The Architecture of Normal Spoken Language Use*

W.J.M. Levelt

1. Introduction

The normal language user's production and understanding of speech involves the highly skilled coordination of myriad processes. When a speaker conceives of some communicative intention, he will select and order information whose expression may realize that intention. He will also formulate that information, *i.e.*, give it linguistic shape. This includes retrieving the appropriate words from memory and assigning them their proper grammatical roles and syntactic positions. The speaker will further compute a phonetic specification for the developing utterance, and use it go guide the articulatory execution that produces overt speech. The addressee will normally try to reconstruct the speaker's communicative intentions. She will perform an acoustic-phonetic analysis on the continuous speech signal in order to segment it into recognizable words and phrases. She will retrieve the syntactic properties and meanings of successive words and *parse* the string into meaning full phrases and sentences. And she will interpret this information in terms of the context of interaction, the purpose of the exchange, the presuppositions about the speaker's intentions, etc.

This complex system operate largely unawares. A speaker concentrates his attention on what is to be said, not on how it is to be done. The rate of fluent speech, some 2 to 3 words per second, is too high for a speaker to ponder over each and every word or syntactic construction. And there is no way for a speaker to consciously prepare some fifteen speech sounds per second. Similarly, the listener normally attends to the content of what is said, not to the phonetic shape of words or the syntactic complexities of the utterance. Like in all other skills, the lower level processes are automatic. They don't use central resources; they come for free.

In the following I will outline the architecture of this skill. This involves the dissection of the speech processing system into component subsystems. It also involves a characterization of the representations computed by the processing components as well as the manner in which these representations are computed. And it requires a specification of how the processing components cooperate in producing their joint product. These matters will be discussed in reference to Fig. 1.1. It is a 'blueprint' of the language user, depicting the main processing components and their connections. We will proceed anticlockwise through the system, beginning at the highest conceptual level at which the speaker generates information to be expressed. We will follow the flow of information down to articulation, the generation of overt speech. We will then turn to the perception and parsing of speech, back to the conceptual interpretative level. After this grand tour some remarks will be made about the modes of cooperation between the various components, in particular about incrementality and autonomy of processing.

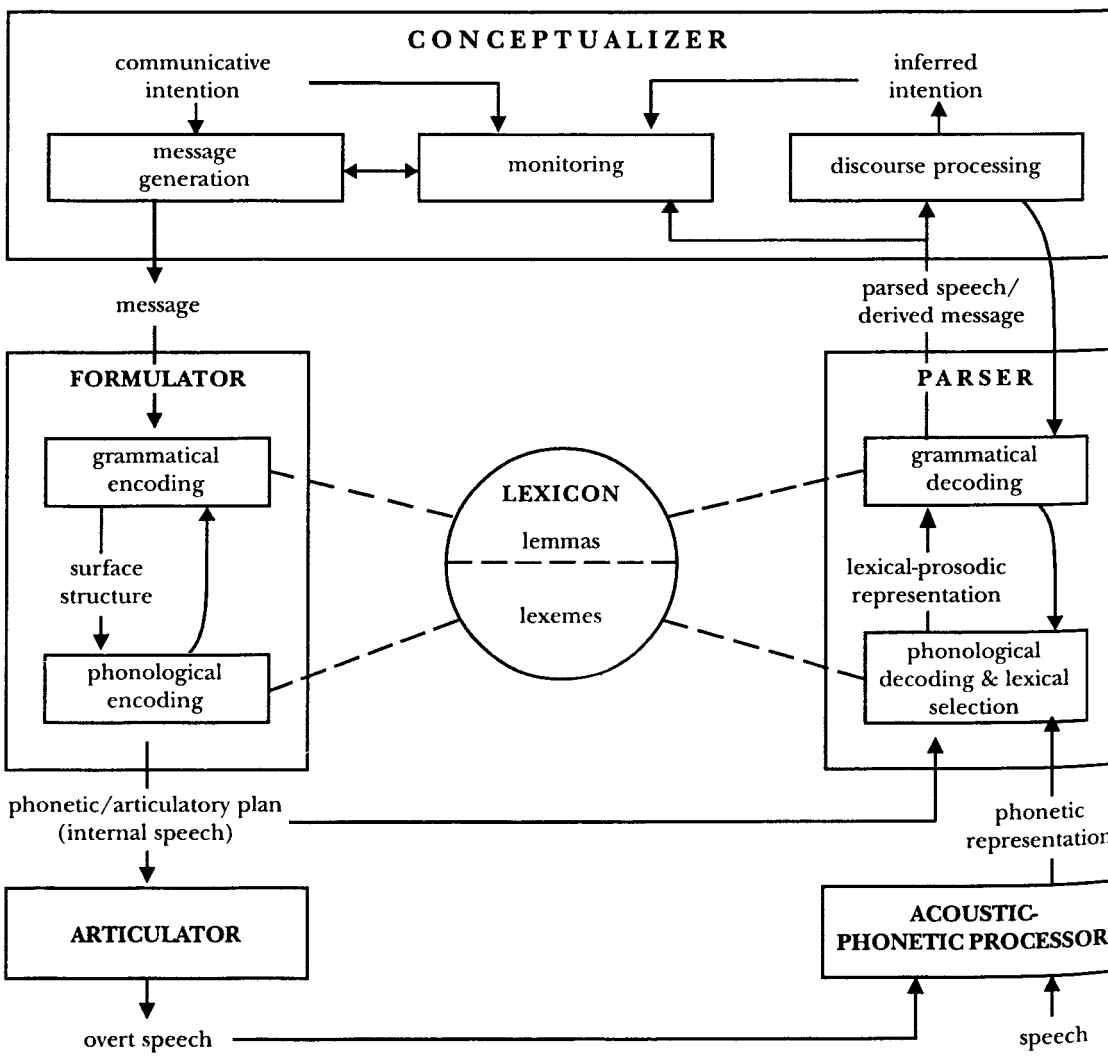
* *Lingistic Disorders and Pathologies*, An International Handbook, 1-15 Berlin, Walter De Gruyter

2. Speaking

2.1 Conceptual Preparation

The conceptual preparation of speech (see Fig. 1.1, 'conceptualizer') begins with the conception of an illocutionary intention, *i.e.*, a communicative intention the speaker decides to express by means of language. A communicative intention is one the speaker wants to be recognized as such by the addresses (Grice 1968). The speaker may want to refer to something, to express some belief or expectation, to commit the addressee or himself to some course of action, etc. In order to make such intentions recognizable by the addressee, the speaker will decide on a speech act and select information whose expression can realize that purpose. So, for instance, if the illocutionary intention is that the addressees will believe X, then one way is to state that X is the case. Or if the intention is that the addressee performs action Y, then the speaker may request that Y be performed. Often the information expressed

Fig. 1.1: Schematic representation of the processing components involved in spoken language use.



indicates the illocutionary intention indirectly. *It is cold in here* may express the request to close the window (Clark 1979). Such indirect speech acts convey an intention that is not literally expressed (as in *close the window).* It is the rule rather than the exception that the information selected for expression entertains an indirect relation to the intention to be conveyed. This is a powerful means to regulate personal relations in verbal interaction, to be polite, ironical, or whatever the situation might call for (Clark/Gerrig 1984). The information expressed invites the addressee to infer the intention, and this may be based on conventions (as in the request interpretation of *please)* and on the mutual knowledge of the interlocutors in a discourse situation (Grice 1975; Clark/Chunk 1980). Meanings are not so much expressed as negotiated between interlocutors. A simple intention to refer to a particular object may take several turns between interlocutors to become realized *(The newspaper— Which newspaper?—The one on the table — On the small table? — Yes,* cf. Clark/Wilkes-Gibbs 1986).

Conveying an intention may involve the planning of a sequence of speech acts. If a speaker decides to comply with a request to describe his apartment in some detail, he will have to retrieve from memory a complex spatial arrangement of spaces and objects. It will involve several statements of fact to express this information. In such cases speakers have to solve the 'linearization problem', *i.e.,* what to say first, what to express next, etc. When the complex information to be expressed is spatial, speakers solve the problem by making an imaginary tour — a body tour or a gaze tour —through the scene. When the information is temporal, as in the description of events, speakers tend to adhere to the chronological order (For an analysis of speakers' linearization principles see Levelt 1989, 138 ff). The speaker's planning of a speech act, his selection of information to be expressed, and his linearization of that information are called 'macroplanning' (Butterworth 1980; Levelt 1989, 123ff).

In addition there is 'microplanning' to be done. First, the information to be expressed should be given 'propositional shape' and 'perspective'. An image of an apartment is not enough for an apartment description: something should be said *about* that image, for instance that the table is in front of the window. The speaker decides what predications have to be made about what referents, and this usually involves taking perspective. Is the spatial relation to be expressed as a property of the table (being in front of the window) or as a property of the window (being behind the table)? Taking perspective is governed by various principles, among them Gestalt principles. Objects tend to be related to backgrounds, not inversely (*i.e.,* the table being in front of the window, rather than the window being behind the table). Small objects are related to big objects, not inversely (a pen on the table, not a table under the pen), etc.

Likewise, the expression of an event requires a predication to be made, and that predication should have a certain perspective. One and the same event can be propositionalized as the mother giving an ice cream to the child, or as the child receiving an ice cream from the mother. The proposition involves the same three arguments: mother, child and ice cream, but in the first case the mother is the topic of the event, in the second case the child is topic. These perspective relations have to be specified in the speaker's conceptual preparation of speech, because they are important determinants of word choice *(give, receive)* and of the assignment of grammatical roles *(the mother* as subject of the sentence or *the child).*

Microplanning also involves the assignment of an 'accessibility status' to the referents in a proposition. If the mother in the above proposition had already been mentioned in the previous sentence, she is likely to be in the focus of the addressee's attention. Knowing this, the speaker gives her a high accessibility status, which means that reduced or anaphoric reference can be made *(She gave the ice cream to the child).* There are various possible degrees of accessibility, each leading to a different choice of referring expression (cf. Levelt 1989, 144 ff). Related to accessibility is the assignment of 'prominence' to a referent. Something that may have particular 'news value' for the addressee can be marked in the message, so it will eventually be prosodically stressed in the utterance. Finally, microplanning involves certain language-specific decisions. When a language (such as English) has a

tense system, it is obligatory to specify the relevant temporal relations, even if they don't contribute to conveying the speaker's intention. A speaker of a tenseless language (such as Chinese) will only plan for the expression of temporal relations where they are relevant for conveying his intentions. Languages may require conceptual prespecifications of various sorts, such as the number (singular, plural) of referents, the degree of distance of objects to the speaker (proximal, medial, distal), the form or material of referents (fork-shaped, clay), etc. Any of these quite arbitrary properties may be 'grammaticized' in a language, *i.e.*, have an effect on morphology or syntax.

The eventual output of a speaker's conceptual preparation is technically called a 'message' (see Fig. 1.1); in cases of linearized information it is a series of messages. It is the information to be expressed, in prepositional format and supplied with perspectives, accessibility statues and language-specific conceptual information. Such messages form the characteristic input to the Formulator (see Fig. 1.1).

The Formulator maps messages onto linguistic form. It performs two relatively independent operations: grammatical encoding and phonological encoding.

2.2 Grammatical Encoding

Grammatical encoding takes a message as input and delivers a surface structure as output. A surface structure is a hierarchical organization of syntactic phrases. Its lowest-level elements are 'lemmas'. These are lexical elements as yet unspecified for phonological form (That specification takes place during phonological encoding). The surface structure is a syntactic, not a conceptual representation. It represents syntactic relations or functions such as 'head of phrase', 'subject of', 'direct object of' etc. In so far, it is what Garrett (1975) called the 'functional level representation'.

The generation of surface structure is lexically driven. Each lemma requires a particular syntactic environment. Grammatical encoding is like solving a set of simultaneous equations: the surface structure must be such that for all lemmas the required syntactic environments are realized.

Lemmas can become available prior to syntactic construction. How are they retrieved? Lemmas are entries in the 'mental lexicon' (see Fig. 1.1). That is the repository of knowledge about the words in one's language. Most lemmas have a semantic specification. This is the set of conceptual conditions under which the lemma can be appropriately used. So, for instance, the semantic specification for the lemma *give* is something like 'an event in which agent X causes object Y to be transferred from agent X to recipient Z' (This can, of course, be given formal expression in some logical language). So, if the speaker's message is that the mother gives the child an ice cream, the conceptual condition for the appropriate use of the lemma *give* is fulfilled: There is an agent X (the mother), who causes the transfer of an object or theme Y (the ice cream) to some recipient Z (the child). This being the case, the lemma *give* is retrieved from the lexicon. How does this take place? This is controversial, or rather unknown. Theories of lexical selection do exist (*i.e.*, Morton's 1969 Logogen theory and Miller and Johnson-Laird's 1976 decision table theory), but they are seriously inadequate and underspecified (cf. Levelt 1989, 198 ff). Research on speech errors (cf. Garrett, 1992) and experimental studies (Levelt/Schriefers/Vorberg *et al.* 1991) indicate that a single concept can (temporarily) trigger the activation of two or more semantically related lemmas (the lemma lexicon is clearly organized according to semantic principles). In most cases, however, a single lemma is selected and determines the further progress of grammatical encoding. (For a further discussion of these issues, see the special issue of 'Cognition' on lexical access in speech production, Levelt 1992.)

Each retrieved lemma contains two kinds of syntactic information: *First*, information specifying the lemma's syntactic category (major category, such as Noun, Verb, Adjective, Preposition, or minor category, such as Auxiliary, Determiner). *Second*, information determining the lemma's mapping of conceptual arguments (also called 'thematic roles') onto grammatical functions. The lemma *give* for instance, specifies that the conceptual agent (X) should be assigned the grammatical role of subject in the surface structure, that the theme (Y) should be realized as a direct object, and the recipient

(Z) as an oblique object. These mapping conditions are met in a sentence such as *The mother gave an ice cream to the child.* The lemma *give* also allows for another argument-to-function mapping, the so-called dative, which figures in a sentence like *The mother gave the child an ice cream.* Many transitive verbs also have a passive mapping, as in *The child was given an ice cream by the mother.* Which of these mappings is chosen during grammatical encoding depends on what the speaker has chosen for the topic of discourse (mother or child in the example); the topic is preferably realized as the subject. It also depends on the order in which the various lemmas become available from the mental lexicon. The dative will be preferred if the recipient lemma *(child* in the example) is available before the theme lemma *(ice cream).* Hence, temporal ordering in grammatical encoding is multiply determined. Three forces compete in this ordering process: the syntactic ordering restrictions set by the grammar; the schedule of lemma retrieval which, in turn, depends on the timing and saliency of the concepts in the message (cf. Bock/Warren 1985); and the topicalization the speaker has decided to achieve.

Various mechanisms of grammatical encoding have been proposed, in particular Kempen and Hoenkamp's Incremental Procedural Grammar (1987), Kempen and Vosse's Incremental Syntactic Tree Formation (1989), and MacWhinney and Bates' Competition Model (1989). See Levelt (1989, 235 ff) for a review.

The output of grammatical encoding is a surface structure, that is, a hierarchical phrase or tree structure with lemmas as terminal nodes. Each phrase corresponds to some predicate, argument or modifier in the message. And each lemma's syntactic environment is an appropriate realization of the lemmas syntax (its syntactic category and its argument-to-function mapping). Finally, lemmas may be marked for prosodic focus, dependent on the prominence of concepts in the message. The (developing) structure forms the input to the process of phonological encoding.

2.3 Phonological Encoding

Phonological encoding is the construction of a phonetic or articulatory plan, given the surface structure. The first step here is to retrieve each lemma's phonological specification, its 'lexeme', from the mental lexicon. These lexemes are not fully specified phonetic templates or motor programs, ready to be executed. In connected speech the shape of words is heavily dependent on their environment. For instance, the words *want* and *to* may become encoded as *wanna* in an utterance such as *Where d'you wanna go?* The context-dependent shape of words has to be created time and gain from abstract phonological specifications. The units *d'you* and *wanna* in the example are called 'phonological words'. They consist of a head word and one or more clitics that are phonological traces of full words.

When the retrieval of a lemma's lexeme is momentarily blocked, the speaker is in a 'tip-of-the-tongue' (TOT) state (cf. Kohn/Wingfield/Menn *et al.* 1987). The speaker may then know the word's initial segment(s) or its accent structure. Normally, however, access to lexeme information is quite fast. But there is evidence that even then segmental information is retrieved 'from left to right' (Meyer 1990; 1991).

This segmental information is to be inserted into metrical frames. The basic metrical frame corresponds to a phonological word. Introducing the utterance *Dick gave it,* two phonological word frames are set up, one for *Dick* and one for *gave it.* The first one is a one-syllable frame, the second one is a two-syllable frame. As the segments /d/, /I/ *and* /k/ become successively available, they fill the one-syllable frame. The segments of the *gave* and *it* lexemes, namely /g/, /eI/, /v/, /I/, and /t/, are successively inserted in the second phonological word frame. The first two segments, /g/ and /eI/, complete the first syllable; /v/, /I/ and /t/ compose the second syllable. The resulting syllabification (/geI-vIt/) thus 'straddles' the lexeme boundary. The output of this 'slot-filling' phase of phonological encoding is a string of syllable specifications. Crompton (1982) suggested that each syllable thus composed is in turn the address of a stored phonetic or articulatory plan for that syllable.

As soon as a syllable address is composed, the corresponding phonetic plan can be retrieved. This model has been adopted and further developed by Levelt (1989), but there are highly respectable alternative views on the transition from phonological to phonetic representations (cf. Browman/Goldstein 1990).

Thus, three processes cooperate in phonological encoding. First metrical frames are generated, whose basic units are phonological words. Very little is known about the mechanisms generating such frames (cf. Levelt 1989, 318 ff for a review). In the second process, segmental information is spelled out and is inserted into the frames. This has been the subject of extensive empirical and theoretical research. Most of this research is based on the analysis of naturally observed and experimentally elicited speech errors. Shattuck Hufnagel (1979) proposed the slot/filler model from which all subsequent work is derived. Dell (1986) added an activation spreading model of segmental spell-out, which provides a unified account of a large range of speech error phenomena. The third process is the mapping of syllabified and metrically specified phonological strings on to phonetic or articularly programs.

The articulatory plan's metrical specification is part of a larger prosodic specification. Phonological words are grouped together into 'phonological phrases'. They have their boundaries right after lexical heads-of-phrase (in particular after nouns in noun phrases, after main verbs in verb phrases and after adjectives in adjective phrases), as in *The detective/suddenly remembered/that the station/could be entered/from the other side/* (slashes mark ends of phonological phrases). In their turn, phonological phrases concatenate into 'intonational phrases', *i.e.*, phrases over which a particular intonation contour is realized. Speakers have great freedom in partitioning an utterance in intonational phrases. The above example can be spoken either as a single intonational phrase; or a break could be made between *remembered* and *that*, so that two phrases result. Intonational phrases are characterized by their *tone*, and each language has a repertoire of meaningful tones. So, for instance, the 'high-fall' as in *Peter has come* has a serious declarative intention, the 'low-rise' as in *Peter has come* has something reassuring to it, whereas the 'high-rise' as in *Peter has come* mildly invites confirmation. But a speaker's intonational repertoire is not limited to these conventional tones. Nothing in speech is so directly expressive of emotion as intonation and prosody in general (cf. Scherer 1986). It is as yet unknown by what mechanisms the formal-representational planning of speech is merged with its directly expressive use.

2.4 Articulation

There is no reason to suppose that the pace at which the phonetic plan is generated is always completely synchronous with the pace of articulation. In fact, there is good evidence for the existence of an 'articulatory buffer' which temporarily stores successive bits of phonetic plan (at least the size of phonological words) until they are to be executed (Morton 1969). The time needed to retrieve a unit from the buffer depends on the total number of units it contains, each additional unit adding some 10 milliseconds to the retrieval time (Sternberg/Monsell/Knoll/Wright 1978). It is, therefore, most efficient for the buffer to be relatively empty. A retrieved motor program first has to be 'unpacked', making the whole hierarchy of motor commands available. The more complex a unit, the more time it takes to unpack it (Sternberg/Wright/Knoll/Monsell 1980). Buffered speech is subjectively experienced as 'internal speech'. Internal speech can to some extent be monitored by the speaker (see section 2.5).

The articulatory program is executed by a motor system consisting of three major parts. The respiratory system provides the acoustic energy for speech by controlling the steady outflow of air. The respiratory cycle during speech is quite different from the normal breathing cycle; inhalation takes no more than 15 per cent of the total cycle and the air pressure gradient is almost constant during the outflow of air. The laryngeal system, with the vocal folds as their central part, controls

voicing and loudness in speech. During voicing the laryngeal system provides a periodic string of puffs with a wide range of high-frequency components. The supralaryngeal system or vocal tract contains three cavities in which this frequency spectrum is modulated, the nasal, the oral and the pharyngeal cavities. They control the timbre of vowels and consonants. The vocal tract can be constricted in different places (dental), palatal, alveolar, velar, uvular, glottal), and there are different manners in which these constructions can be released (plosive, fricative, affricative, lateral, etc.). The combinations of places and manners provide the wide variety of speech sounds that the world's languages display.

One of the most striking features of speech motor control is its context-dependency. The same speech sound can be produced in different ways, and speakers almost instantly adapt to changing physical contingencies in the vocal tract. Speech produced by a speaker with a pipe in his mouth hardly differs from unhampered speech. It is, therefore, unlikely that phonetic programs are detailed description of articularly gestures. Rather, they are 'tasks' for the articulators to perform; they prescribe the speech sounds to be generated. Recent theories of articulation suppose the existence of 'model-referenced' control (Arbib 1981). There is an internal model which relates an equivalence class of articulatory gestures to a particular target sound. Through proprioceptive feedback the model is also informed about the contingencies in force. It then automatically chooses the least energy consuming way to produce the sound structure dictated by the 'task'. The set of muscles that cooperate as a unit in the context-dependent execution of a particular phonetic task is called a 'synergism' (Lennerberg 1967) or a 'coordinative structure' (Saltzman/Kelso 1987).

2.5 Self-monitoring and Repair

Speakers are their own hearers. They can monitor anything in their own speech that they can monitor in the speech of others. They can attend to both the meaning and the form of their utterances. And when they detect trouble, they may interrupt themselves and make a repair, such as in *Left to pink — er straight to pink* (speaker tries to describe a pattern of coloured dots) where a lexical error is repaired or in *What are this kid — is this kid going to say incorrectly?* where syntactic agreement is re-established.

How is this self-monitoring performed? There are three loci of self-control in Fig. 1.1 First, the speaker can attend to his own message planning. He may revise a planned message before it gets formulated: This is probably happened in *Tell me, er what — d'you need a hot sauce?* Whatever the speaker intended to say here was revised before it was fully delivered. Hesitations in speech are often due to such preformulating revisions. Second, the speaker can attend to his internal speech. Speech in the articulatory buffer is accessible to attention in just the same way as overt speech. It can be parsed by the speaker's own language comprehension system. The speaker may then become aware of trouble (*i.e.*, at the conceptual level) and decide to interrupt himself. This is the 'internal monitoring loop'. As long as the buffered internal speech has not been articulated, such an interruption can prevent the production of an error. This probably happened in *A. v- ... a horizontal line,* where the speaker was on the verge of uttering *vertical* instead of *horizontal.* Third, the speaker can attend to his own overt speech, using his normal speech understanding system. This is the 'external monitoring loop'. The internal loop is faster than the external loop (cf. Lackner/Tuller 1979; Levelt 1989, 467 ff), but research on close shadowing (Marslen-Wilson 1985) shows that the external loop can, in exceptional cases, work in as few as 250 milliseconds, *i.e.*, a syllable's duration.

There is good evidence that speakers interrupt their speech as soon as serious trouble is detected (Levelt 1983). But detection can be late because the speaker attends mainly to the planning of speech, not to his output. This means that one or more words can follow the trouble item before the speaker detects it. This happened in the utterance *And from green left to pink — er from blue left to pink.* Here *green* was wrong, but the speaker delivered three more words before interrupting his speech. The place of interruption is unprincipled; it can violate any clause, phrase, word, or syllable boundary. There is a tendency, however, to complete a word that is not itself the trouble item (Levelt 1983).

After interruption, the speaker often utters some 'editing expression'. When the occasion for repair is a real error, the dominant editing term is *er*, especially when the interruption is immediate. But it can also indicate a more explicit rejection of what was said, such as *no* or *sorry*. The use of editing terms is different when the repair is for appropriateness, not for error. In *To the right — further to the right is yellow* the repair involves a further specification, not a rejection of what was said. In such cases editing terms are used infrequently and they are of a different type, not indicating rejection, such as *I mean*.

Though there is no linguistic systematicity in where a speaker interrupts his speech, resuming speech is, in fact, systematic. It is governed by a linguistic well-formedness rule (Levelt 1983; 1989, 486) which says that repairing follows the rules of coordination. A repair like *Is the doctor seeing - er - the doctor interviewing patients?* is linguistically ill-formed, and so is the corresponding coordination. *Is the doctor seeing patients or the doctor interviewing patients?* The repair *is the nurse - er - the doctor interviewing patients?*, however, is well-formed and so is the corresponding coordination *Is the nurse or the doctor interviewing patients?* Linguistically ill-formed repairs are probably as infrequent in spontaneous speech as are other linguistic deviances, but they can be experimentally elicited (van Wijk/Kempen 1987). The linguistic systematicity of repairing shows that the speaker keeps the interrupted formulation in abeyance during the planning of repair. The repair proper is then grafted on this still available grammatical structure.

Repairs are also often elicited by the interlocutor. She can make the speaker aware of some error or unclarity by saying *what?*, by raising her eyebrows, or by other signals to the speaker. In normal conversation, speakers capitalize on this cooperative feedback. They do not have to be too precise in planning their speech, since the interlocutor will react when necessary.

3. Speech Understanding

In normal conversation the listener's objective is to discover the speaker's communicative intentions. Several speech understanding mechanisms cooperate to accomplish this. There are, first the mechanisms that perform an acoustic-phonetic analysis of the speech signal; they produce phonetic representation of the signal. It is the code for accessing the lexicon and for deriving the metrical structure of the utterance. The recognized words and the prosodic information are then used to perform syntactic and semantic parsing of the utterance. And finally, the listener will interpret this linguistic structure in terms of the ongoing discourse in order to derive the speaker's communicative intentions.

We will consider these processes by following the right side of Fig. 1.1 from bottom to top.

3.1 Acoustic-Phonetic Analysis

Most theories of speech perception assume the existence of a 'front end' processor that receives the acoustic signal as input and that delivers some kind of featural representation as output. Opinions differ about the character of these features. Liberman/Mattingly (1986) suggest that they are representations of the speaker's intended articulatory gestures. This would bring the output of acoustic-phonetic analysis quite close to what we called above the speaker's articulatory plan. Others stay closer to the spectral properties of the speech signal, and suggest the existence of detectors for onsets and spectral peaks, and for frequencies and motions of formants. These, in turn, are used to derive the presence, absence, or degree of phonetic features such as voicing, nasality, coronality, vowel height, stridency, sonorance, etc., as well as their temporal distribution (see especially Stevens 1986). These patterns of features are the access codes to words in the lexicon (Lahiri/Jongman 1990). They are called 'phonetic representations' in Fig. 1.1.

There is still much controversy about the character of the phonetic representations and the way they are derived from the speech signal; see the two excellent reviews of acoustic-phonetic processing by Pisoni/Luce (1987) and Klatt (1989). A major problem to be dealt with is the huge variability in

the speech signal. A word's acoustic shape depends on its linguistic context (e.g., the phonological word and phrase in which it partakes, its stress and its intonation), the rate of speech, the dialect and sex of the speaker, the reverberation and noise in the room, and so on. Still, there is an increasing conviction that there are relational patterns in the speech signal that are robust and from which the presence of phonetic features can be reliably derived (cf. Stevens/Blumstein 1981; Stevens 1986; Zue 1986).

Whatever the precise nature of the phonetic representation, it must be such that the listener's linguistic parsing can be based on it. The parser contains two major processing components. One deals with phonological decoding and lexical access, the other one with grammatical (both syntactic and semantic) decoding.

3.2 Phonological Decoding and Lexical Selection

The phonetic representation of an utterance forms the access code to the lexicon. When the listener is exposed to connected speech, a first major problem is how to segment it. To recognize a word, one must know where it begins. But connected speech does not have the nice spaces between words that we encounter in written language. And there is no reason to suppose that phonetic representations, *i.e.*, temporal distributions of phonetic features, make this task any easier. There are two ways in which the listener can approach the segmentation problem. The first one is to use cues in the signal itself. Cutler/Ladd (1983) reviewed studies showing that English listeners prefer to perceive word boundaries before strong syllables (*i.e.*, syllables that carry word accent). There is, for instance, Taft's (1984) finding that listeners dominantly perceive such speech strings as [lettuce] as a single word, *lettuce*, not as two words *let us*, whereas words such as [invests] are preferably heard as two words, *in vests*, not *invests*. Cutler (1989) computed that this segmentation strategy will be about 85 percent correct for English open class or content words. Other cues could be of a phonotactic nature. Frazier (1987a) argued that certain strings can only be word-final, such as [-arpt] in English. That would tell the listener that what is about to follow must be a new word. A second approach is to recognize a word before it ends. The listener can then predict the end of the word, and hence the beginning of the next one. This would be a powerful strategy given the cohort theory, to which we will presently turn. Still, the perennial problems in automatic speech segmentation show that a satisfactory theory of segmentation is still long in coming.

But assuming that the listener knows where in the phonetic representation a new word is about to begin, how is that word identified? Domanant here is Marslen-Wilson's 'cohort theory' (for recent publications, see Marslen Wilson 1987; 1989). For each lexical item the recognition lexicon contains an abstract phonological code (probably not unlike the phonological code in the production lexicon. Notice that Figure 1.1 does not distinguish between a perception and a production lexicon; opinions differ as to whether such a distinction has to be made). This phonological code specifies a word's non-redundant distinctive features. For English, a phonetic feature like [+ aspiration], as after [p] in *pot*, will not appear in the lexicon because it is not distinctive (there are no two different words *p'ot* and *pot*). Similarly, for a nasal consonant the feature [+ voice] will not be specified, because it is redundant: all nasals are voiced (but not inversely). Finally, a feature is only specified in the word's representation if it is marked. Thus, nasality is marked and needs specification. But non-nasality is the default case; it needs no specification in the lexicon. See Lahiri/Marslen-Wilson (1991) for further details of this representational theory.

After a small stretch of sensory input (corresponding to about two segments of the input word) has been received, its phonetic feature pattern activates all lexical items whose word-initial phonological specification matches the input. So, if the input word is *trespass*, and the phonetic features of [tr] have become available, not only the lexical item *trespass* will be activated, but also items such as *tree, tremble, trestle, trombone*, etc. This set of initially activated items is called the 'word-initial cohort'.

As the incoming speech signal proceeds, the word-initial cohort is successively reduced. As soon as a feature appears that contradicts the specification of /o/ in *trombone* or /i/ in *tree, trombone* and *tree* disappear from the cohort. Next, the fricative feature of /s/ will exclude a candidate like *tremble*. The reduction of the cohort proceeds until a single candidate is left. In the example case this happens when the initial stretch *tresp-*has been received. The item *trespass* is the only one in the recognition lexicon whose phonological specification matches the phonetic pattern of that initial stretch.

The time course of word recognition is jointly determined by the developing information in the speech signal, and by the ensemble of competitors to the target word. The ensemble of competitor items determines a word's 'uniqueness point', the point at which it becomes different from all competitors in the lexicon. For *trespass* the uniqueness point is at /p/. The empirical issue is whether listeners in fact recognize a word when its uniqueness point is reached, *i.e.*, does the recognition point coincide with the uniqueness point? The experimental evidence is surprisingly assuring here (cf. Marslen-Wilson 1987), though certain factors such as semantic/syntactic context (Zwitserlood 1989) or word frequency can additionally affect the position of the recognition point. Contrary to predictions of alternative theories, such as the connectionist model TRACE by Elman/McClelland (1984), there is no evidence that activated competitors actively inhibit one another (Frauenfelder/Segui/Dijkstra 1990).

Though accessing the lexicon is the main target of phonological decoding, another major task is prosodic decoding. In particular, the listener will have to recognize the metrical groupings of the words recognize the metrical groupings of the words recognized. This will facilitate subsequent syntactic processing. For instance, the last element in a phonological phrase is usually the head of a syntactic phrase. Also, the listener will have to recognize the tone of an intonational phrase because this carries important communicative information. Little is known about these aspects of prosodic recognition, but see the volume edited by Cultler/Ladd (1983).

The final output of phonological decoding/lexical selection is what has been called a 'lexical/prosodic representation' in Fig. 1.1. It forms the input to the higher-order interpretative processes.

3.3 Grammatical Decoding

As lexical items and metrical structure become successively available, the listener will immediately try to interpret these materials. The incrementality of this process (see section 4) implies that syntactic and semantic processing proceed 'on-line' with the incoming information. There is solid experimental evidence that this in fact is the case in speech understanding (Marslen-Wilson/Tyler 1980; Tyler/Warren 1987).

Though syntactic analysis and semantic interpretation develop hand in hand, each follows its own principles. And there is a certain degree of autonomy in syntactic processing; it proceeds even in the face of semantic implausibility or anomaly. Let us first consider some aspects of syntactic processing.

The time pressure under which a listener must perform her syntactic analysis of the incoming materials, makes her follow a principle of syntactic analysis that Frazier (1987 b) called 'minimal attachment'. It means that new materials are syntactically attached in such a way that a minimum number of nodes is to be added to the existing syntactic structure. For example, when the listener has heard *The spy saw the cop* and now takes in *with binoculars*, then she will preferably attach the later prepositional phrase to the already existing verb phrase node. This would lead to the interpretation that seeing the cop (by the spy) was done with binoculars. The non-preferred alternative, however, is that the noun phrase *the cop* is deleted from the verb phrase node, a new additional noun phrase node is created for *the cop with binoculars*, which then in turn gets attached to the verb phrase node. This would imply the interpretation that the cop was equipped with binoculars. The minimal attachment principle predicts that listeners will tend to 'garden path' on sentences where minimal attachment leads to the wrong solution. According to this assumption, a sentence like *The horse raced passed the barn fell in a puddle* must be harder to understand than *The horse raced past the barn and fell in*

a puddle, and in fact it is (Ferreira/Clifton 1986). Regrettably, the minimal attachment principle has almost exclusively been studied in reading tasks. Whether it is equally valid for spoken language comprehension, where prosody provides important cues, is an open issue.

That speech is syntactically processed even if it is semantically anomalous was demonstrated by Marslen-Wilson/Tyler (1980). In a more recent experiment, Tyler/Warren (1987) confirmed and extended this finding. They used a word monitoring task, in which the listener had to push a button as soon as she heard a given target word. For example, the listener is given KITCHEN as target word, and is presented with the following normally spoken sentence: *The maid/ was carefully peeling/the potatoes / in the garden / because during the summer / a hot KITCHEN / is unbearable to work in* (here the slashes mark phonological phrase boundaries). For such sentences listeners push the button about 300 milliseconds after onset of the target word. When a semantically anomalous text was presented (such as *An orange dream / was loudly watching / the house /during smelly nights / because within these signs / a slow KITCHEN /snored / with crashing leaves)* reaction times were, not surprisingly, slower (by some 60 milliseconds). If, however, the local syntax around KITCHEN was also disrupted (by replacing a *slow KITCHEN* by *slow very KITCHEN),* latencies increased (by another 45 milliseconds). Thus, local syntactic parsing apparently proceeds where it can, even when the speech is semantically anomalous.

However, this is not the case for global syntactic parsing. When the phonological phrases in the anomalous sentence are scrambled (as in *Because within these signs/during smelly lights/was loudly watching/ the house /an orange dream/a slow KITCHEN / snored /with crashing leaves)* so that global (but not local) syntax is disrupted, the reaction times are the same as for the original semantically anomalous text. Tyler and Warren could show that global syntax is only effective when the text is meaningful. And this should not be surprising. The eventual output of global syntactic parsing is a conceptual structure whose thematic roles (see section 2.2. above) correspond to the function-argument mappings of the retrieved lexical items. If these mappings are violated, as in semantically anomalous text, global parsing will be blocked.

The following picture emerges: First, as words are successively recognized, their syntactic and semantic information becomes available to the parser. The syntactic information is concerned with a word's category (Noun, Verb, etc.) and its argument-to-function mapping (see section 2.2. above). Of these, the category information is probably enough to initiate a second process, local syntactic analysis (Frazier 1989). We have no difficulty in parsing nonsense such as *The beer slept the guitar.* Here we construct a verb phrase, *slept the guitar,* with *the guitar* as direct object of *slept,* in spite of the fact that this contradicts the argument-to-function mapping of *sleep* (an intransitive verb which does not take a direct object). The initial syntactic parsing can also ignore semantics (the meaning of *sleep* dictates that the verb requires an animate subject, but *beer* is inanimate). However, initial local parsing is very dependent on the intactness of phonological phrases. Tyler/Warren (1987) found a dramatic increase of reaction times when the phonological phrase was broken up (e.g., *the slow KITCHEN* being distributed over two phrases, with a break after *the slow).* In short, initial local syntactic analysis uses the metrical output of the acoustic-phonetic component plus the syntactic category information of the recognized words. But a third process, global syntactic parsing is, at least in part, based on semantic analysis. Grammatical functions such as 'subject', 'direct object', or 'indirect object' are translated into the thematic roles they are attached to in the lexical representations. For instance, the indirect object of the verb *give* in a sentence will be interpreted as the recipient of the action of transfer denoted by *give.* This is syntactically based semantic interpretation. The final output of global parsing, then, is a conceptual structure, akin to the speaker's 'message' (see section 2.1). It is therefore referred to as a 'derived message' in Fig. 1.1.

Only one aspect of global parsing has been studied in some detail for spoken language: the resolution anaphora. Syntax can restrict the interpretation of an anaphor. In the sentence (I) *The boxer told the skier that the doctor of the team would blame himself for the recent injury,* the reflexive pronoun *himself* can

only refer back to *the doctor*, not to *the boxer* or *the skier*. Will the listener reactivate *the doctor* at encountering the anaphor, in correspondence with syntax, or will the other referents also be reactivated, ignoring syntax? Nicol (1988) studied this question in a so-called cross-modal priming experiment. The subject listened to a sentence of this kind. Right after the pronoun, either a word (like *nurse*) or a non-word (like *murve*) was presented visually. The subject's task was to decide whether the visual probe was a word or a non-word, and to indicate this by pushing either a yes-or a no-button, accordingly. The latency of this lexical decision was measured. If the pronoun reactivates the syntactically indicated antecedent (e.g., *the doctor*), this may lead to a more rapid response to an associate probe (e.g., *nurse*) as compared to a neutral probe (e.g., *house*), and in fact it did. In contrast, no facilitation was obtained for associates of the other potential antecedents (e.g., *the boxer* or *the skier*).

What happens when there is no syntactic restriction on the interpretation of a pronoun? Consider sentence (ii), *The boxer told the skier that the doctor for the team would blame him for the recent injury*. Here the pronoun *him* can refer back to *the boxer* or *the skier* (but not to *the doctor*). And Nicol now found reactivation of both *the boxer* and *the skier* (but not of *the doctor*), in complete agreement with the syntax.

Even null-anaphors produce essentially the same effect. In *The policeman saw the boy that the crowd at the party accused t of the crime, t* stands for a null-anaphor (or *trace*) that refers back to *the boy*. And indeed, by probing at this position Swinney/Ford/Frauenfelder/Bresnan (unpublished) showed that *the boy* was reactivated there, but not *the crowd*. For an excellent review of this and related work, see Nicol/Swinney (1989). Notice also that in examples like these the listener is making essential use of the verb's function-argument structure. The verb *accuse* requires a direct object, so the listener knows that there must be a direct object trace after *accuse*. Global syntactic processing takes the verb-argument structure into account (*i.e.*, different from local syntactic processing).

Though apparently automatically restricted by syntax, the listener's binding of anaphors is a semantic activity. Cloitre/Bever (1988), for instance, found that when the syntactic antecedent is a concrete noun it is easier reactivated by a pronoun than it is an abstract noun.

Though the derived message is a conceptual interpretation of the utterance, it is one at a rather shallow level. As a rule, it will not be the ultimate interpretation. In example (ii) above, for instance, the derived message does not specify whether the anaphor refers back to *the boxer*, *the skier*, or somebody else mentioned earlier in the discourse. Solving such problems of interpretation, as well as many others, are part of 'discourse processing', to be discussed after a final remark on the output of the parser.

Though derived messages form the parser's main output, they are not the only information it makes available to the conceptualizer. The parser can also transmit certain intermediary results. The listener can, for instance, attend to phonetic or syntactic properties of the utterance. She can, for instance, notice the speaker's dialect, or grammatical errors. The term 'parsed speech' in Fig. 1.1 includes these aspects, as well as the main output of parsing, the derived message.

3.4 Discourse Processing

The last step towards deriving the speaker's communicative intention is called discourse processing. It is the subject of extensive research; here we can touch on a few aspects only. The first one is 'identifying referents' and refers to the listener's attempt to determine what entities the speaker is talking about. Clark/Schreuder/Buttrick (1983) showed subjects a photograph depicting president Reagan and David Stockman (a high, but not well-known official). When they asked *You know who this man is, don't you?*, most subjects answered that it was Reagan. But when asked *Do you have any idea at all who this man is?*, nobody said Reagan; they all made guesses about the other person on the photograph. The referent chosen for *this man* depends on the listener's presupposition about the speaker's knowledge, namely that Reagan is better known than the other depicted person.

An important mechanism in the interpretation of discourse, and in particular in the assignment of reference, is the construction of a 'discourse model;. It is a mental model (Johnson-Laird 1983) about the state of affairs discussed. Typically, when the listener encounters an indefinite description (as in *There is a baby here),* she will set up a new address in the model (in this case an address for the baby entity). However, when she encounters a definite description (as in *The baby is crying,* or *It is crying),* she will infer that the address is already present, and the predication is interpreted as being about the entity at that address (cf. Seuren 1993 for a theory of reference assignment in discourse models).

Earlier we saw that anaphoric reference can still be undecided in the derived message (as was the case for *him* in example (ii) above). The listener can then choose the most salient entity in the discourse model as a referent for the pronoun (Clark/Schreuder/Buttrick 1983; Morrow 1986). Alternatively, the listener may select the entity mentioned last *(the skier* in example (ii)). This is also dependent on the 'genre' of the discourse, as Morrow established.

Inferring referents is devilishly complicated. How does a listener (waitress in a restaurant) compute the person referent for her colleague's remark *The hamburger wants the bill* (Nunberg 1979)? It must be the person who was served a hamburger. Nunberg argued that there is some 'referring function' that maps the demonstratum *(the hamburger)* onto the intended referent (see also Fauconnier 1985). But how does the listener compute such arbitrary referring functions?

Such cases of indirect reference are by no means the only instances of non-literal interpretation in speech understanding. Other paramount cases are the interpretation of indirect requests (see section 2.1), the interpretation of metaphor (Seperber/Wilson 1986), irony (Clark/Gerrig 1984), hyperbole (Grice 1975) under other figures of speech. In these latter cases prosody, and especially the tones of intonational phrases, may cue the listener to infer the speaker's intention.

Finally, the 'denotation of words' surpasses the shallow semantic interpretation of the derived message. Hormann (1983) showed that quantifiers such as *some or many* are interpreted in relationship to what they quantify. For example, there are less people in *some people* than grains of sand in *some grains of sand.* Morrow and Clark (1988) extended this study by showing that the distance denoted by the verb *approach* depends on what is approaching and what is being approached. A tractor approaching a farmhouse, for instance, is at a greater distance than a mouse approaching a farmhouse. The distance is even smaller, when a mouse approaches a piece of cheese. In all these cases the listener's discourse model, or image of the situation is decisive of the inference.

Discourse processing requires the listener's full attention. The derived message, which come rather automatically with the incoming speech, is only a cue to the interlocutor's intentions. There is no limit to the variety of discourse situations, and every utterance is to be interpreted in the light of the current situation, the knowledge the interlocutors share, the social relation they are engaged in the lay-out of their physical environment, and so on. Recovering the speaker's intentions, not speech perception or parsing, is the listener's ultimate goal.

4. Incrementality, Autonomy and Interaction

It is not the intention of Fig. 1.1 to suggest that processing components work one after another. On the contrary, parallel processing does take place both in speaking and in listening. However, it occurs in a specific way that Kempen and Hoenkamp (1982) called 'incremental'. Though all components work in parallel, they work on different bits and pieces. Concretely, as soon as the speaker has conceived of an initial concept to be expressed, it is delivered to the Formulator. Instead of waiting until the whole message has been planned, the first concept is grammatically encoded — a word is retrieved from the lexicon and receives case. That bit of surface structure is then subjected to phonological encoding, and so on. Meanwhile other fragments follow the same route. As a consequence, a speaker can begin articulating a sentence long before he has completed the planning of his message. It is a

'roofing tile' organization of processing. Similarly in speech perception, as soon as the initial fragment of the speech signal has been received, acoustic-phonetic processing of that bit of signal begins. And as soon as its first two or three segments have been phonetically analyzed a word-initial cohort is activated. Syntactic local parsing can start as soon as the first word is recognized, and so can semantic processing and discourse interpretation. Each processor works incrementally on successive fragments it receives from the previous one. Thus, there is both staging and parallelness of computation.

Parallel computation is only possible because most components are fully automatic. If the language user were to pay attention to all component processes, parallel processing would be infeasible, Language users can normally limit their attention to conceptual processing: planning messages, self-monitoring and discourse interpretation. The fluency of speech production and the 'on-line' character of speech comprehension are essentially based on a combination of automaticity and incremental production.

The staging inherent in the notion of incremental processing is ultimately an empirical issue. And indeed, it has been, and still is a topic of much debate (cf. Garfield 1987). In the ideal case of staging, a later processing component down the line cannot affect the mode of processing of an earlier component. There is only feed forward of information, no interaction between components.

Research in speech production has produced evidence of feedback from phonological to grammatical encoding (Dell 1986; Bock 1987), but this feedback is highly limited (Levelt 1989, 275ff.; Levelt/Schriefers/Vorberg *et al.* 1991). There is at present no hard evidence for any other direct feedback interaction between speech production components.

The picture for speech understanding is less straightforward. Though there is good evidence that a word's acoustic-phonetic processing is not affected by lexical or higher level processes (Frauenfelder/Segui/Dijkastra 1990), the effects of context on word recognition have been heavily debated. It is rather safe now to say that when a speaker hears an ambiguous word (like *bank),* both of its meanings are temporarily retrieved from the lexicon, even if one is ruled out by the context (as in *I withdrew my money from the bank).* (Seidenberg/Tanenhaus/Leiman/Beinkowski 1982). The inappropriate reading is then quickly lost (at least within 200 milliseconds, cf. Jones 1989). In other words, there is no feedback from semantic parsing to lexical access here. The story is more complicated for context effects on the recognition of non-ambiguous words. Using the cross-modal priming technique, Zwitserlood (1989) showed that (i) the structure of the word-initial cohort is independent of biasing semantic or syntactic context, but also that (ii) context begins to have an effect on lexical selection just before the stimulus information has reached the word's uniqueness point (cf., section 3.2 above). This finding contradicts Marslen-Wilson's (1989) claim that not only lexical activation but also lexical selection is exclusively stimulus driven.

As far as grammatical decoding is concerned, it probably involves a relatively autonomous syntactic subcomponent. Whereas the initial, local syntactic analysis proceeds independent of semantic context, global syntactic parsing seems to interact with both semantic and discourse interpretation.

References

Arbib, M.A. (1981). Perceptual structures and distributed motor control. In V. Brooks (Ed.), *Handbook of physiology. The nervous system: Vol. 2. Motor control.* 1449-1480. Bethesda, MD: American Physiological Society.

Bock, J. K. (1987). An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language,* **26**, 119-137.

Bock, J. K. & Warren, R. K. (1985). Conceptual accessibility and syntactic structures in sentence formulation. *Cognition,* **21**, 47-67.

Browman, C. P. & Goldstein, L. (1990). Articulatory gestures as phonological units. *Haskins Laboratories Status Report on Speech Research, SR-99/100.* 69-101.

Butterworth, B. (1980). Evidence from pauses in speech. In B. Butterworth (Ed.), *Language production: Vol. 1 Speech and talk. 155-176.* London: Academic Press.

Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive Psychology,* **4**, 430-477.

Clark, H. H. & Gerrig, R. J. (1984). On the pretense theory of irony. *Journal of Experimental Psychology: General,* **133**, 121-126.

Clark, H.H., Schreuder, R. & Buttrick, S. (1983). Common ground and the understanding of demonstratives. *Journal of Verbal Learning and Verbal Behavior,* **22**, 245-258.

Clark, H. H. & Schunk, D. H. (1980). Polite responses to polite requests. *Cognition,* **8**, 111-143.

Clark, H. H. & Wilkes-Gibbs. D. L. (1986). Referring as a collaborative process. *Cognition,* **22**, 1-39.

Cloitre, M. & Bever, T. G. (1988). Linguistic anaphors, levels of representation, and discourse. *Language and Cognitive Processes,* **3**, 293-322.

Crompton, A. (1982). Syllables and segments in speech production. In A. Cutler (Ed.). *Slips of the tongue and language production. 109-162.* Berlin: Mouton.

Cutler, A. (1989). Auditory lexical access: Where do we start? In W. D. Marslen-Wilson (Ed.), *Lexical representation and process. 342-356.* Cambridge, Mass.: MIT Press.

Cutler, A. & Ladd, D. R. (Eds.) (1983). *Prosody: Models and measurements.* Heidelberg: Springer.

Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production, *Psychological Review,* **93**, 283-321.

Elman, J. L. & McClelland, J. L. (1984). Speech perception as a cognitive process: The interactive activation model. In N. Lass (Ed.), *Speech and language, vol. 10, 337-374.* New York: Academic Press.

Fauconnier, G. (1985). *Mental spaces: Aspects of meaning construction in natural language.* Cambridge, Mass: MIT Press.

Ferreira, F. & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language,* **25**, 75-87.

Fraudenfelder, U. H., Segui, J., & Dijkstra, T. (1990). Lexical effects in phoneme processing: Facilitatory or inhibitory? *Journal of Experimental Psychology: Human Perception and Performance,* **16**, 77-91.

Frazier, L. (1987 a). Structure in auditory word recognition. *Cognition,* **25**, 157-187.

Frazier, L. (1987b). Sentence processing: A tutorial review. In M. Coltheart (Ed.). *Attention and performance XII. 559-586.* Hillsadale, N.J.: Lawrence Erlbaum.

Frazier, L. (1989). Against lexical generation. In W. D. Marslen-Wilson (Ed.), *Lexical representation and process. 505-528.* Cambridge, Mass: MIT Press.

Garfield. J. L. (Ed.) (1987). *Modularity in knowledge representation and natural-language understanding.* Cambridge, Mass: MIT Press.

Garrett, M. F. (1975). The analysis of sentence production. In G. Bower (Ed.), *Psychology of learning and motivation Vol. IX. 133-177.* New York: Academic Press.

Grice, H. P. (1968). Utterer's meaning, sentence meaning and word meaning. *Foundations of Language,* **4**, 25-242.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: 3. Speech acts. 41-58.* New York: Academic Press.

Hörmann, H. (1983). The calculating listener or how many are *einige, mehrere,* and *ein paar (some, several* and *a few)?* In R. Bäuerle, C. Schwarze, & A. von Stechow (Eds.), *Meaning, use, and interpretation of language.* 221-234. *Berlin:* De Gruyter.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge: Cambridge University Press.

Jones, J. L. (1989). Multiple access of homonym meanings: An artifact of backward priming? *Journal of Psycholinguistic Research,* **18**, 417-432.

Kempen, G. & Hoenkamp. E. (1982). Incremental sentence generation: Implications for the structure of a syntactic processor. In J. Horecky (Ed.). *Proceedings of the Ninth International Conference on Computational Linguistics. 151-156.* Amsterdam: North-Holland.

Kempen, G. & Hoenkamp, E.(1987). An incremental procedural grammar for sentence formulation. *Cognitive Science,* **11**, 201-258.

Kempen, G. & Vosse, T. (1989). Incremental syntactic tree formation in human sentence processing: An interactive architecture based on activation decay and simulated annealing. *Cahiers de la Foundation Archives Jean Piagel.* Geneva.

Klatt, D. H. (1989). Review of selected models of speech perception. In W. D. Marslen-Wilson (Ed.), *Lexical representation and process 169-226.* Cambridge, Mass: MIT Press.

Kohn, S. E., Wingfield, A., Menn, L., Goodglass. H., Berko Gleason, J., & Hyde, M. (1987). Lexical retrieval: The tip of the tongue phenomenon. *Applied Psycholinguistics,* **8**, 245-266.

Lackner, J. R. & Tuller. B. H. (1979). Role of efference monitoring in the detection of self-produced speech errors. In W. E. Cooper, E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett. 281-294.* Hillsdale, N. J.: Lawrence Erlbaum.

Lahiri, A. & Jongman, A. (1990). Intermediate levels of analysis: features or segments? *Journal of Phonetics,* 18, 435-443.

Lahiri, A. & Marslen-Wilson, W. D. (1991). The mental representation of lexical from: A phonological approach to the recognition lexicon. *Cognition,* **38**, 245-294.

Lenneberg, E. H. (19967). *Biological foundations of language.* New York: Wiley.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition,* 14, 41-104.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, Mass: MIT Press.

Levelt, W. J. M. (Ed.) (1992). Lexical access in speech production. Special issue of *Cognition, **42**.*

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production. A study of picture naming. *Psychological Review,* **98**, 122-142.

Liberman, A. M. & Mattingly, I. G. (1968). The motor theory of speech perception revised. *Cognition,* **21**, 1-36.

MacWhinnney, B. & Bates, E. (1989). *The crosslinguistic study of sentence processing.* Cambridge: Cambridge University Press.

Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. *Speech Communication,* **4**, 55-73.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition,* **25**, 71-102.

Marslen-Wilson, W. D. (1989). Access and integration: Projecting sound onto meaning. In W. D. Marslen-Wilson (Ed.), *Lexical representation and process. 3-24.* Cambridge, Mass: MIT Press.

Marslen-Wilson, W. D. & Tyler, L. (1980). The temporal structure of spoken language understanding. *Cognition,* **8**, 1-71.

Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language,* **29**, 524-545.

Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language,* **30**, 69-89.

Miller, G. A. & Johnson-Laird, P. N. (1976). *Language and perception,* Cambridge, Mass: Harvard University Press.

Morrow, D. G. (1986). Places as referents in discourse. *Journal of Memory and Language,* **25**, 676-690.

Morrow, D. G. & Clark, H. H. (1988). Interpreting words in spatial descriptions. *Language and Cognitive Processes,* **3**, 257-291)

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review,* **76**, 165-178.

Nicol, J. (1988). *Coreference processing during sentence comprehension.* Ph.D. Dissertation. Cambridge, Mass.: M.I.T.

Nicol, J. & Swinney, D. (1989). The role of structure in conference assignment during sentence comprehension. *Journal of Psycholinguistic Research,* **18**, 5-19.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy,* **3**, 143-184.

Pisoni, D. B. & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition,* **25**, 21-52.

Saltzman, E. & Kelso, J. A. S. (1987). Skilled actions: A task-dynamic approach. *Psychological Review,* **94**, 84-106.

Scherer, K. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin,* **99**, 143-165.

Seidenberg. M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1992). Automatic access of the mentions of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology,* **14**, 489-537.

Seuren, P. M. (1993). Principles of discourse semantics. *Behavioral and Brain Sciences,* **16**

Shattuck - Hufnagel, S. (1979) Speech errors as evidence for a serial order mechanism in sentence production. In W.E. Cooper & E.C.T. Walker (Eds.) *Sentence processing: Psycholinguistic studies presented to Merrill Garrett. 295-242.* Hillsdale, NJ: Lawrence Erlbaum.

Sperber, D. & Wilson, D. (1986). *Relevance: Communication and cognition.* Oxford: Backwell.

Sterberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (Ed.). *Information processing in motor control and learning. 118-152.* New York: Academic Press.

Sternbeg. S., Wright, C. E. Knoll, R. L., & Monsell, S. (1980). Motor programs in rapid speech. Additional evidence. In R. A. Cole (Ed.), *Perception and production of fluent speech. 507-534.* Hillsdale, N. J.: Lawrence Erlbaum.

Stevens, K. N. (1986). Models of phonetic recognition II: A feature-based model of speech recognition. In P. Mermeelstein (Ed.). *Proceedings of the Montreal Satellite Symposium on Speech Recognition,* Twelfth International Congress of Acoustics.

Stevens, K. N. & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P.D. Eimas & J. L Miller (Eds.), *Perspectives on the study of speech. I-38.* Hillsdale; Lawrence Erlbaum.

Taft, L. (1984). *Prosodic constraints and lexical parsing strategies.* Ph.D. Thesis. University of Massachusetts.

Tyler, L. & Warren, P. (1987). Local and global structure in spoken Language comprehension. *Journal of Memory and Language,* **26**, 638-657.

Van Wijk, C. & Kempen, G. (1987). A dual system for producing self-repairs in spontaneous speech: Evidence from experimentally elicited corrections. *Cognitive Psychology,* **19**, 403-440.

Zue, V.W. (1986). Models of speech recognition III. The role of analysis by synthesis in phonetic recognition. In P. Mermelstein (Ed.), *Proceedings of the Montreal Satellite Symposium on Speech Recognition.* Twelfth International Congress on Acoustics.

Zwitserlood, P. (1989). The effects of sentential-semantic context in spoken-word processing. *Cognition,* **32**, 25-64.