors send various signals to the speaker which tell him that something wasn't clear (*eh?*), or that he should go on (*mhm*), or that one waits for him to take the turn, and so on. Much of this can be done by gaze or gesture. A speaker, while delivering his utterance, is continuously monitoring himself and his interlocutors, and this feeds back to what he is doing.

The student's utterance may not have helped him much in the interview, but it has been most helpful for us in distinguishing various steps in a speaker's production of an utterance. There is the initial choice of purpose ("conceiving the intention") and there is selection of the means to make this intention apparent to the interlocutor. These conceptual processes depend on the speaker's state of motivation, the knowledge shared with the interlocutors, and especially the speaker's discourse record. They create a "message" to be expressed. Furthermore, there are more specifically linguistic steps to be taken. Words have to be accessed. Syntactic forms that map the concepts and their relations onto a grammatical surface structure have to be constructed. These surface structures, in turn, have to be developed into phonetic plans that serve to instruct the articulatory apparatus of the speaker. On top of all this, the speaker apparently manages to monitor and, where necessary, improve what he is doing.

In the next section a framework will be proposed in which these processing notions are brought together.

1.2 A Blueprint for the Speaker

Figure 1.1 proposes a partitioning of the various processes involved in the generation of fluent speech. It consists of a number of processing components, each of which receives a certain kind of input and produces a certain kind of output. The output of one component may become the input for another. In the subsequent sections some preliminary motivation will be given for proposing the flow of information depicted in the figure, but first the different processing components will have to be introduced.

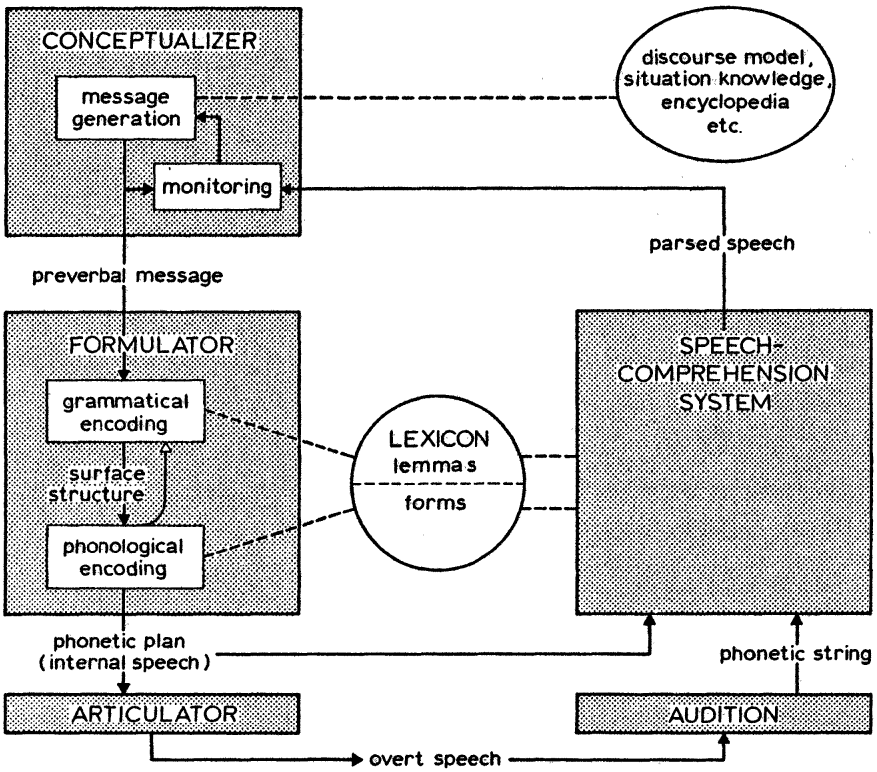

Figure 1.1
A blueprint for the speaker. Boxes represent processing components; circle and ellipse represent knowledge stores.

1.2.1. Conceptualizing

Talking as an intentional activity involves conceiving of an intention, selecting the relevant information to be expressed for the realization of this purpose, ordering this information for expression, keeping track of what was said before, and so on. These activities require the speaker's constant attention. The speaker will, moreover, attend to his own productions, monitoring what he is saying and how (see subsection 1.2.4). The sum total of these mental activities will be called *conceptualizing*, and the subserving processing system will on occasion be called the *Conceptualizer* (in full awareness that this is a reification in need of further explanation—we are, of course, dealing with a highly open-ended system involving quite heterogeneous aspects of the speaker as an acting person). The product of conceptualizing will be called the *preverbal message*.

In order to encode a message, the speaker must have access to two kinds of knowledge.

The first kind is *procedural* knowledge; it has the format IF X THEN Y. For instance:

IF the intention is to commit oneself to the truth of *p*, THEN assert *p*. Here *p* is some proposition the speaker wants to express as being the case, and the indicated procedure is to build an assertion of that proposition. The Conceptualizer and its message generator can be thought of as a structured system of such condition/action pairs (to which we will return in section 1.3). These procedures can deposit their results in what is called *Working Memory* (Baddeley 1986). Working Memory contains all the information currently accessible to the speaker, i.e., all the information that can be processed by message-generating procedures or by monitoring procedures. It is the information *attended to* by the speaker.

The second kind of knowledge is *declarative* knowledge. A major kind of declarative knowledge is *propositional* knowledge. The variable *p* above could, for instance, be given the value

"Manhattan is dangerous".

This is a unit of propositional knowledge. The speaker has access to a huge amount of declarative knowledge. That knowledge is, in the first place, available in *Long-Term Memory*—the speaker's structured knowledge of the world and himself, built up in the course of a lifetime (and also called *encyclopedic knowledge*). But there is also declarative knowledge of the present discourse situation. The speaker can be aware of the interlocutors—where they are and who they are. The speaker, moreover, may be in the perceptual presence of a visual array of objects, of acoustic information about the environment, and so forth. This *situational knowledge* may also be accessible as declarative knowledge, to be used in the encoding of messages. Finally, the speaker will keep track of what he and the others have said in the course of the interaction. This is his *discourse record*, of which only a small, focused part is in the speaker's Working Memory. Figure 1.1 represents declarative knowledge within circles. Procedural knowledge is not represented independently in the figure; it is part of the processors themselves, which are given rectangular shape.

When the speaker applies the above IF X THEN Y procedure to the proposition "Manhattan is dangerous", the message will be the assertion of this proposition. The message generated is not only the output of the Conceptualizer; it is also the input to the next processing component, which will be called the *Formulator*. As we will see in subsection 4.4.5, the Formulator can handle only those messages that fulfill certain language-specific conditions. Hence, the adequate output of the Conceptualizer will be called a *preverbal* message. It is a conceptual structure that can be accepted as input by the Formulator.

We have already distinguished two stages in the planning of a preverbal message: macroplanning and microplanning. Macroplanning involves the elaboration of some communicative goal into a series of subgoals, and the retrieval of the information to be expressed in order to realize each of these subgoals. Microplanning assigns the right propositional shape to each of these "chunks" of information, as well as the informational perspective (the particular topic and focus) that will guide the addressee's allocation of attention.

1.2.2 Formulating: Grammatical and Phonological Encoding

The formulating component, or *Formulator*, accepts fragments of messages as characteristic input and produces as output a *phonetic* or *articulatory plan*. In other words, the Formulator translates a conceptual structure into a linguistic structure. This translation proceeds in two steps.

First, there is *grammatical encoding* of the message. The Grammatical Encoder consists of procedures for accessing lemmas, and of syntactic building procedures. The speaker's lemma information is declarative knowledge, which is stored in his mental lexicon. A lexical item's lemma information contains the lexical item's *meaning* or *sense*, i.e., the concept that goes with the word. Two examples of such information are that *sparrow* is a special kind of bird and that *give* involves some actor X causing some possession Y to go from actor X to recipient Z. Also, the *syntax* of each word is part of its lemma information. The lemma *sparrow* is categorized as a count noun; the verb *give* is categorized as a verb (V) which can take a subject expressing the actor X, a direct object expressing the possession Y, and an indirect object expressing the recipient Z (as in *John gave Mary the book*); and so forth. A lemma will be activated when its meaning matches part of the preverbal message. This will make its syntax available, which in turn will call or activate certain syntactic building procedures. When, for instance, the lemma *give* is activated by the conceptual structure of the message, the syntactic category V will call the verb-phrase-building procedure. This procedural knowledge (stored in the Grammatical Encoder) is used to build verb phrases, such as *gave Mary the book*. There are also procedures in the Grammatical Encoder for building noun phrases (e.g. *the sparrow*), prepositional phrases, clauses, and so on.

When all the relevant lemmas have been accessed and all the syntactic building procedures have done their work, the Grammatical Encoder has produced a *surface structure*—an ordered string of lemmas grouped in phrases and subphrases of various kinds. The surface string *John gave Mary the book* is of the type "sentence," with the constituents *John* (a noun

phrase which is the sentence's subject) and *gave Mary the book* (a verb phrase which is its predicate). The verb phrase, in turn, consists of a main verb and two noun phrases: the indirect object and the direct object. The grammatical encoding procedures can deposit their interim results in a buffer, which we will call the *Syntactic Buffer*.

Second, there is *phonological encoding*. Its function is to retrieve or build a phonetic or articulatory plan for each lemma and for the utterance as a whole. The major source of information to be accessed by the Phonological Encoder is *lexical form*, the lexicon's information about an item's internal composition. Apart from the lemma information, an item in the lexicon contains information about its morphology and its phonology—for instance, that *dangerous* consists of a root (*danger*) and a suffix (*ous*), that it contains three syllables of which the first one has the accent, and that its first segment is /d/. Several phonological procedures will modify, or further specify, the form information that is retrieved. For instance, in the encoding of *John gave Mary the book*, the syllable /buk/ will be given additional stress.

The result of phonological encoding is a *phonetic* or *articulatory plan*. It is not yet overt speech; it is an internal representation of how the planned utterance should be articulated—a program for articulation. Not without hesitation, I will alternatively call this representation *internal speech*. The term may, of course, lose some of its everyday connotation when used as an equivalent for the technical term "phonetic plan." In particular, the speaker will, in the course of fluent speech, often not be aware of his phonetic plan. The term "internal speech," however, entails a certain degree of consciousness (McNeill 1987). A more precise way to put things would be to say that internal speech is the phonetic plan as far as it is attended to and interpreted by the speaker—i.e., the phonetic plan as far as it is *parsed* by the speaker (see below). I will ignore this fine distinction where it is without consequence. This end product of the Formulator becomes the input to the next processing component: the Articulator.

1.2.3 Articulating

Articulating is the execution of the phonetic plan by the musculature of the respiratory, the laryngeal, and the supralaryngeal systems. It is not obvious that the Formulator delivers its phonetic plan at just the normal rate of articulation. In fact, the generation of internal speech may be somewhat ahead of articulatory execution. In order to cope with such asynchronies, it is necessary that the phonetic plan can be temporarily stored. This storage device is called the *Articulatory Buffer*. The Articulator retrieves successive

chunks of internal speech from this buffer and unfolds them for execution. Motor execution involves the coordinated use of sets of muscles. If certain muscles in a set are hampered in their movement, for instance when the speaker chats with a pipe in his mouth, others will compensate so that roughly the same articulatory goal is reached. In other words, though the articulatory plan is relatively independent of context, its execution will, within limits, adapt to the varying circumstances of articulation. The product of articulation is *overt speech.*

1.2.4 Self-Monitoring

Self-monitoring involves various components that need no detailed treatment in a book on language production since they are the processing components of normal language comprehension. A speaker is his own listener. More precisely, a speaker has access to both his internal speech and his overt speech. He can listen to his own *overt* speech, just as he can listen to the speech of his interlocutors. This involves an *Audition* processing component. He can understand what he is saying, i.e., interpret his own speech sounds as meaningful words and sentences. This processing takes place by means of what is called the *Speech-Comprehension System* in figure 1.1. It consists, of course, of various subcomponents, which are not at issue here and hence not indicated in the figure. The system has access to both the form information and the lemma information in the lexicon, in order to recognize words and to retrieve their meanings. Its output is *parsed speech,* a representation of the input speech in terms of its phonological, morphological, syntactic, and semantic composition.

The speaker can also attend to his own *internal* speech (Dell 1980). This means that parsed internal speech is representable in Working Memory. How does it get there? Figure 1.1 expresses the assumption that internal speech is analyzed by the same Speech-Comprehension System as overt speech. In this way the speaker can detect trouble in his own internal speech before he has fully articulated the troublesome element. This happened, presumably, in the following self-correction (from Levelt 1983).

(1) To the left side of the purple disk is a v-, a horizontal line

There is reason to assume (see chapter 12) that the speaker of these words intercepted articulation of the word *vertical* at its very start. Presumably, the plan for *vertical* was internally available, understood, and discovered to have a nonintended meaning. In other words, the monitor can compare the meaning of what was said or internally prepared to what was intended. But it can also detect form errors. The Speech-Comprehension System allows

us to discover form errors in the speech of others. In the same way, it is able to notice self-generated form failures. This is apparent from a self-correction such as the following (from Fay 1980b).

(2) How long does that has to – have to simmer?

Dell (1980) found that speakers also discover form failures in their own internal speech. In short, speakers monitor not only for meaning but also for linguistic well-formedness (Laver 1973).

When the speaker detects serious trouble with respect to the meaning or well-formedness of his own internal or overt speech, he may decide to halt further formulation of the present utterance. He may then rerun the same preverbal message or a fragment thereof, create a different or additional message, or just continue formulation without alteration, all depending on the nature of the trouble. These processes are not of a different nature than what is going on in message construction anyhow.

The speaker no doubt also monitors messages *before* they are sent into the Formulator (see chapter 12), considering whether they will have the intended effect in view of the present state of the discourse and the knowledge shared with the interlocutor(s). Hence, there is no good reason for distinguishing a relatively autonomous monitoring component in language production. The main work is done by the Conceptualizer, which can attend to internally generated messages and to the output of the Speech-Comprehension System (i.e., parsed internal and overt speech).

1.3 Processing Components as Relatively Autonomous Specialists

The architecture in figure 1.1 may, on first view, appear to be rather arbitrary, and at this stage it is. There is no single foolproof way of achieving the partitioning of a complex processing system. There are always various empirical and theoretical considerations that have to be taken into account before one decides on one partitioning rather than another. It doesn't help much at this stage to say that the blueprint reflects earlier proposals by Garrett (1975), Kempen and Hoenkamp (1987), Bock (1982, 1987a), Cooper and Paccia-Cooper (1980), Levelt (1983), Dell (1986), and others. In fact, it is one of the aims of this book to argue that these proposals make sense. The present chapter can only give some background considerations for deciding whether a particular partitioning of the system is more attractive than another.

A first argument for distinguishing a particular processing component is that it is *relatively autonomous* in the system. The central idea is that a pro-

cessing component is a *specialist*. The Grammatical Encoder, for instance, should be a specialist in translating conceptual relations into grammatical relations; no other component is able to build syntactic phrases. Moreover, in order to execute these specialized procedures, the Grammatical Encoder needs only one kind of input: preverbal messages. That is its characteristic input. And in order to do its work, it need not consult with other processing components. The characteristic input is necessary and sufficient for the procedures to apply. More generally, it makes no sense to distinguish a processing component A whose mode of operation is continuously affected by feedback from another component, B. In that case, A is not a specialist anymore, it won't come up with the right result without the "help" of B. There is only one component then: AB.

There is another way in which the idea of components as autonomous specialists can be ducked, namely by assuming that all components receive as characteristic input the output of all other components (plus feedback of their own output). In that way each component has access to all information in the system. But this is tantamount to saying that components have no *characteristic* input—that they are general problem solvers that weigh all the available information in order to create their characteristic output. The Grammatical Encoder, for example, would access one lemma rather than another not only on the basis of the concept to be expressed, but also taking into consideration the morphology assigned to the previous word, the intonation pattern of the current sentence, the next intention the speaker has just prepared, and so forth. Some theorists like such models, which make each component an intelligent homunculus. The problems are, of course, to define the algorithm the component applies in considering this wide variety of information, and to realize this algorithm by a processing mechanism that can work in real time.

Generally speaking, one should try to partition the system in such a way that (a) a component's characteristic input is of a maximally restricted sort and (b) a component's mode of operation is minimally affected by the output of other components.

The combination of these two requirements is sometimes called *informational encapsulation* (Fodor 1983). In the blueprint of figure 1.1, these two requirements are met. Each component is exclusively provided with its characteristic input: the Grammatical Encoder with preverbal messages, which are conceptual structures; the Phonological Encoder with surface structures, which are syntactic entities; the Articulator with internal speech, which consists of phonetic representations; and so forth. The functioning of these processors is affected minimally, or not at all, by other

input. There is no feedback from processors down the line (except for some Formulator-*internal* feedback). The Articulator, for instance, cannot affect the Formulator's subcomponents. The only feedback in the system is via the language-comprehension components. This makes self-monitoring possible. But there is not even any *direct* feedback from the Formulator or the Articulator to the Conceptualizer. The Conceptualizer can recognize trouble in any of these components only on the basis of feedback from internal or overt speech.

These are strong and vulnerable hypotheses about the partitioning of the system. If one could show, for instance, that message generation is directly affected by the accessibility of lemmas or word forms, one would have evidence for direct feedback from the Formulator to the Conceptualizer. This is an empirical question, and it is possible to put it to the test. Studies of this kind will be reviewed in section 7.5. So far, the evidence for such feedback is negative.

A processing component may itself consist of subcomponents of varying degrees of autonomy. The Formulator in figure 1.1, for instance, consists of two subcomponents, which may be less autonomous than the Formulator as a whole. There is, in fact, convincing experimental evidence in the literature for the possibility of feedback from phonological to grammatical encoding (Levelt and Maassen 1981; Dell 1986; Bock 1987b; see also chapters 7 and 9 below).

And partitioning can even go further. It will, for instance, be argued in the course of this book that both of these subcomponents consist of even smaller building blocks, such as a noun-phrase processor and a verb-phrase processor within the Grammatical Encoder.

On the notion that a processing component is a relatively antonomous specialist, the following questions should be asked for each component that is proposed:

1. What are the characteristic kinds of information, or types of representation, the component accepts as input and delivers as output?
2. What sort of algorithm is needed to transform the input information into the characteristic output representation?
3. What type of process can execute that algorithm in real time?
4. Where does the input information come from, and where does the output information go to? A component can have one or more sources of input, and can transmit information to one or more other components.

In the course of this book these questions will return like the main theme of a rondo. For each component to be discussed, the nature of the target

output will be considered first. One cannot specify the operations of a component without an explicit characterization of the representation it computes. The Grammatical Encoder, for instance, produces what we called "surface structures" as output. Making a theory of grammatical encoding requires one to be explicit about what kinds of objects surface structures are. They are the target representations of the syntactic building operations (the grammatical encoding algorithm). Question 1 above has a certain priority over questions 2 and 3. In the following chapters I will honor that priority by considering a component's output representation before turning to its operations. Still, the questions are, in fact, interdependent. One may have good independent reasons for assuming a particular kind of operation. Speech errors, as we shall see, reveal much about the processes of grammatical encoding. We will naturally prefer to conjecture an encoding algorithm that does justice to such empirical observations. But the choice of algorithm, in turn, limits the kind of target representations that can be generated. Processes and representations cannot be studied independent of one another.

A component's output representation is, at the same time, the characteristic input for the next processor down the line. For each processor, we must ask whether there are circumstances under which it can be affected by information other than its characteristic input (issue 4 above). I have already mentioned the issue of feedback; in the next section I will discuss components' sensitivity to "executive control."

In subsequent chapters, a discussion of a component's output representation will always be followed by a treatment of its algorithm and its processes (i.e., issues 2 and 3 above). This will involve reviewing both theoretical proposals and empirical research. In all cases the depth of treatment is crucially dependent on the amount of detail provided by the existing literature.

The procedures an algorithm consists of are taken to be *productions* in the sense defined by Newell and Simon (1972) and used extensively by Anderson (1983). It was mentioned earlier that these productions are condition/action pairs of the kind IF X THEN Y, where X is the condition and Y the action. The example given was the conceptual procedure *IF the intention is to commit oneself to the truth of p, THEN assert p.* Here the IF clause is the condition. It states the speaker's prevailing intention. The THEN clause states the action. A speech act of the type "assertion" is to be made (such as *Manhattan is dangerous*, and not *Is Manhattan dangerous?*; the latter would be a question). Productions can contain variables, such as *p* in the example. That gives them generality, and it keeps them apart from

declarative knowledge (i.e., from the set of insertable propositions). Not only can conceptual algorithms be stated in the formal language of productions; the same holds for the algorithms of grammatical and phonological encoding.

Algorithms must run in real time. This means that the brain must be able to execute an algorithm—in fact, the sum total of all algorithms involved in speaking—in such a way that fluent speech results (see issue 3 above). This will make certain proposals for algorithms less attractive than others. Take, for instance, an algorithm for the planning of speech melody or intonation. It is known that speech melody bears some relation to the syntactic structure of the sentence. One might therefore be tempted to propose an algorithm that inspects the full surface structure before generating the appropriate melody for a sentence. But such an algorithm would violate the real-time requirement. Since no word can be pronounced without melody, the full surface structure of a sentence would have to be stored in the Syntactic Buffer before the sound form of its first word could be generated. This would create huge dysfluences between sentences, except when one would make the unlikely assumption that the speaker articulates sentence i while formulating sentence $i + 1$. I will return to this issue in the next paragraph, but first let me state that a main real-time restriction on speech planning should be that it run "from left to right" with very little lookahead.

One might want to go one step further and propose neural-network structures that could run the algorithm. This is still a very long shot for the algorithms involved in the process of speaking. Still, proposals of a quasi-neurological sort are being made for various aspects of language processing. These are the "*connectionist*" or "*spreading activation*" accounts (I will use the more accurate term "activation spreading"). In these accounts an algorithm is implemented in a network of connected nodes. The nodes can be in various states of activation, and they can spread their activation to the nodes with which they are connected. Figure 1.2 gives an example. It represents, in a highly simplified way, how the above procedure of accessing a lemma's corresponding sound form could be implemented in detail. Each lemma node in the lexicon is connected to a set of syllable nodes. The figure represents this state of affairs for just two lemmas, *construct* and *constrain*. The network connections are relatively permanent. What varies is the states of activation of the nodes. When the lemma *construct* is part of the surface structure, its node is in a state of high activation. The node "fires" and spreads its activation to its two constituent syllable nodes in the form lexicon: *con* and *struct*. Initially, *con* should be more highly activated

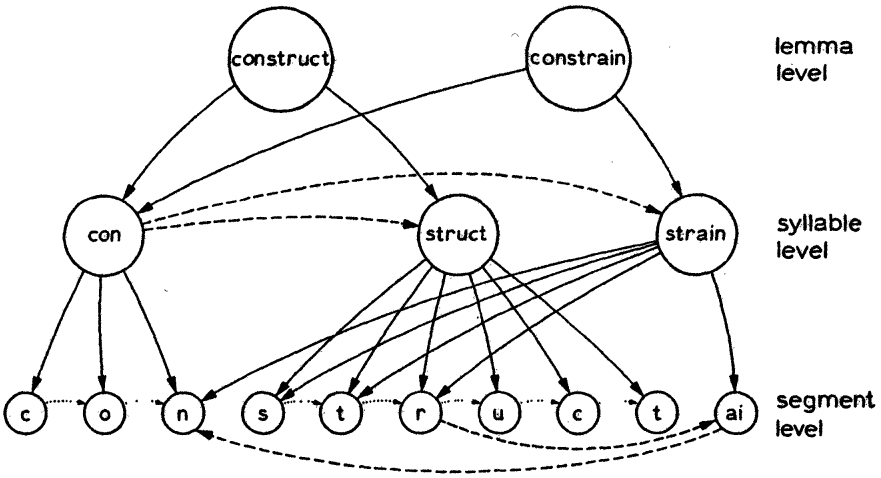

Figure 1.2
Example of an activation-spreading network.

than *struct*; otherwise the speaker may happen to say *structcon* instead of *construct*—a type of slip that is absent from collections of speech errors. In order to realize this, a directed inhibitory connection (dotted line) can be supposed to exist between the syllable nodes *con* and *struct*. When the syllable node *con* is sufficiently activated, it will, in turn, fire and spread its activation to the so-called segment nodes, *c*, *o*, and *n*. And again, their ordering of activation has to be controlled by a system of inhibitory connections. If everything runs well, the segment nodes will fire in the right order. It can further be assumed that a node, after spreading its activation, returns to a low state of activation. When this happens to the syllable node *con*, the inhibition on *struct* will fade away so that it can reach threshold activation and fire. Its constituent segments will be activated, and the inhibitory mechanism will make them fire in the right order. The lemma *constrain* is connected to a large part of the same network, and so are other lemmas that share syllables or segments with *construct*.

This is not meant to be more than an example. Activation-spreading or connectionist accounts vary enormously in detail (Anderson 1983; Dell 1986; Rumelhart et al. 1986; MacKay 1987). They differ in the kinds of nodes, the use of excitatory and inhibitory connections between nodes, the directions of spreading, the time characteristics of activation spreading, the summation function of input activations to a node, the possible range of activation states of a node, and the nodes' output function. They also differ in the control of timing and order. It may or may not be the case that *any* explicit process can be implemented in a "spreading activation" network.

Connectionism is, in the first place, a *formal language* for the expression of cognitive processes. It is not a *theory about* cognitive processes. Theories, whether expressed in the language of spreading activation or in the language of production systems, are coherent sets of principles which *restrict* the domain of possible processes. In other words, a theory *forbids* certain states of affairs to occur, whereas a sufficiently rich formal language doesn't. A formal language is a vehicle for the expression of interesting theoretical principles, which can be more or less convenient. And it can provide a complexity measure for the output generated by an algorithm. The connectionist formal language is especially convenient for the representation of principles of parallel processing, and there is much parallel processing in the generation of speech. In the course of this book we will meet certain restricted theoretical proposals that use the connectionist language of parallel distributed processing—in particular, Dell's (1986, 1988) theory of phonological encoding.

1.4 Executive Control and Automaticity

Speaking is usually an intentional activity; it serves a purpose the speaker wants to realize. An intentional activity is, by definition, under central control (Carr 1979; Bock 1982; Fodor 1983). A speaker can decide on one course of verbal action rather than another on the basis of practically any sort of information: his state of motivation, his obligations, his believing this rather than that, his previous speech acts or other actions, and so forth. The speaker will invest his attention on matters of this sort in planning what to say next.

Given the existence of central or *executive* control, an important question is to what degree the various processing components are subject to such control. When a component is not subject to central control, its functioning is *automatic*. The distinction between controlled and automatic processing is fundamental to cognitive psychology, and is based in a firm research tradition (LaBerge and Samuels 1974; Posner and Snyder 1975; Schneider and Shiffrin 1977; Flores d'Arcais 1987a).

Automatic processes are executed without intention or conscious awareness. They also run on their own resources; i.e., they do not share processing capacity with other processes. Also, automatic processing is usually quick, even reflex-like; the structure of the process is "wired in," either genetically or by learning (or both). This makes it both efficient and, to a large extent, inflexible; it is hard to alter automatic processes. Since auto-

matic processes do not share resources, they can run in parallel without mutual interference.

Controlled processing demands attentional resources, and one can attend to only a few things (the items in Working Memory) at a time. Attending to the process means a certain level of awareness of what one is doing. Human controlled processing tends to be serial in nature, and is therefore slow. But it is not entirely fixated in memory. In fact, it is highly flexible and adaptable to the requirements of the task.

Let us now look again at the components of the blueprint in figure 1.1. Clearly, the Conceptualizer involves highly controlled processing. Speakers do not have a small, fixed set of intentions that they have learned to realize in speech. Communicative intentions can vary in infinite ways, and for each of these ways the speaker will have to find new means of expression. This requires much attention. And introspection supports this. When we speak, we are aware of considering alternatives, of being reminded of relevant information, of developing a train of thought, and so forth. Message construction is controlled processing, and so is monitoring; self-corrections are hardly ever made without a touch of awareness. The speaker can *attend* to his own internal or overt speech. The limited-capacity resource in conceptualizing and monitoring is Working Memory. The system allows only a few concepts or bits of internal speech to be highly active, i.e., available for processing (Miller 1956; Broadbent 1975; Anderson 1983). On the other hand, not all processing in message encoding is under executive control. An adult's experience with speaking is so extensive that whole messages will be available in long-term memory and thus will be retrievable. Many conversational skills (such as knowing when and how to take or give a turn in conversation and deciding how direct or how polite one's speech act should be) have been acquired over the course of a lifetime and are quite directly available to the speaker. They are not invented time and again through conscious processing. Still, even these automatic aspects of conceptualizing are easily attended to and modified when that is required by the conversational situation. They are not "informationally encapsulated."

All the other components, however, are claimed to be largely automatic. There is very little executive control over formulating or articulatory procedures. A speaker doesn't have to ponder the issue of whether to make the recipient of GIVE an indirect object (as in *John gave Mary the book*) or an oblique object (as in *John gave the book to Mary*). Neither will much attention be spent on retrieving the word *horse* when one wants to refer to

the big live object that is conventionally named that way. These things come automatically without any awareness. They also come with very high speed. Speech is normally produced at a rate of about two to three words per second. These words are selected at that rate from the many tens of thousands of words in the mental lexicon. There is just no time to consciously weigh the alternatives before deciding on a word. Articulation runs at a speed of about fifteen phonemes per second. One should be grateful that no attention need be spent on the selection of each and every individual speech sound. Formulating and articulating are "underground processes" (Seuren 1978) that are probably largely impenetrable to executive control even when one wishes otherwise. (See Pylyshyn 1984 for more on cognitive impenetrability.)

There may be marginal forms of executive control, however. They are evidenced, for instance, in the fact that a speaker can abruptly stop speaking when he detects an error (Levelt 1983). The sentence or the phrase is then typically not completed. One can stop a word in the middle of its articulation, even ignoring syllable boundaries. It is apparently possible to send an executive "halt" signal to the individual processing components. Maybe similar signals can be sent to control other global aspects of processing, such as speaking rate, loudness, and articulatory precision.

The notions of automaticity, informational encapsulation, and cognitive impenetrability also figure centrally in the ongoing "modularity" discussions (Fodor 1983, 1985, 1987; Garfield 1987; Marshall 1984). The issue is whether, in an interesting number of cases, automatic components of processing also show several other features, such as being genetically given to the species, being located in specialized neurological tissues, and showing highly specific breakdown patterns. It is by no means excluded that some or all of these additional features have a certain applicability to the automatic processing components that underlie speech production. Only man can speak; there are dedicated neurological substrates for the production of speech in the left hemisphere; their disruption creates specific disorders such as agrammatism; and in the course of this book we will observe a multitude of characteristic breakdown patterns for different processing components, in particular speech errors. A processing component that shares most of these features is called a *module*. Whether the automatic components proposed in the blueprint above share the additional features that would make them modules will, however, not be a major issue in this book; hence, we will not call them modules.

1.5 Units of Processing and Incremental Production

1.5.1 Units of Processing

Much ink has been spilled on the question of what units of processing are involved in speech production, and in part for the wrong reasons. Many authors have tried to delineate *the* unit of speech, and this search for the Holy Grail has enriched the literature with an astonishing gamma of units. Others, surely, have recognized that there is no single unit of speech production, but have spent much attention on one particular unit. Here are some of the units one regularly encounters in the literature, with references to selected sources:

cycle (Goldman-Eisler 1967; Beattie 1983)
deep clause (Ford and Holmes 1978)
idea (Butterworth 1975; Chafe 1980)
information block (Grimes 1975)
information unit (Halliday 1967a; Brown and Yule 1983)
I-marker (Schlesinger 1977)
message (Fodor, Bever, and Garrett 1974)
phonemic clause (Boomer 1965)
phrase (Bock 1982)
proposition or propositional structure (Clark and Clark 1977; Herrmann 1983)
sentence (Osgood 1971, 1980; Garrett 1980a)
spurt (Chafe 1980)
surface clause (Hawkins 1971)
syntagma (Kozhevnikov and Chistovich 1965; McNeill 1979)
tone group (Halliday 1967a)
tone unit (Lehiste 1970).
total conception [*Gesamtvorstellung*] (Wundt 1900)
turn-constructional unit (Sacks, Schegloff, and Jefferson 1974).

And one can easily double or triple the length of this list. Foss and Hakes (1978) correctly remark that "speech has many planning units: words, syllables, phonological segments, and even phonological features."

The empirical evidence marshaled for one unit rather than another has been very diverse, including pause patterns, intonational structure, speech errors, and speech-accompanying gestures. Much of this evidence will be reviewed in the following chapters. The point to be stressed here is that there is no single unit of talk. Different processing components have their own characteristic processing units, and these units may or may not be

preserved in the articulatory pattern of speech. If, for instance, grammatical encoding involves units such as "noun phrase," "verb phrase," "sentence," and "clause," then these units need not be preserved in the prosody of the utterance. Later stages of processing—particularly the stage of phonological encoding—may undo these units of surface structure and impose a somewhat different organization (one more appropriate for fluent articulation). Still, the presumed absence of syntactic-clause boundaries in an utterance's prosody has been used as argument against multistage models of speech generation (McNeill 1979).

1.5.2 Incremental Production

A major reason for some theorists to object to multistage models and to prefer "multi-faceted single-stage speech production" (McNeill 1979) may be what Danks (1977) calls the "lock-step succession" of processing stages. It would indeed be disturbing if processing were strictly serial in the following way: First, the speaker generates the complete message to be communicated. Then, he generates the complete surface structure for the message. Next, the speaker starts building a phonetic plan for the utterance. Only after finishing this can the speaker begin to work on the articulation of the first word of the utterance. After completion of the utterance, the speaker can start preparing the next message. This would, of course, create serious dysfluences in discourse.

There is, however, nothing in stage models that requires this kind of seriality. Even though there can be no formulating without some conceptual planning, and there can be no articulating without a phonetic plan, message encoding, formulating, and articulating can run in parallel. Fry (1969) and Garrett (1976) made the obvious assumption that the next processor can start working on the still-incomplete output of the current processor (i.e., can start working before the current processing component has delivered its complete characteristic unit of information). Kempen and Hoenkamp (1982, 1987) called this *incremental processing*. All components can work in parallel, but they all work on different bits and pieces of the utterance under construction. A processing component will be triggered by any *fragment* of characteristic input. As was already noted in section 1.3, this requires that such a fragment can be processed without much lookahead—i.e., that what is done with the fragment should not depend on what will be coming in later fragments. Intoning the first few words of a sentence, for instance, should not depend on the way in which the sentence will finish. Some lookahead is, of course, necessary in certain cases. A speaker who is going to say *sixteen dollars* should not pronounce

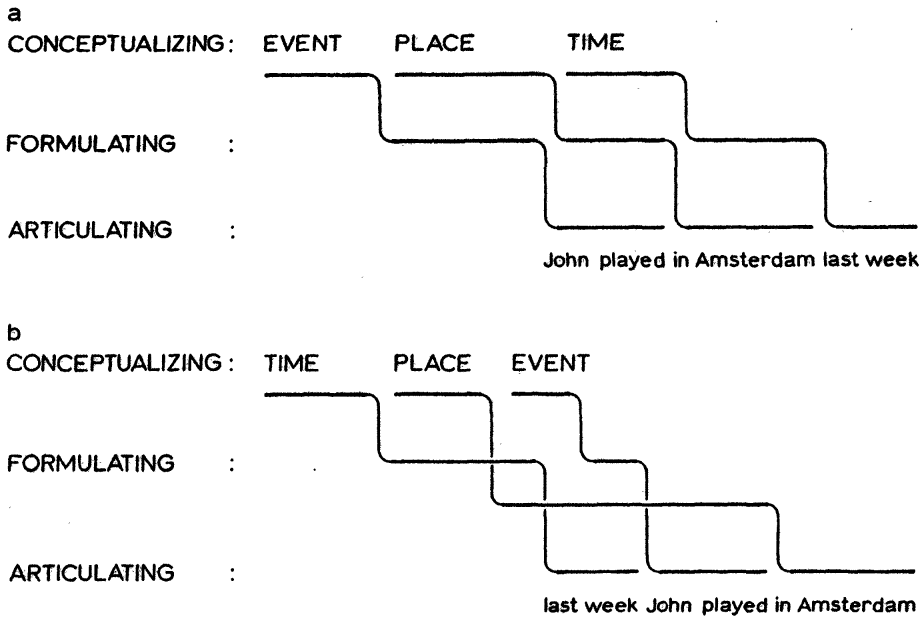

Figure 1.3
Incremental production without (a) and with (b) inversion of order. (After Kempen and Hoenkamp 1987.)

sixTEEN (a correct accentuation of the word) and then *DOllars*; rather, he should say *SIXteen DOllars*, with word stress shifted to SIX. In other words, in order for the right stress pattern to be generated for the first word, the stress pattern of the second word must be available. This is lookahead. But in order to make incremental processing possible, this lookahead should, for each processor, be quite limited. This puts interesting restrictions on the kind of algorithm that is allowable for each component.

It should immediately be added that a processing component will, on occasion, have to reverse the order of fragments when going from input to processed output. Figure 1.3 depicts incremental processing without (a) and with (b) inversion of fragment order. The first case is meant to represent an instance in which the speaker conceptualizes for expression an EVENT (John's playing before "now"), then the PLACE of the event (it took place in Amsterdam), then the TIME of the past event (during last week). When the first fragment of the message (the EVENT) becomes available, the Formulator starts working on it. While the Formulator is encoding the EVENT, the Conceptualizer generates the next piece of the message (the PLACE information). It is sent to the Formulator, which has just completed *John played.* This piece of the phonetic plan is now up for articulation. While this articulation proceeds, the Formulator encodes the PLACE information. At the same time, the Conceptualizer generates the

message fragment concerning the TIME information. And so on. This is pure incremental processing.

But the order of words doesn't always follow the order of thoughts. Figure 1.3(b) gives the case where the message fragments come in the order TIME, PLACE, EVENT. The formulation and articulation of the TIME information can follow the normal course, leading to the articulation of *last week*. But the Formulator cannot deliver its encoding of the PLACE information before having encoded the EVENT information; this would produce *last week in Amsterdam John played.* The Formulator, which is built to produce English syntax, will reverse the order and come up with *last week John played in Amsterdam.*

Other languages will have other ordering problems. A speaker of German, for instance, will have to swap fragments in the formulation depicted in figure 1.3(a), where they come in the order EVENT, PLACE, TIME, and should cast the sentence as *Hans spielte letzte Woche in Amsterdam.* It is obvious that, where such reversals are necessary, certain fragments must be kept in abeyance. In other words, components must have storage or buffering facilities for intermediate results. Three such facilities have already been mentioned: Working Memory (which can store a small number of message fragments as well as fragments of parsed speech), the Syntactic Buffer (which can store results of grammatical encoding), and the Articulatory Buffer (which can store bits of the phonetic plan). These buffers will, at the same time, absorb the asynchronies that may arise from the different speeds of processing in the different components.

Although there is a need for a theory of processing that will handle ordering problems of this kind, the main job will be to do as much as can be done with strictly incremental production. This is a time-honored principle in psycholinguistics. Wundt (1900) said that word order follows the successive apperception of the parts of a total conception [*Gesamtvorstellung*]. Of course Wundt added that this can hold only to the degree that word order is free in a language, but the principle is there. Let us call it *Wundt's principle*, but broaden it somewhat for the present purposes: *Each processing component will be triggered into activity by a minimal amount of its characteristic input.* In the following chapters we will, time and again, have to consider how small that minimal amount can be. How large a fragment of surface structure is needed for phonological encoding to do its work? How much of a phonetic plan must be available for articulation to be possible? And so forth. When these amounts are all small, articulation can follow on the heels of conceptualization.

But the theoretical assumption of incremental processing (i.e., of parallel processing activity in the different components of speech generation) hinges on automaticity. Only automatic processors can work without sharing resources and, thus, work in parallel. If each processor were to require access to attentional resources (i.e., to Working Memory), we would be in the situation that Danks (1977) called "lock-step succession." Then speaking would be more like playing chess: an overt move now and then, but mostly silent processing.

Summary

The intentional use of speech is subserved by information-processing skills that are highly complex and little understood. A case analysis of a single utterance appearing in a natural conversation gave a first impression of the intricacy of the processing that underlies speech. It also suggested a variety of kinds of information and of processing steps involved in the generation of an utterance.

How to partition such a system in a psychologically meaningful way? There is no single foolproof approach to this issue. This chapter's sketch of a framework for such a partitioning will be filled in and elaborated in subsequent chapters. The blueprint for the speaker consists of the following components: (i) *A Conceptualizer, which generates preverbal messages.* These messages consist of conceptual information whose expression is the means for realizing the speaker's intention. (ii) *A Formulator consisting of two subcomponents.* The Grammatical Encoder retrieves lemmas from the lexicon and generates grammatical relations reflecting the conceptual relations in the message. Its output is called "surface structure." The Phonological Encoder creates a phonetic plan (or "internal speech") on the basis of the surface structure. It has access to the form information in the lexicon, and it also incorporates procedures for generating the prosody of an utterance. (iii) *An Articulator, which unfolds and executes the phonetic plan as a series of neuromuscular instructions.* The resulting movements of the articulators yield overt speech. (iv) *The Speech-Comprehension System,* which makes self-produced internal and overt speech available to the conceptual system; this allows the speaker to monitor his own productions.

Each of these components, we assume, is an autonomous specialist in transforming its characteristic input into its characteristic output. The procedures apply largely without further interference or feedback from other components. The theoretical task ahead of us is to describe, for each component, what kind of output representations it generates, and by what

kind of algorithm. Also, we will have to consider how that algorithm is implemented as a mechanism operating in real time.

Next, the distinction between controlled and automatic processing was applied to these components. Message generation and monitoring were described as controlled activities requiring the speaker's continuing attention. Grammatical encoding, form encoding, and articulating, however, are assumed to be automatic to a large degree. They are speedy and reflex-like, require very little attention, and can proceed in parallel.

The proposed architecture allows for a mode of processing which Kempen and Hoenkamp (1987) called *incremental.* It combines serial and parallel processing in the following way: Each fragment of information will have to be processed in stages, going from the conceiving of messages to articulation. Still, all processing components can work in parallel, albeit on different fragments. If the fragments are small (i.e., if the components require little lookahead), incremental processing is efficient, producing fluent speech without unintended interruptions. That it is sufficient for a processing component to be triggered into activity by only a minimal fragment of characteristic input was called "Wundt's principle."

Intermediate representations, such as preverbal messages, surface structure, and the phonetic plan, have their own kinds of units; there is no *single* unit of processing in the production of speech. There must be storage facilities for buffering such intermediate representations as they become available. Working Memory can store messages and parsed internal speech. A Syntactic Buffer can store bits of surface structure. And an Articulatory Buffer can store stretches of articulatory plan for further execution as motor programs.

The chapters to follow will trace the blueprint of figure 1.1 from message generation to self-monitoring, considering the kinds of representations generated by the processors, the algorithms involved, and the real-time properties of these algorithms. In the course of this journey, certain parts of the blueprint can be worked out as theoretical statements with predictive potential. In many more cases, however, we will be able to do no more than "zoom in" a little closer on the details of the architecture, and in particular on empirical studies of these details. But first, we will give attention to the speaker as interlocutor—in particular, to his role in conversation.

Chapter 2
The Speaker as Interlocutor

The most primordial and universal setting for speech is conversational, free interaction between two or more interlocutors. Conversation is primordial because the cradle of all language use is the conversational turn-taking between child and parent (Bruner 1983). Unlike other uses of language, conversation is also universal; it is the canonical setting for speech in all human societies. The speaker's skills of language use cannot but be tuned to the requirements of conversation. Of course, this does not mean that they can be derived from or explained by conversational usage. One cannot deduce a car's construction from the way it does its canonical job of driving along the road, but it would be silly to ignore that behavior when studying the car's internal construction and operations. Similarly, one cannot dissect the speaker's skill into components without carefully considering the tasks these components, alone and together, have to perform. We know that they should at least allow the speaker to converse. The present chapter will review some essential aspects of a speaker's participation in conversation.

Conversation is, first, a highly contextualized form of language use. There is, on the one hand, the participant context. A speaker will have to tune his talk to the turns and contributions of the other persons involved; his contributions should, in some way or another, be relevant to the ongoing interaction. There is, on the other hand, a spatio-temporal setting, shared by the interlocutors, which serves as a source of mutual knowledge. By anchoring their contributions in this shared here and now, interlocutors can convey much more than what is literally said. Nonconversational forms of speech are usually less contextualized. The addressees may be scattered (as in radio reporting), the spatial setting may not be shared (as in telephone talk), the temporal setting may also not be shared (as in tape-recording), there may be turn taking without the other party's talking (as in