Levelt, W. (1999). A Blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), The neurocognition of language (chapter 4). Oxford Press.

4 *Producing spoken language: a blueprint of the speaker*

Willem J. M. Levelt

4.1 Design by evolution

The ability to speak is one of the basic ingredients of human life. We are social animals, deeply caring for the cohesion of our closest kin and for harmony in our daily personal contacts. From this perspective, the copious time idled away on chatting and gossiping is well spent. In all cultures, human bonding is largely achieved and maintained through speech. This is, clearly, species specific. Our closest relatives in nature, the Old World primates, regulate much of their bonding by way of grooming. And they don't stint on it, just as we don't stint on conversation: there are baboons that spend no less than 20% of their waking day on grooming. Dunbar (1996) showed that the amount of time devoted to social grooming is directly related to group size. How much time should *Homo sapiens* be spending on grooming if we had continued that linear trend? That depends on estimations of our typical group size. Hunter–gatherer societies are characteristically partitioned in clans of about 150 persons; in a clan all members know one another. The same number seems to hold for the first agricultural settlements. On this estimate, we should be grooming about 40% of our waking day in order to maintain group cohesion. That would be excessive, especially for an ape with so little fur. Dunbar argues that here the other pre-existing communicative system, the vocal one, began to accumulate increasing functionality in the management of social cohesion, ultimately developing into language. Speech, after all, is so much more effective in transmitting the intentions and motivations that shape our social mesh than is grooming. Chatting is not limited to dyads; it can be practised in larger groups. Talking is information sharing. The 'aboutness' of language enables us to jointly attend to the current state of coalitions and conflicts, to the intentions and deceptions of those present or absent. And, inherited from the old vocal call systems, the prosody of speech is richly expressive of emotion. We can only guess what the many intermediate evolutionary steps have been that bridge the enormous gap between the vocal call systems of Old World primates and the speech/language ability of our species. But

there have been two landmark developments. First, the development of supralaryngeal articulation under neo-cortical control. As Ploog (1990) and Müller-Preuss and Ploog (1983) have shown, primate call systems are largely controlled by caudal midbrain structures; they are directly expressive of the animal's emotion, such as fear, aggression, alarm, contact seeking. The only neocortical input is from the (limbic) anterior cingulate gyrus. The latter makes calling marginally conditionable, as Sutton *et al.* (1974) have demonstrated in macaques; amplitude and duration of innate calls are to some extent malleable. Speech, however, is fully under neocortical control. Larynx, pharynx, tongue, and lip movements in speech are controlled by left and right primary motor cortex, which is an evolutionary novelty. In addition, the function of the supplementary/premotor area became vastly expanded as a repository of articulatory gestural programmes.

The old call system is largely one of phonation, involving the modulation of vocal fold activity. This prosodic–emotional call system became overlaid with a rich supralaryngeal system of modulation in the time/frequency domain, involving pharynx, tongue, oral and nasal cavities, and lips. MacNeilage (1998) argued that this articulatory control developed from pre-existing ingestion-related cyclicities such as chewing, sucking, licking, which attained communicative significance as tongue and lip smacks, etc. The resulting ability to articulate in rhythmic, syllabic patterns is at the heart of all spoken languages. It allows us to pack an elaborate code of temporally overlapping distinctive information from multiple sources (such as glottis, pharynx, velum, and oral cavity) into the time/frequency domain (Liberman 1996).

This first landmark development involves the evolution of a rich species-specific articulatory system, which can function under intentional control. The old vocal system is not lost, but integrated. Prosody keeps being expressive of emotion, controlled by the limbic system. But, in addition, we have direct control over the voice from the larynx motor area. It not only allows us to sing, but also to do such things as feigning emotion in speech.

The second landmark development in evolution is one of social competence. The emergence of Theory of Mind. One of the most noticeable differences between human brains and those of other primates is the much larger relative size of neocortex in man. Still, there is no obvious ecological variable (such as size of territory) that can account for this difference. Dunbar (1996) found one surprisingly reliable predictor of relative neocortex volume : group size. The human data nicely fit the general log/log trend. This invites the interpretation that a major function of neocortical expansion in hominids has been to refine social competence. And, indeed, the vast neocortical areas in our brains dedicated to person recognition (face, voice), to the recognition of intention (facial expression), and to the processing of speech and language seem to support that interpretation. How has *Homo sapiens* dealt with the ever growing social complexity of its clan? It was not enough to interpret actions of group members as intentional, as goal directed. This ability we share with chimpanzees. But in order to make intentional behaviour predictable and malleable, we developed the ability to interpret that behaviour as caused by beliefs, wishes, hopes, that is in terms of mental states that we attribute to the agents around us. In Premack and Woodruff's (1978)

terms, we acquired a 'Theory of Mind' (ToM). Since Wimmer and Perner's (1983) germinal paper on this issue, a flood of research has demonstrated that already at the age of four, but probably earlier, children do attribute beliefs, wishes, and fears to others in order to explain and predict their behaviour. In contrast, chimpanzees show no more than rudiments of this ability (see Bogdan 1997 for a review). ToM allows us to build up complex knowledge structures about our social environment. Over and above registering Who did What to Whom, we encode such complex states of affairs as 'A knows that B did X', 'A believes B did X', 'A hopes B does X', 'A fears that B does X', 'A erroneously believes that B did X', but also 'A believes that B knows X', 'A doesn't know that B hopes X', and so on. And we act on such knowledge, as appears from our remarkable ability to cheat, feign, mislead, and lie.

These two landmark developments are still reflected in the ontogenesis and design of our speech producing system (Levelt 1998). There is, on the one hand, the innate articulatory system. It begins to mature around the seventh month, when the infant utters its first exemplars of repetitive and alternating babbles. Babbles are simple syllables and initially they are not specific to the mother tongue. In fact, even deaf children have a short, transient babbling period. But in the next four or five months, children build up a quite elaborate syllabary, that is increasingly tuned to the syllable repertoire of the native language (De Boysson-Bardies and Vihman 1991; Elbers 1982). On the other hand, there is the very early development of social competence. Like the perception of causality (Leslie and Keeble 1987), the perception of intentionality already matures during the first year of life and, as mentioned above, ToM is up and running by the age of four (Premack and Premack 1995). But what is most remarkable is that these two competences initially mature independently. The elaborate system of social and physical knowledge that the infant acquires during the first year of life simply doesn't interact with the maturation of the syllabary. Babbles are, initially, devoid of any meaning. It is purely articulatory–motor activity, reinforced by auditory feedback. The initially diffuse state of this articulatory system appears from the floundering of arms and feet that accompanies all babbling. It takes months before these motor systems become independently controllable. There is, apparently, enormous plasticity here. As Petitto and Marentette (1991) have shown, deaf children of deaf, signing parents develop 'hand babbling' during roughly the same period. In the absence of auditory feedback, gestural feedback stimulates the adjacent motor system to take over.

It is only around the age of 12 months that first, hesitant links are created between the articulatory and meaning systems. First spoken words are probably 'borrowed' from already established meaning relations in the auditory domain. As Elbers (1982) has shown, first spoken words are usually pre-existing babbles that resemble already meaningful spoken words in the infant's perceptual repertoire.

Even after the two systems become increasingly linked during the second year of life, their further development is controlled by system-internal pressure in the first place. When the articulatory system has acquired some 50 different proto-words, the child slowly but surely gets overwhelmed by keeping ever more similar articulatory patterns apart. The fascinating solution is to 'phonologize' the initial lexicon (C. Levelt 1994).

The child begins to focus on initial, final, and middle parts of proto-words, freely varying their place and manner of articulation. This creates a rich segmental/featural bookkeeping system which allows us to keep apart unlimited amounts of spoken word patterns. In other words, the articulatory system becomes bipartitioned into something like the original syllabary, a repository of articulatory–motor gestures, and a generative phonological coding system for keeping the record.

Similarly, the semantic system begins to get overtaxed during the third/fourth year of life. The child's initial multiword utterances easily express the focused semantic relations (who does what to whom, who possesses what, etc.) by word order; usually a functor word plus one or two argument terms will do. But inevitably, the child's messages become ever more complex. The emergence of ToM probably plays a major role here. There is, first, an increasing awareness of what information is shared with the interlocutor and what not. Not only focused, but also non-focused arguments may need expression; the child's utterances become less elliptical. Second, there is increasing similarity of semantic roles to be expressed in the same utterance. To express *A thinks that B knows X*, the roles of A and B are very similar; they are not easily mapped on the old agent/action type word order. The, again fascinating, development here is the 'syntactization' of semantics. Semantically similar roles are all mapped onto a very lean system of syntactic categories (nouns and verbs, and their modifiers, adjectives, adverbs to start with), and each word gets a (language-specific) syntactic frame, specifying how semantic roles should be assigned to various syntactic functions and allowing for the expression of recursive states of affairs that are so typical for social conceptualizations. Like the articulatory system, the semantic system becomes bipartitioned. Syntax develops as 'the poor man's semantics' for the child to systematize the expression of semantic roles, just as phonology is 'the poor man's phonetics', a lean system for keeping track of the subtle infinitude of articulatory patterns.

These two bipartitioned processing systems play drastically different roles in speech generation. The semantic/syntactic system is there to map the conceptualization one intends to express onto some linear, relational pattern of lexical items ('lemmas'), a 'surface structure', for short. The function of the phonological/phonetic system is to prepare a pattern of articulatory gestures whose execution can be recognized by an interlocutor as the expression of that surface structure, and hence of the underlying conceptualization. I will call it the 'articulatory score'. Although the skilled adult speaker normally shows fluent co-ordination of these two underlying systems, the rift between them never disappears entirely, as I will discuss in subsequent sections.

4.2 The blueprint

The pair of bipartitioned systems emerging from evolution and ontogeny form the core of the adult speech producing apparatus. They figure centrally in the 'blueprint of the speaker depicted in Fig. 4.1. From top to bottom the processing components (rectangles) perform the following functions:

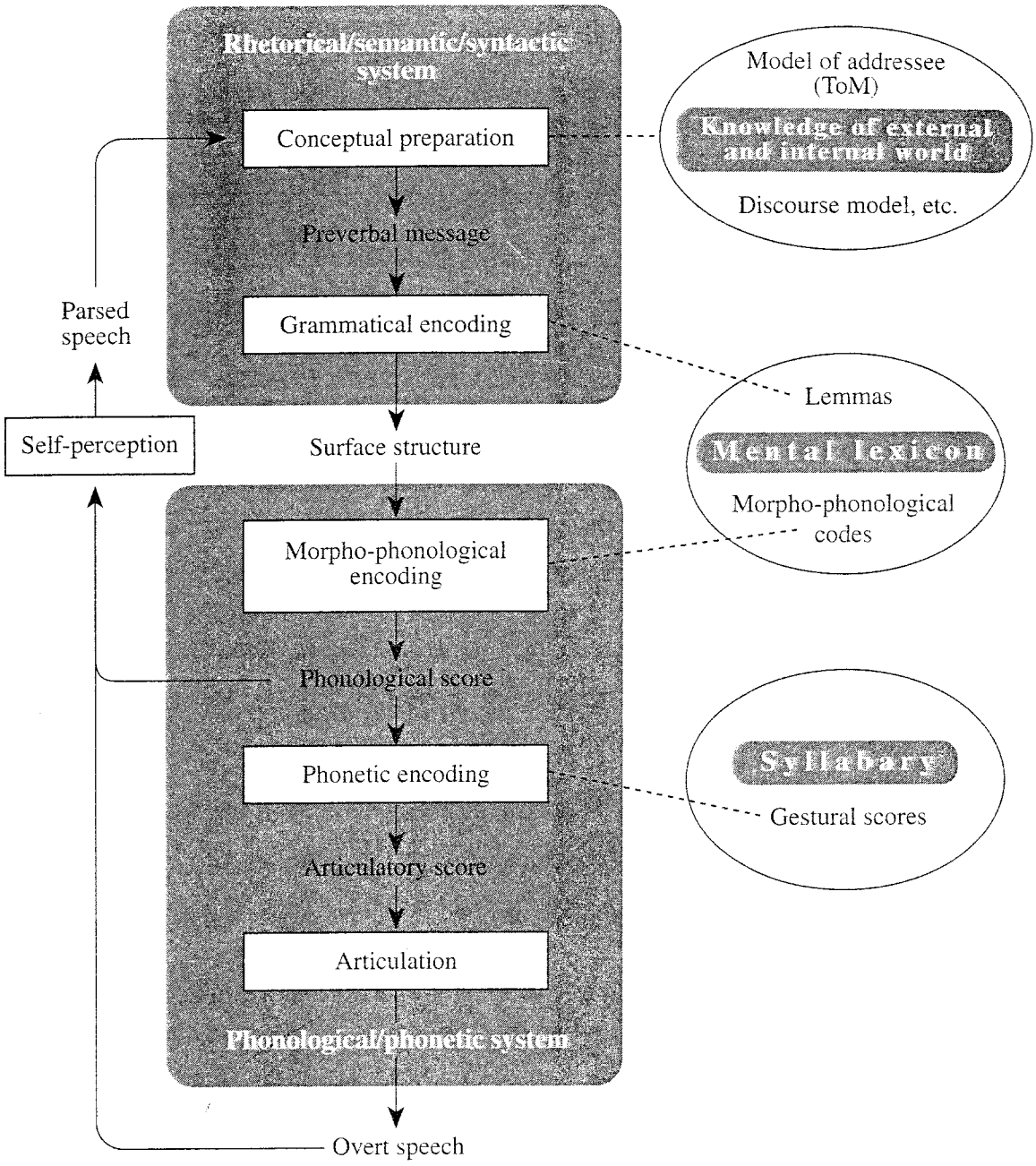


Fig. 4.1 A blueprint of the speaker.

Conceptual preparation Alone, or interactively with the interlocutor, the speaker generates a message, whose expression may affect the interlocutor as intended. Messages are conceptual structures of the kinds described above. In preparing a message, we exercise our social competence, minding the knowledge shared with our interlocutors, directing their attention to what is new or relevant, etc. This is accomplished by skilfully accessing various knowledge sources (knowledge sources are diagrammed as ellipses). The ultimate message is a conceptual structure, consisting of lexical concepts, that is concepts for which there are words in the language.

In this respect the message is more specific than just any conceptual structure. Not all concepts that we can entertain are lexical (think of a dead tree). But a message must eschew those, because it must be expressible in words. This is captured in the term 'preverbal message'.

Grammatical encoding The lexical concepts in the message will activate the corresponding syntactic words ('lemmas') in the mental lexicon. Their selection makes the syntactic frames available that should correspond to the semantic functions and arguments in the message. In grammatical encoding, the speaker uses this lexical–syntactic information to build up the appropriate syntactic pattern, the 'surface structure'. And this is roughly done incrementally, that is 'from left to right'. This completes the processing of the first core system.

Morpho-phonological encoding As soon as a lemma is selected, its form code becomes activated. The speaker gets access to the item's morphological and phonological composition. This is the basic material for building up phonological words. In particular, it is used to generate a word's syllabification in its syntactic context. For instance, the word *comprehend* is syllabified differently in the phrase *I-com-pre-hend* than in the phrase *I-com-pre-hen-dit*. In phonological encoding, the 'phonological score' of the utterance—its syllabified words, phrases and intonation pattern—is built up incrementally, dogging the steps of grammatical encoding.

Phonetic encoding Each of the syllables in the phonological score must trigger an articulatory gesture. Here we finally reach the repository of syllabic gestures that the infant began to build up by the end of the first year of life. Sometimes new or infrequent syllables have to be composed, but mostly speakers can resort to their syllabary. Phonetic encoding is the incremental generation of the articulatory score of an utterance.

Articulation The execution of the articulatory score by the laryngeal and supralaryngeal apparatus ultimately produces the end product: overt speech.

Self-perception When we speak we monitor our own output, both our overt speech and our internal speech. This output monitoring involves the same speech comprehension system that we use for listening to others (see Cutler and Clifton, Chapter 5). If we notice trouble in the speech we are producing, in particular trouble that may have communicative consequences, we can stop and correct ourselves.

This blueprint has a dual function. It is, first, a way of framing of what can be called a basic consensus in the language production literature. There is not much disagreement among researchers about the existence of such mechanisms as grammatical or phonological encoding. Neither is there much disagreement about the general flow of information from component to component. In particular, the notion of *incremental production* (Fry 1969; Garrett 1976; Kempen and Hoenkamp 1987) is generally accepted. It says that the next processing component in the general flow of information can start working on the still incomplete output of the current processor. A processing component will be triggered into action by any *fragment* of its characteristic input. As a consequence, the various processing components are normally simultaneously active, overlapping their processing as the tiles of a roof. When we are uttering a

phrase, we are already organizing the content for the next phrase, etc. There are, certainly, many disagreements about details of the organization. This holds in particular for the amount and locus of feedback and interaction among components. But this doesn't affect the consensus on the general architecture of the system.

The second function of the blueprint is to frame a research programme. The ultimate aim of this research programme is to explain how we speak. The agenda can be read from the blueprint. We will have to produce and empirically test working models of the various functions performed by the speaker. How does grammatical encoding work? How does morpho-phonological encoding work? And so on. Also we will have to produce accounts for how the various processing components co-ordinate their activities in the generation of fluent speech. One thing should be clear about this research programme. Its advance will be measured by how well we succeed in producing empirically viable working models for smaller or larger aspects of the main processing components involved.[1]

In the following sections, I will discuss the various component functions in the above order, without losing sight of the ultimate purpose of the system, to map communicative intentions onto fluent speech.

4.3 Conceptual preparation in context

It is one thing to claim that language evolved for the management of cohesion in ever larger groups of humans, but quite another thing to specify in detail how that function is exercised in actual language use. In fact, that problem is horrendously complex, just as complex as the myriad linguistic transactions we perform in modern society. It cannot be the purpose of a working model to account for this complexity, just as it cannot be the purpose of a theory of thermodynamics to predict the weather. Still, advances in the analysis of language use provide an important sounding board for theories of speech production. The one major recent publication on language use, Clark (1996), analyses language use as a form of joint action. Participants in joint activities are aware of some goal of the activity and of their common ground. (In the above terms: they exercise their ToM to monitor the mutually shared state of information.) A production model should at least be 'on speaking terms' with core aspects of the QJ;co-ordination of action, such as details of turn-taking, managing politeness, inviting or initiating repair. A speaker's decision *what* to say, in our terms the speaker's *message*, should be understandable in terms of the current state of joint action.

The recent advances in the analysis of language use are, regrettably, not matched by similar advances in working models of conceptual preparation. In fact, the situation is hardly different from the state of affairs sketched in Levelt (1989). The progress has mostly been in the engineering of natural language generation, the development of models for artificial text generation (see, for instance, Pereira and Grosz 1994). Here I will only present a bare minimum of machinery that should go into the development of any working model, relating to the two core processes in the conceptual generation for speech: *macroplanning* and *microplanning*.

4.3.1 Macroplanning

This is the process by which the speaker decides what to say next. A working model will, of course, not deal with speakers' potential topics of discourse (see the *Encyclopaedia Britannica* for a short list). It will rather implement general principles of how subsequent moves within and between participants are sequenced. The central notion here is *discourse focus*. Given the communicative intention, the speaker will focus attention on something specific to be expressed (the 'current focus'). In moving to the next focus, the speaker's ToM is at work. The speaker will, normally, try to guide the intended focus shift of the interlocutor. Focus shifting is attention management at two levels. First, the speaker will monitor whether what should be said for realizing the communicative intention will be said. Second, the speaker will monitor whether the interlocutor is following the speech act.

The management of attention can be represented by the combination of a 'focus tree' (McCoy and Cheng 1991) and a stack. An overly simple example is presented in Fig. 4.2. When a speaker has as a goal to inform an interlocutor about the layout of the figure in the left panel, starting at the star, his focus tree may develop as shown in the right panel (in fact, it *will* develop that way, as numerous experiments have shown, cf. Levelt 1989). The ensuing text will, for instance, be:

> There is a star at the bottom. It has a line connection straight up to a triangle. From the triangle there is to the left a line to a square. Back to the triangle, there is a connection to the right to a circle. And the circle connects straight to the right to a diamond. That's it.

The focus tree has STAR at the top, the first focus of the speaker. The speaker's attention then moves to TRIANGLE. The speaker formulates how the triangle is placed with respect to the star. Now, the speaker has a choice, turning to the square or to the circle. The square is attended to first, but the speaker should keep in mind that a return must be made to the triangle. Hence, TRIANGLE is put on the stack. After mentioning the square, there are from there no further connections to attend to.

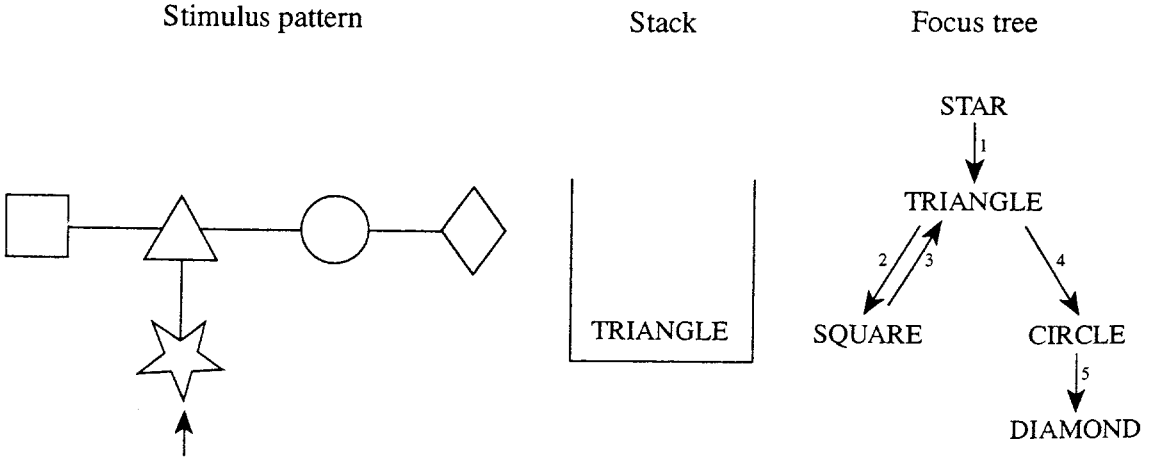


Fig. 4.2 Focus tree and stack for the description of a visual pattern.

The stack pops up TRIANGLE, and the speaker proceeds by describing the right branch of the figure.

What kind of principles govern the construction of focus trees? They are well-known for the description of networks such as the example case here (and much more complex ones). For that domain focus shift follows just three principles: *connectivity* (focus as the next item one that is connected to the currently focused item), *stack* (in the absence of a connecting item, turn to the top of the stack), and *simplest first* (if there is a choice, attend to the simplest item first). It is easy to see that the above description is predictable from these three principles. The working model can be found in Levelt (1982). Other types of discourse involve more and different principles, but hopefully it is a finite set (see especially Hovy 1994). As in the example, it is often the case that attention moves over a set of subgoals that must be fulfilled in order to realize the 'grand' communicative intention. Also, we will normally make an effort to guide the attention of our listeners in such a way that they can make the corresponding connections. For instance in the above description the phrase *Back to the triangle* is essential given the state of the listener's mental model—it was not yet known that TRIANGLE had become stacked. In fact, speech partners often intrude for clarification, redirecting attention to other parts of the focus tree, or inviting the growing of entirely new branches.

4.3.2 Microplanning

Conceptual preparation involves more than deciding what to say and in what order. Each bit of information needs further shaping in order to be formulated. Remember that the message is a particular kind of conceptual structure. In order for it to be expressible in words, its terminal elements must be lexical concepts. Also, it should incorporate the kind of semantic relations that are expressible in language, in particular function/argument and modification relations. Many conceptual structures don't have these properties. If they are focused for expression, we must somehow cast them in propositional form. Let us consider another spatial example, the state of affairs depicted in Fig. 4.3. Assume we intend to inform an interlocutor about this scene. Here are two of many possible descriptions:

(1) There is a house with a tree to the left of it.

(2) There is a tree with a house to the right of it.

In the first description, the position of the tree is related to that of the house; in the second description it is the other way round. But notice that the spatial scene itself is entirely neutral with respect to what should be related to what; it is the speaker's free choice to do it one way or another. The important point here is that *some* choice should be made. The speaker *must* take some perspective on the scene in order to express it in language. It should be cast as a propositional relation and the two options discussed here are LEFT (TREE, HOUSE) and RIGHT (HOUSE, TREE). The speaker may have pragmatic reasons for taking the one perspective rather than the other. For instance, a previously described scene showed a house with a man to the left of it.

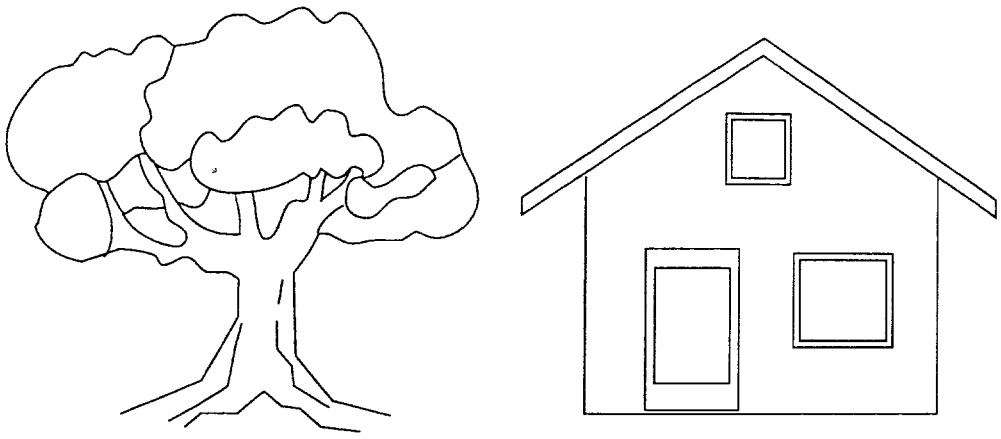


Fig. 4.3 Two different conceptualizations of a visual scene.

In that case description (1) is more appropriate to focus the listener on the difference with the previous scene (and the speaker will stress 'tree').

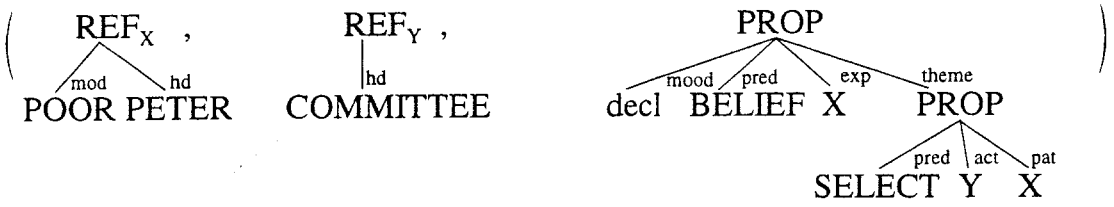
But there is more to perspective taking. There is also freedom in the choice of the lexical concepts that go into the propositional relation. That is easily demonstrated from still another description of the same scene:

(3) There is a house with a tree to the right of it.

How can both (1) and (3) be veridical descriptions of the same scene? Does 'left' mean 'right'? No, it doesn't. The difference is in the kind of perspective the speaker takes. For description (1) the speaker takes so-called 'deictic' perspective, which is a relation between the perceiving speaker, the relatum (the house), and the referent (the tree). From the speaker's vantage point, the tree is to the left of the house. But for description (3) the speaker takes 'intrinsic' perspective. The relatum (the house) has an intrinsic orientation. It has, in particular, a front and a right and a left side. The tree is on the house's right side and this holds independently from the speaker's point of view. Hence, the same spatial relation between relatum HOUSE and referent TREE can be veridically expressed in terms of two converse lexical concepts, LEFT and RIGHT. And again, the speaker may have good pragmatic reasons for taking one or the other perspective (see Levelt 1996 for a full analysis).

In considering this example, we have not been dealing with a peculiar property of the terms 'left' and 'right', or of spatial descriptions in general. Rather, the example demonstrates an entirely general property of conceptual preparation for speech. Whatever the information to be expressed, there is always perspective taking. The information must be cast in propositional form (see below) and in terms of pragmatically appropriate lexical concepts. I can express the same kinship relation as *John is Peter's father* or as *Peter is John's son*; it will depend on what I want to focus as the new information for my listener. Also, I can refer to the same person as *my brother*, *my neighbour*, *my colleague*, etc., depending on whichever of my relations to the referent I want to highlight for my listener. Perspective taking is at the very core of all conceptual preparation for speech (Clark 1997).

What exactly is the propositional format of a message? There are various proposals in the literature (see, for instance, Levelt 1989; Zock 1997; Kempen 1999). The choice largely depends on the details of one's computational theory, which is not at issue in this chapter. But the information that goes into a message is essentially of four kinds, which can be exemplified from the speaker preparing the following utterance: *Poor Peter believes that the committee selected him.* This utterance is, first, about particular referents, namely Peter and the committee. The message should specify the referents and link them to the relevant 'state of affairs', that is in the discourse model. Second, there is some predication made about these referents (in particular that Peter believes something, namely that the committee selected him, Peter). We call this 'argument structure'. Arguments fulfil 'thematic roles' in the predication. Peter, for instance, is the *experiencer* of believing, and the *patient* of selecting. Other roles are *agent* (the one who causes something to happen), *actor* (the one who does something), *theme*, *source*, and *goal* (as in *the ball rolled from the chair to the table*), etc. Third, there may be specifications or modifications in a message. In the example message, Peter is further specified or modified as being pitiful or poor. An important kind of specification is quantification. A speaker could, for instance, refer to some apples or to all cows. Fourth, each message has a *mood*. It can be declarative, imperative, or interrogative. It is declarative when the speaker intends to assert something; it is imperative when the speaker wants to express the desirability of some state of affairs, and it is interrogative when the speaker wants to invite the interlocutor to provide some specific information. There is more that goes into a message (cf. Levelt 1989 for a fuller treatment), but this suffices for the present purposes. So, for *Poor Peter believes that the committee selected him*, the underlying message is something like this:[2]



It says that there are two referents (X and Y), which are the arguments or thematic roles in a complex declarative proposition, where the predicate BELIEF has as experiencer X (*poor Peter*) and as theme argument the proposition that Y (*the committee*) selects X (*poor Peter*).

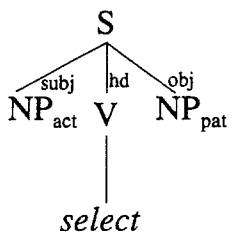
One final aspect of microplanning should be mentioned. Conceptual preparation is not language-independent. Languages differ, first, in their range of lexical concepts. Tzeltal, for instance, has no lexical concepts for LEFT and RIGHT, but only a superordinate concept TRAVERSE. A Tzeltal speaker's perspective taking for expressing a scene such as the one in Fig. 4.3 will therefore be different from that of a native Dutch or English speaker. Second, languages differ in the conceptual information that is *obligatorily* expressed. In a tense-marking language, such as English,

the speaker must always think of the temporal properties of a state or event before expressing it. It is not enough for a speaker of English to prepare the above example message, because grammatical encoding will block on tense assignment. Should it become *Poor Peter believed the committee selected him* or *Poor Peter believes the committee selected him* or *Poor Peter believes the committee will select him*, etc? The speaker should mark the prevailing temporal relations (such as 'past') in the message, whether or not it is of any communicative significance. Speakers of Chinese or Javanese do not carry that conceptual burden, because their languages are not of the tense-marking kind. Slobin (1987) usefully called these language-dependent aspects of conceptual preparation 'thinking for speaking'.

4.4 Grammatical encoding

The blueprint in Fig. 4.1 depicts three properties of grammatical encoding: it takes preverbal messages as input, it produces surface structures as output, and it has access to the mental lexicon. Surface structures are syntactic in nature. They have a 'left-to-right' ordering of syntactic words ('lemmas' for short, such as nouns or verbs) that is incrementally generated from the emerging preverbal message. These lemmas are not evenly spread, but tend to be grouped in smaller or larger phrases. If a phrase contains a tensed verb, we call it a clause. Languages differ markedly in the kinds of syntactic relation they encode in a surface structure, but 'subject of' or various kinds of 'object of' are popular. One should not forget that syntax is the poor man's semantics. There are obviously different ways in which similar thematic role structures can be mapped onto a small number of syntactic relations. Also, languages differ in how they encode syntactic relations. Some languages, such as English, prefer to encode them in terms of phrasal relations and order relations within a sentence. Other languages prefer to mark lemmas in the surface structure for their syntactic function. Neither order nor hierarchy matter much, which leaves these features of surface structure available for pragmatic functions (such as directing the hearer's attention to particular elements in the sentence).

Whatever the differences between languages, the generation of surface structure is, for a large part, lexically driven. This means that in grammatical encoding a major operation is this: a lexical concept in the message (for instance SELECT in the above example) activates the corresponding lemma (*select*) in the mental lexicon. Upon its selection, the lemma's syntactic properties become available for further syntactic construction. The syntax of the lemma *select* is something like this:



It says that *select* is a verb that should be the head of a sentence; it should have a subject NP and an object NP. Also, it specifies how these NPs should correspond to the thematic roles in the concept SELECT: the subject NP should link to the *actor* role in the message and the object NP to the *patient* argument.

Each lemma is the terminal node of such a syntactic tree and grammatical encoding consists of connecting these retrieved syntactic trees to form a surface structure that matches the input message. In a way grammatical encoding is like solving a set of simultaneous equations. Each lemma requires particular syntactic constraints from its environment and the emerging syntactic structure should simultaneously satisfy all these constraints.

But the mental lexicon contains more than just single-word lemmas. Some lexical concepts, or rather 'idiom concepts' map onto idioms of one kind or another. Idioms such as *to throw in the towel* are encoded by going from a single concept to a complex idiom lemma with its own syntactic properties. For instance, *to throw in the towel* is a verb lemma, but it doesn't allow for passivization (Jackendoff 1997). Probably, the amount of idiom and collocation in the mental lexicon is of the same order of magnitude as the number of words (a good source on idiom is Everaert *et al.* 1995).

Given that grammatical encoding is largely lexically driven (in this broader sense), I will first discuss lemma selection and then turn to further syntactic composition.

4.4.1 Lemma selection

Recent years have seen important progress in the theory of lemma access. Levelt (1989) still painted a bleak picture of inadequate theories, that all run into the so-called *hyperonym problem*. When the semantic conditions are met for selecting some lemma (for instance *horse*), the selection conditions are also met for selecting all of its hyperonyms (such as *mammal*, *animal*). But that hardly ever happens. Roelofs (1992, 1993) proposed a new model of lemma selection that does not run into this problem, and that also accounts for a wide range of old and new reaction time results. Meanwhile the computational model, now called WEAVER, has been extended to incorporate morpho-phonological encoding as well (Roelofs 1997*a*,*b*). Together, these developments have given us a new handle on the production lexicon. A comprehensive statement of this new theory of lexical access and its empirical foundations can be found in Levelt *et al.* (1999). Here I will present a small fragment of the production lexicon as modelled in WEAVER and then discuss how lemma selection is handled in the model. In later sections, other aspects of word production will also be discussed in reference to this fragment.

Figure 4.4 presents the lexical item 'select' in the lexical network. At the top, conceptual level the central node represents the lexical concept SELECT with its two thematic role slots X and Y for the one who selects and the entity selected. The semantics of the concept is represented by the set of labelled relations to other

concepts in the network (both lexical and non-lexical ones). For instance, SELECT has CHOOSE as a superordinate concept (to select is to choose from among a number of similar entities), and has ELECT (to select by some democratic procedure) as a subordinate concept. The lexical concepts in the network are connected to the next stratum in the network, the lemma stratum. The concept node SELECT, for instance, is connected to a node at the lemma stratum that represents the lemma *select*. Its syntactic properties are represented by labelled connections to various nodes at this level. The network shows, for instance, that *select* is a transitive verb with two syntactic arguments x and y onto which the thematic roles X and Y should be mapped. In addition, it has a set of (variable) diacritic features (tense, aspect, number, and person) that can get fixed in various ways during grammatical encoding. At this level there are also nodes for all other lemmas, such as for *choose* and *elect*.

Lemma selection is modelled as follows. In the conceptual network, the target concept is in a state of activation. Its activation spreads to all semantically related concepts (for empirical evidence, see Levelt *et al.* 1991). Each active lexical concept also spreads part of its activation down to 'its' lemma, down in the lemma stratum. Lemma selection now becomes a probabilistic affair. During any smallest interval in time the probability of selecting the target lemma is its degree of activation divided by the total activation of all active lemmas in the stratum. This is called 'Luce's rule'. This probability allows one to compute the expected selection latency, which is the prediction tested in reaction-time experiments. An important property of the original model is that it will not make selection errors. The reason is that selection of any lemma must meet the condition that it entertains the correct sense relation to the conceptual level. If *examine* happens to win out by Luce's rule, it will not be selected, because its sense relation is not to SELECT.

The typical reaction-time experiment to test the model is one of picture naming. The subject names a picture of an action (such as a man drinking water) or of an object (such as a dog). But at some moment during the trial a distractor word is presented, either visually (in the centre of the picture) or acoustically as a spoken word. The distractor can be semantically related to the target word (for instance *eat* when the target word is 'drink', or *horse* when the target word is 'dog'), or it can be an unrelated word (such as *work* or *chair*, respectively). These distractors are supposed to activate 'their' lemmas in the lexical network and hence to reduce the Luce ratio. And indeed, reaction latencies are typically longer when there are distractors in the experiment. But the model further predicts that interference should be larger for semantically related distractors than for unrelated ones. Another prediction is that the difference will be maximal when picture and (visual) distractor coincide in time, diminishing with increasing stimulus onset asynchrony. The model gives an excellent fit both for the classical picture/word interference data of Glaser and Düngelhoff (1984) and for myriad new data obtained in further experiments (Roelofs 1992, 1993).

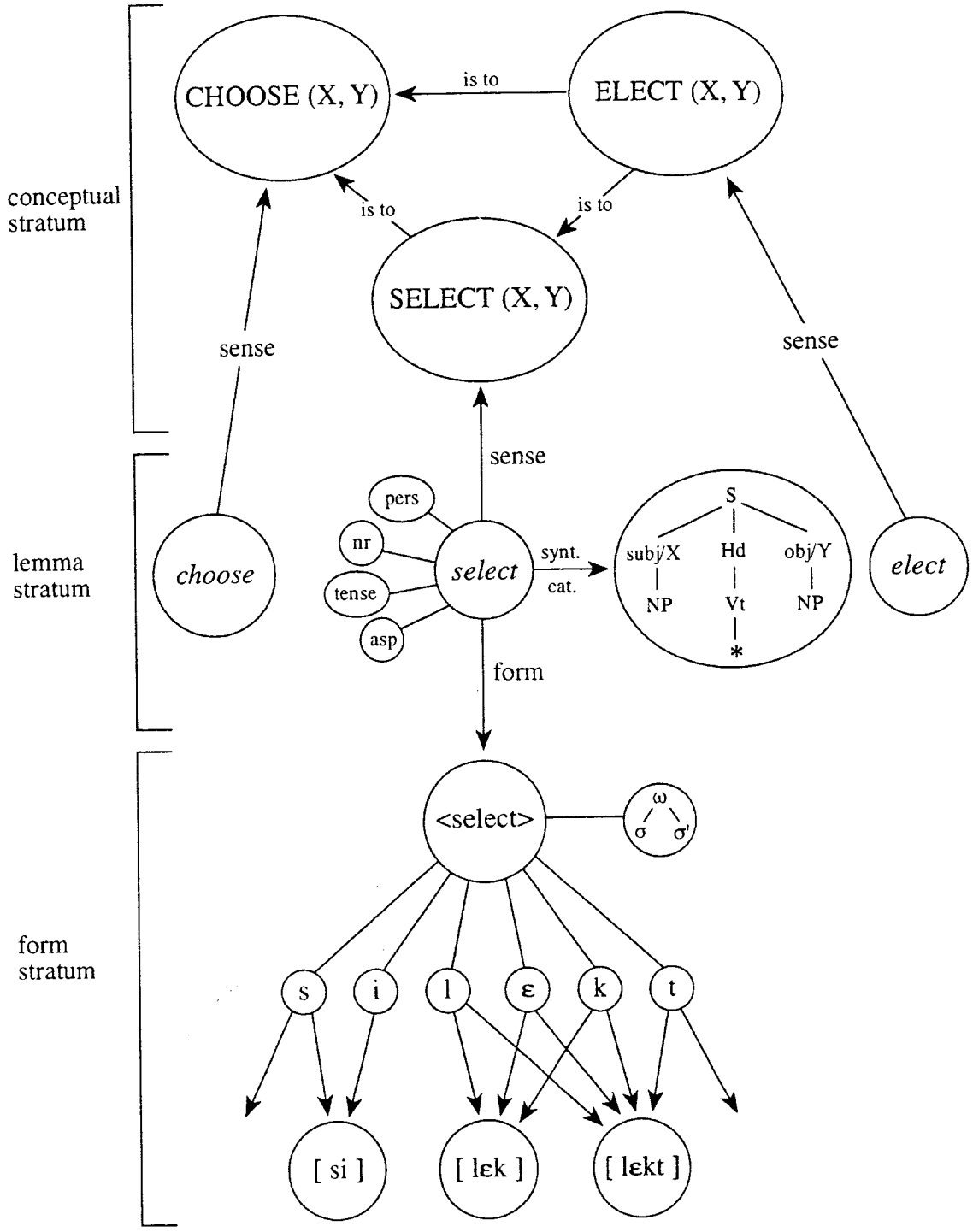
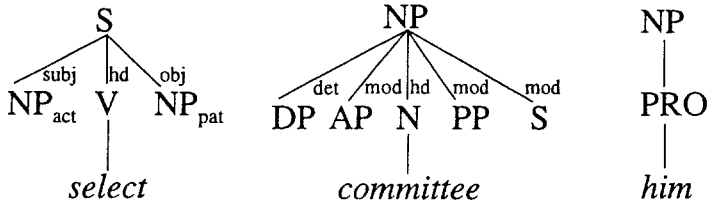


Fig. 4.4 Fragment of a lexical network.

4.4.2 Syntactic composition

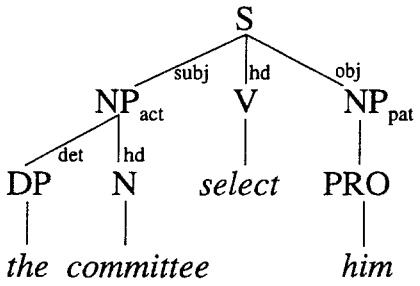
As lemmas become available, triggered by the message, syntactic composition will be initiated. It consists essentially in coupling the syntactic fragments that come with the

lemmas. This process is called 'unification'.[3] Let us consider how the syntactic fragments underlying *the committee selected him* are unified. Three lemmas are active here, *select*, *committee*, and *him*. Here are their syntactic tree fragments:



The tree fragment for *select* was introduced above. It has the node S as a root, and two NP nodes as feet. The syntactic fragment for *committee* is typical for any full noun. It has NP as the root node, which means that it must become the head of a noun phrase, and it has several feet. It allows, in particular, for a determiner phrase (in the present case the determiner will be the definite article *the*, whose selection I won't discuss here—but see Levelt 1989, p. 236 ff.). It can combine with an adjectival phrase (AP), as in *the big committee*, with a prepositional phrase, as in *the committee of the school*, and with a relative clause, as in *the committee that runs the soccer club*. The fragment for *him* is also head of an NP. How does the lemma *him* get triggered by the message? Remember that it refers back to referent X, POOR PETER. In the message, one of the occurrences of argument X will be marked as 'in focus'. That will tell the grammatical encoder that it should select a reduced, pronominal lemma for that occurrence of the lexical concept. Schmitt (1997) demonstrated experimentally that the full noun lemma does get selected in the process of pronominalization. In her model, the 'in focus' feature makes the connected pronoun lemma 'take over'.

Unification now consists in connecting roots to feet. In the example, the root node of *committee* can unify with the first NP foot of *select*. Similarly the root node of *him* can unify with the second NP foot of *select*. Feet that don't receive a unification get trimmed. If all this is done for our example, the following syntactic structure emerges:



But how come that the NP fragment of *the committee* doesn't attach to the second NP foot of *select*? This is because of the linkage between syntactic functions and thematic

roles. In the message COMMITTEE is the thematic actor role. The syntax of *select* requires that the subject NP expresses the actor. Also notice that *him* is the accusative pronoun, not *he* or *his*. When a pronoun unifies with an NP, it inherits the case of that NP. The object NP of *select* carries accusative case.

As everything else in speech production, the generation of syntax is an incremental process. As soon as one or a few fragments of the message, lexical concepts, become available the lemmas get selected and unification begins. The resulting syntax is, therefore, to some extent determined by the order in which lexical concepts come available. Highly accessible concepts tend to come first. In turn, they tend to 'claim' prominent syntactic positions. In particular human and animate referents are often the salient actors or agents in a message. For most verb lemmas these map onto the subject function (as in the *select* example). Less salient or less accessible concepts tend to end up with less prominent syntactic functions, such as direct object, indirect object, or oblique object. For more extensive reviews of grammatical encoding, see Levelt (1989) and Bock and Levelt (1994).

4.5 Morpho-phonological encoding

As lemmas become selected and positioned in the emerging surface structure, their morpho-phonological codes become available to the second main system involved in speech production, a system specialized in generating articulatory scores. Remember that in ontogeny the infant's articulatory system, the beginning syllabary, gets overtaxed when more and more protowords are acquired. The 'phonologization' of the articulatory memory codes solves this problem by providing the child with a discrete generative bookkeeping system for accessing the ever more similar articulatory codes. In the mature speech-producing system the articulatory score is accordingly generated in two steps. The speaker first uses the discrete memory codes to generate a 'phonological score', a score in terms of discrete segments and features, with phonological syllables as its basic units and with a simple hierarchy in terms of phonological words and phrases. Then these syllables are given gestural shape in their phrasal context, usually by retrieving their gestural scores from the old syllabary. It is at this second step that the limbic system still exerts direct control over speech generation. In this section we will consider the first, discrete step in the process.

4.5.1 Generating phonological words

Phonological words are the domains of syllabification. These domains may be larger or smaller than lexical words. For instance, most compound words syllabify per morpheme. Take *popart*, which is syllabified as *pop-art*, respecting the integrity of its morphemes 'pop' and 'art'; here the second /p/ is syllable-final and not aspirated. Compare this to the monomorphemic word *coupon*, which is syllabified as *cou-pon*, with the syllable-initial *p* aspirated. But the domain of syllabification is larger in so-called 'cliticization'. For instance, in the utterance *They will select us for the*

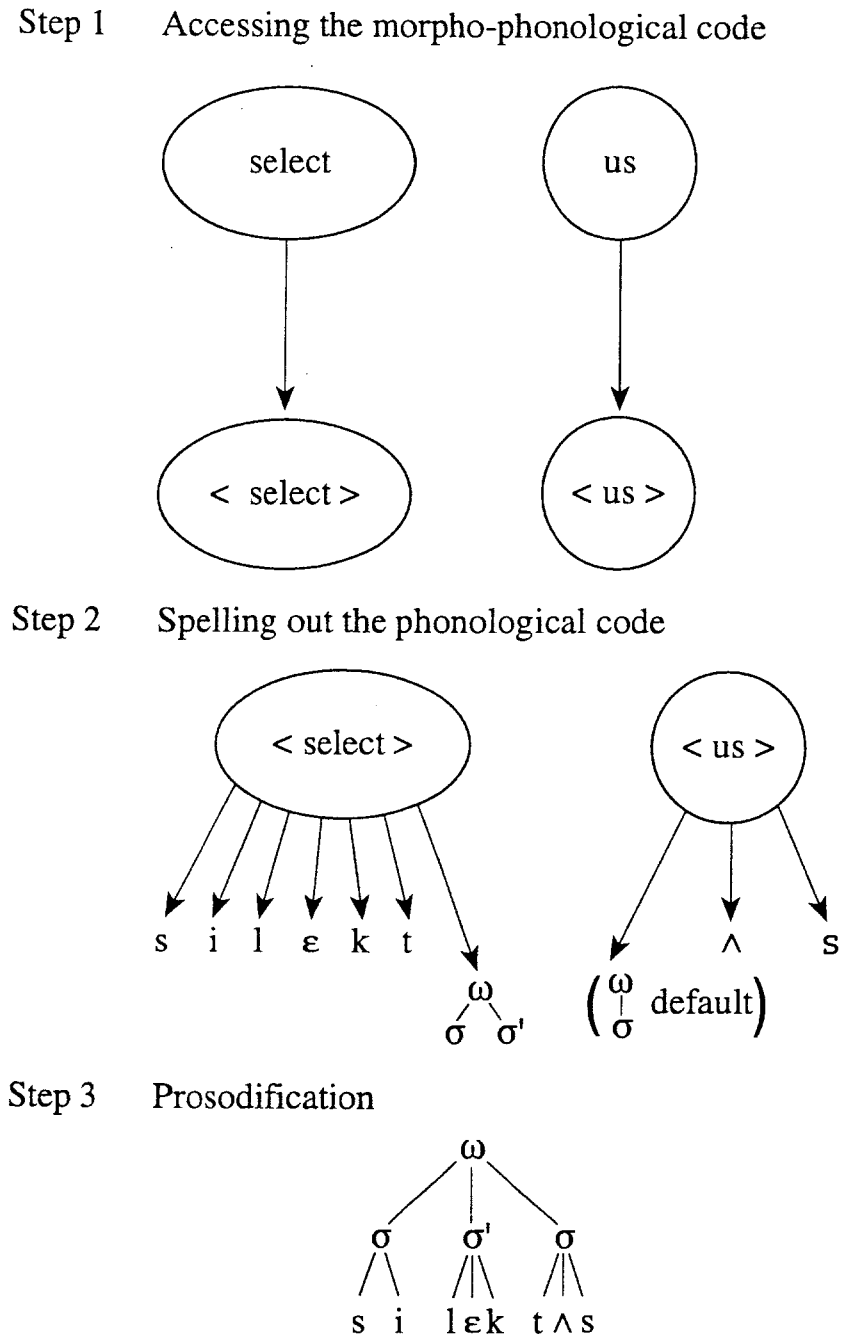


Fig. 4.5 Three steps in morpho-phonological encoding.

competition, the phrase *select us* is syllabified as *se-lec-tus*, ignoring the lexical boundary between *select* and *us*. I will use this latter example to discuss the generation of phonological words in stress-assigning languages such as English or Dutch.

Figure 4.5 presents a schema of phonological word generation. It involves three operations. First, as soon as a lemma gets selected for grammatical encoding, it spreads its activation to its morpho-phonological code in the lexicon. If the word is

multimorphemic, such as 'popart', all morphemes get activated (namely both ⟨pop⟩ and ⟨art⟩). In the case of a monomorphemic word, such as 'select', the morpho-phonological code addressed is a single morpheme, ⟨select⟩ in this case. Second, the phonological code is spelled out. This involves two kinds of information. There is a spell-out of each morpheme's segments. For ⟨select⟩ the spelled out segments are /s/, /ɪ/, /l/, /ɛ/, /k/, and /t/. For ⟨us⟩ they are /ʌ/ and /s/. And there is a spell-out of a word's metrics, except when the metrics has default value. The spelled-out metrics consists of the word's number of syllables and the position of the stressed syllable. For ⟨select⟩ the metrics is σσ'. For ⟨us⟩ the metrics is just σ, but it is not spelled out because it is default metrics. What is default metrics? For stress-assigning languages a word has default stress if stress is on the first full-voweled syllable. For instance, the following words have default stress in English: *post*, *photo*, *marzipan*, but also *arrest*, *cadaver*, *potato*, whose first vowel is pronounced as a schwa. This has been called default metrics (by Meyer *et al.*, in preparation, see also Levelt *et al.* 1999) because most word tokens produced are of that type (85 per cent for English, 91 per cent for Dutch). Third, the spelled-out segments are incrementally grouped into syllables that attach to the spelled-out or composed metrics of the phonological word. It is only at this level of processing that (phonological) syllables appear in speech production. Syllables are not stored in the mental lexicon, because they are highly context-sensitive. For instance, the stressed part of the word 'select' will be syllabified as *lect* in *they will select Peter*, but as *lec* in *they selected Peter* or in *they will select us*. This context-sensitivity has the clear function to create optimally pronounceable utterances (imagine how hard it would be to say *they-se-lect-ed-Pe-ter*).

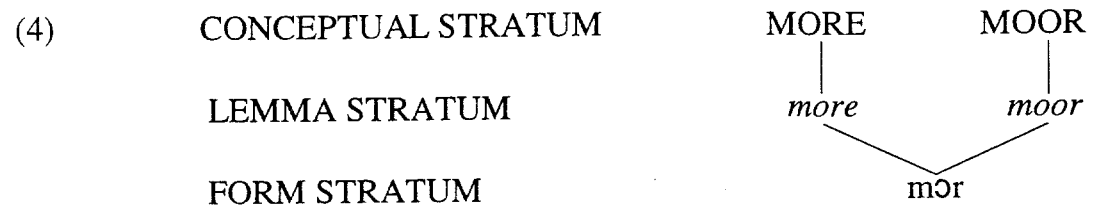
Let us now consider these three steps in somewhat more detail.

4.5.1.1 Accessing the morpho-phonological code

The first step in phonological encoding is most interesting from the neuroscience perspective. It involves 'bridging the chasm' between two evolutionary distinct systems that come to meet during the first few years of life. There are several phenomena in adult speech production that still betray the underlying rift. A first such phenomenon is the so-called word-frequency effect. The phenomenon, discovered by Oldfield and Wingfield (1965), is that pictures with low-frequency names (such as *broom*) have longer naming latencies than ones with high-frequency names (such as *boat*). Wingfield (1968) showed that this was a genuine effect of lexical access; the latency differences didn't show up in picture recognition tests. With the development of a more detailed theory of lexical access (as exemplified in Fig. 4.4), it became important to find out at which stage of lexical access the word-frequency effect is generated. In a series of experiments, Jescheniak and Levelt (1994) showed that the effect is entirely due to word form access. The main problem in that study was to distinguish between the levels of lemma selection and of word form access as possible loci for the word frequency effect. A core experiment involved the production of homophones. Homophones are different words that sound the same. For most dialects of English high-frequency *more* and low-frequency *moor* are

homophones. In our theory, their lexical analysis would be this:

(4)



Now consider the latency of generating low-frequency *moor*. If the word-frequenc effect resides at the lemma level, low-frequency *moor* should have a relatively lon naming latency (i.e. as compared to the latency of high-frequency *more*). If, however the word-frequency effect arises in accessing the word's phonological code /mɔr/, *mor* and *moor* should have the same naming latency and, paradoxically, low-frequenc *moor* should behave as a high-frequency item, because it inherits the accessing speed c its high-frequency twin *more*. In the reaction-time experiments the latter, quite nor trivial result was obtained. Hence, the word-frequency effect arises precisely in th speaker's effort to 'cross the rift' from the semantic/syntactic system to the phonolc gical/articulatory system. In this connection it is particularly relevant that at least pai of the word-frequency effect is in fact an age-of-acquisition effect (Carroll and Whit 1973; Morrison *et al.* 1992; Snodgrass and Yuditsky 1996; Brysbaert 1996). Crossin the rift is apparently easier for words that were acquired early, in part independently c their frequency of usage. These early, more stable connections were established in brain with great plasticity.

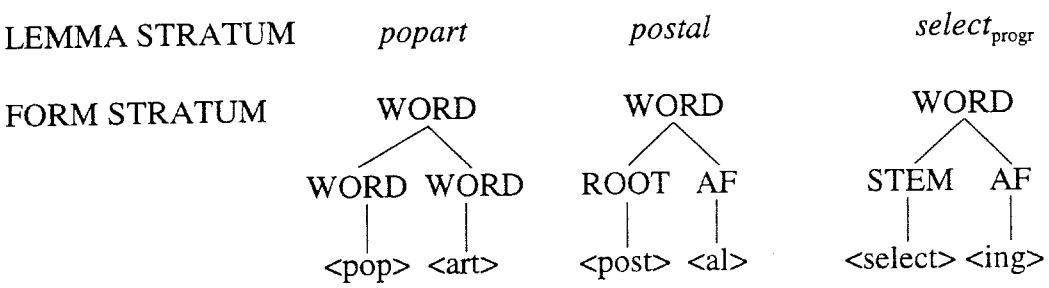
Another well-known phenomenon also emerges at this step. It is the so-called tip-o the-tongue (TOT for short) phenomenon. It can, at any time, happen in spontaneou speech that one suddenly blocks on a name of a person, plant, animal, instrument, c whatever. One knows that one knows the name, and one can even be aware of th word's beginning, stress pattern, or number of syllables. Again, the question is whethe the effect arises at the lemma level or at the level of form access. Levelt (1989) pointe out that if TOT is a problem in accessing the word's form information, the speake should have accessed the word's lemma. For gender-marking languages such as Dutc or Italian this means that in a TOT state the speaker might have access to the gram matical gender of a target noun. This is because gender is a lemma-level syntacti property of a noun. Viggiocco *et al.* (1997), in an elegant series of experiments, hav shown that the prediction is borne out for Italian speakers; the finding was replicate by Caramazza and Miozzo (1997).

A related phenomenon in pathology is anomia. Anomic patients are handicapped i naming objects and they frequently enter TOT states when they speak. Badecker *et a* (1995) tested an Italian patient who could hardly name any pictured object. But in a cases the patient knew the grammatical gender of the blocked target word. Anomia, c at least this particular kind of anomia, is a rupture of the apparently still somewha fragile connection between the two main underlying systems in speech production. C course, the TOT state in both healthy and anomic speakers is an 'off-line' state. Aft

apparent trouble in on-line word access, the speaker is asked to ruminate about the lost word's gender or phonology. We do not know what exactly is involved in these metalinguistic processes; it is certainly premature to draw strong conclusions about the on-line process from whatever emerges in the off-line, metalinguistic state. The only way to find out whether lemma information (such as gender) is retrieved before word-form information (such as word initial segments) is to measure on-line. That is what Van Turennout *et al.* (1998) did in their study of lateralized readiness potential manifestations (LRPs) of gender and phoneme access in retrieving a picture's name. That study showed unequivocally that gender access precedes phoneme access, even in situations where that is disadvantageous to task performance.

All examples so far concerned the access of monomorphemic word forms. But what when a word is multimorphemic, such as *popart*? The present state of our theory is that all multimorphemic words have multiple morpho-phonological codes at the form level. Here are three examples, a compound, a derivation, and an inflection:

(5)



Notice that what is accessed from the lemma is not just a pair of morphemes, but an entire morphological structure to which the morphemes are attached. Levelt (1989, p. 321) called this 'morphological spell-out'. A decade ago the main evidence for the reality of such morphological structures in speech generation came from speech errors. For instance, in the exchange error *I hate* rain*ing on a* hitch*y day* (Shattuck-Hufnagel 1979), the stem *rain* and the root *hitch* got exchanged, leaving the affixes in place. But recently, Roelofs (1996*a*,*b*) began to study the generation of morphological structure by means of reaction-time experiments. In particular, he demonstrated that a word's morphemes are phonologically encoded in incremental fashion (e.g. first *pop*, then *art*). In addition, Janssen, Roelofs and Levelt (submitted) have shown that the spelled-out morphological structure functions as a frame to which the successive morphemes get attached. This holds at least for inflectional morphology, which specifies frames with slots for number and tense affixes.

4.5.1.2 Spelling out the phonological code

After having successfully traversed the Rubicon, the speaker can now begin to spell out the phonological codes of the monomorphemic or multimorphemic words that are involved in generating the phonological word. As mentioned above, the code consists of two parts, a segmental and a metrical code. In phonological speech errors, segmental errors (such as in *if you can change the* p*irst part* in which p is anticipated) are by far

the most frequent. In the following, I will treat segments as basic units of spell-out. It should be added, though, that speech errors can involve consonant clusters, in particular when they are phonologically coherent (Berg 1989), such as in *steady state* stowel. Therefore, Dell (1986) proposed that occasionally whole consonant clusters are spelled out. How 'phonologically complete' are spelled-out segments? Stemberger (1983, 1991*a*,*b*) and others have provided evidence that spelled-out segments can be phonologically underspecified or rather *abstract*. For instance, in one (of many) error induction experiments, Stemberger (1991*a*) showed that on target word pairs such as *sole foe* subjects more frequently erred in the direction of producing *fole* than in producing *soe*. Alveolar /s/ is phonologically unspecified for place. But in spelling out /f/ the marked place feature [labial] comes available. The 'unspecification' of /s/ cannot be inherited by /f/, but the [labial] specification of /f/ can be inherited by /s/, creating the error /f/. Whatever the precise characteristics of underspecification or 'abstractness' of spelled-out segmental units, they come with their contrastive features (the codes that the child develops during phonologization). This accounts for the robust finding that target segments and errors tend to share most of their distinctive features—such reflecting the underlying storage code.

The spell-out of segments can be primed. Schriefers *et al.* (1990) showed this by picture/word interference experiments. Here is an example. The subject has to name a picture of a sheep. At some moment during the trial, beginning with picture onset or a bit earlier or later, the subject hears a prime word. The prime can be phonologically related to the target (*sheet*) or unrelated (*nut*). A major finding was that naming latencies were shorter when the prime was related than when it was unrelated. The explanation is that the related prime (namely *sheet*) activates the corresponding segments in the target word's phonological code (/ʃ/ and /i:/), accelerating their spell-out. Meyer and Schriefers (1991) showed that not only begin-related primes (such as *sheet* for target 'sheep'), but also end-related primes (*deep* for 'sheep') facilitated the naming response. The same held for bisyllabic target words where either the first or the second syllable was shared with the prime (for instance *tailor* and *noble* for target 'table'). The gain in speed of spell-out is 'cashed in' later, during phonetic encoding, as will be discussed below.

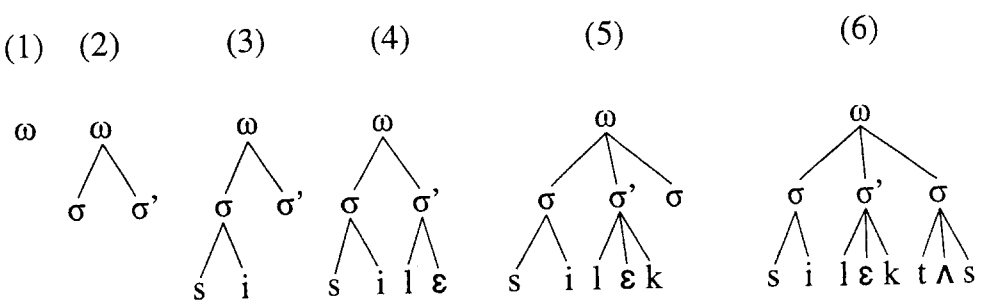
Turning now to the spelling out of the metrical code, it should first be noticed that the code proposed above is rather lean. In the tradition of speech-error based research, the metrical code or 'frame' was supposed to be syllabified, with dedicated slots for onset, nucleus, and coda of each syllable in the word (see Levelt 1989 for a review of this position). The major argument for this view was the syllable position effect; in speech errors syllable onsets tend to exchange with syllable onsets, nuclei with nuclei, and codas with codas. If spelled-out segments would be marked for their slots in the frame (i.e. onset, nucleus, coda), they would automatically end up in the right syllable position, even in the case of error. But there are good reasons for not jumping to this conclusion. First, the syllable position constraint may be an epiphenomenon. Most segment errors (about 80 per cent in English) are word onset errors and word onsets are syllable onsets. Of the remaining 20 per cent a large part can be explained by the simple circumstance that when a consonant moves into the nucleus position the syllable will

usually be unpronounceable (hence, that error will not occur). Finally, the above mentioned feature similarity between error and target will increase the probability that an error ends up in the same syllable position as the target. Vowels are more similar to vowels than to consonants, and syllable-final consonants are more similar to syllable-final consonants than to syllable-onset consonants.

The second reason is that a marking of spelled-out segments for their target position in a syllable will quite regularly interfere with phonological word encoding. How should the /t/ in ⟨select⟩ be marked? Marking as syllable-final would be alright for the encoding of *Whom do we select?*, with the syllabification *se-lect*. But it goes wrong for *They will select us*, with the syllabification *se-lec-tus*. Syllable positions are too variable and context dependent to be fixed codes in memory. Béland *et al.* (1990) suggested, as one of a few possible alternatives, that there is no frame whatsoever. And indeed, one should seriously consider whether a phonological word's metrical structure wouldn't automatically emerge from concatenating successive segments into (weak or strong) syllables. This might in particular work for a language such as French, which has word-final stress across the bank. There are, however, empirical arguments (Roelofs and Meyer 1997; Levelt *et al.* 1999) to assume that for stress-assigning languages such as Dutch and English, the spelled-out metrical frame does play a role.

4.5.1.3 Prosodification

The final step in phonological word construction is the incremental generation of its syllabic and metrical structure. How does this work for the phonological word *select us* in the utterance *They will select us*? Spell-out of the two lexical elements *select* and *us* left us with the following ingredients: two ordered sets of segments: /s/, /i/, /l/, /ɛ/, /k/, /t/ and /ʌ/, /s/, and one non-default metrical pattern for *select*: σσ'. That *select* and *us* should form one phonological word, that is *us* should cliticize to the head word *select* is syntactically conditioned in the surface structure. The procedure consists in incrementally attaching the ordered string of spelled-out segments to syllable nodes, either nodes in the non-default spelled-out metrical frame, or new nodes to be created on-line. This is, in short, the course of action for the present example (for more detail, see Levelt and Wheeldon 1994):



First, (1) the phonological word root ω is set up. Then (2) the spelled out metrics of *select* is attached to this root. In case there is no spelled-out metrical pattern, that is in

the default case, a first syllable node is attached to the root. More specifically, a condition for opening a new syllable node is that there is a (further) vowel coming up: the system can look ahead up till the next vowel. Next (3) /s/ is attached to the leftmost syllable node, followed by /i/. Then (4) /l/ is to be attached, but attachment of a consonant is syllable-initial by default (this is called the 'maximization of onset' rule). Hence, it should attach to the next syllable. In case there would be no next syllable node, that is in case of default metrics, a new syllable node is opened. This is allowed because there is a further vowel in the offing. The new vowel element /ɛ/ will as a nucleus attach to the same syllable. Then (5) /k/ is up for attachment. Default attachment of a consonant is to syllable onset. A new syllable node is created in view of the upcoming vowel /ʌ/ down the line. However, /k/ cannot be attached to the syllable onset position, because /kt/ is not a legal syllable onset in English (it violates the so-called sonority gradient rule). Hence /k/ attaches as offset to the current syllable. Next (6) /t/ will attach as onset to the next syllable, followed by vowel /ʌ/. No new syllable node can be set up to attach the final consonant /s/ to, because there is no further vowel in the offing and /s/ attaches as offset to the current syllable.

The example shows how successive phonological syllables are created on the fly as successive segments attach to syllable nodes. Syllable nodes have either been spelled out or they are newly created every time a further vowel is coming up. Spelled-out segments are not *a priori* marked for syllable positions. For instance, the example shows that though consonants have a predilection for syllable-onset positions, they may well end up in syllable-final position depending on the prevailing context.

What is the evidence for this incremental prosodification process? There are, in particular, two claims in the theory. The first one is that the process is incremental, segment by segment, syllable by syllable. The second is that it makes use of spelled-out metrical information in case the metrics is not default. The evidence for these two claims stems from a host of experiments by Meyer and Roelofs. It is, however, beyond the scope of the present chapter to review that work in detail. The reader is referred to the comprehensive review in Levelt *et al.* (1999).

This completes our consideration of morpho-phonological encoding. The output of this complex process is a phonological, syllabified word. Usually, the phonological word is part of a larger utterance, as in the worked-out example above; the phonological word *select us* appears in the larger utterance *They will select us*. We will now turn to this larger context.

4.6 Generating utterance prosody

Phonological words are parts of phonological phrases, and phonological phrases are parts of still larger units, intonational phrases. In the following we will consider the production of phonological and intonational phrases, respectively. These are complex issues, deserving extensive treatment, but at the same time relatively little is known about the underlying generating process. For both types of phrase I will address just two points: what kind of unit is the phrase, and can it be incrementally generated?

4.6.1 Generating phonological phrases

Any sentence-like utterance is a concatenation of metrical units that are called phonological phrases. The first such phrase starts at the beginning of the utterance and ends right after the first lexical head of a noun phrase (NP), a verb phrase (VP), or an adverbial phrase (AP). The next phonological phrase begins just there and ends after the next such lexical head, and so recursively; any remaining tail after the last lexical head is added to the last phonological phrase. Here is an example (from Nabokov's *Bend Sinister*):

> *Claudina*[1]*/ was standing*[2]*/ quite still*[3]*/ in the middle*[4]*/ of the dining room*[5]*/ where he had left her*[6]*/.*

In the surface structure of this sentence we have the following heads of NP, VP, or AP: *Claudina* (head of NP), *standing* (head of VP), *still* (head of AP), *middle* (head of NP), *dining room* (head of NP), *he* (head of NP), *left* (head of VP), and *her* (head of NP). However, *he* and *her* are anaphors; they are not 'full' lexical heads. Each full lexical head ends a phonological phrase, except for the last one, *left*. Here the remaining tail (*her*) is added to the last phrase. Phonological phrases are metrical units in utterance production, phonological output packages, as Bock (1982), Garrett (1982), and Van Wijk (1989) have suggested.

A characteristic property of this metrical unit is its so-called 'nuclear stress'; the head word in a phonological phrase receives more stress than any of the others. That can be quite informative for the listener, because these heads-of-phrase are the syntactic 'pegs' for the sentence's interpretation. But a few qualifications are necessary. First, nuclear stress can be overridden by focal stress. For instance, in the third phrase above, *quite still*, the speaker can focus *quite*, which will then receive more stress than *still*. Second, phonological phrases are rather 'soft' packages. Selkirk (1984) and others have argued that boundaries between phonological phrases vary in depth and that speakers often blend adjacent phonological phrases with shallow borders into larger ones. In the example it would be quite normal for a speaker to pronounce *in the middle of the dining room* as a single phonological phrase. In other words, phonological phrase boundaries are break *options* rather than breaks.

Time and again we have discussed that we speak incrementally; a processing component will be triggered into action by any *fragment* of its characteristic input. The characteristic input for the generation of utterance prosody is the growing surface structure. How far should the speaker minimally look ahead in the surface structure to generate a phonological phrase? Not very far. When a new phrase begins, the speaker can process lemma after lemma without any look ahead. As soon as a head-of-phrase lemma appears, nuclear stress should be assigned and normally the phrase should be completed. The one complication is the tail. Coming to the end of a sentence, the speaker should not open a new phonological phrase after the last lexical head word. That means that there must be so much surface structure in the window that the appearance of a new lexical head word can be excluded. Levelt (1989) argues that that is a very short stretch. This being said, it does not mean that speakers *cannot* or *will not*

look ahead further than a lemma or two. In fact, they probably often do, as will be discussed in the next section.

4.6.2 Generating intonational phrases

An intonational phrase is characterized by its pitch movement, and pitch movement is produced by the vocal tract. As we have seen, the vocal tract is the 'old' system in primate sound communication. Our closest relatives in evolution can phonate but hardly articulate. Their phonation, moreover, is emotional in character; it is under the control of the limbic system. Although our phonation has further evolved as a voluntary system under the control of the cortical face area and the supplementary motor area, our dorsal midbrain area (around the anterior sulcus cinguli) that, just as in other mammals, mediates vocal fold movements, has not lost its old connection to the limbic system. Although emotion can be expressed at all levels, from the semantic to the prosodic, the most immediate expression of emotion in speech is through pitch movement. The voluntary control of pitch movement makes it possible for us to feign emotion in our intonation, but the reverse is much more common; what we can hide in our wording is easily given away in our intonation.

The intonational phrase (IP) is a sense unit. It consists of one or more phonological phrases and often spans the whole utterance. It is characterized by its pitch contour. The intonational meaning of an IP is largely carried by what is called its *nuclear tone*. The nucleus of an IP is the syllable that receives the most prominent *pitch accent*. This pitch movement is the beginning of the nuclear tone. The tone ends at the last syllable of the IP with a *boundary tone*. These two pitch movements can be several syllables apart, or they can be made on the same (final) syllable of the IP. Figure 4.6a and b show these two cases for the sentences *They've a bear* and *They've a polar bear, I believe*, for instance uttered in response to the question *Isn't that a very small zoo?* This tone is called the 'fall–rise'. It begins slightly up at the nuclear syllable, drops over that syllable, only to rise up again at the final, boundary syllable. This tone expresses some reservation. There is a contrast to an apparent opinion that the speaker disagrees with. Compare this to the tone in Fig. 4.6c, which is called 'high–fall'. It could be a response to the question *What's their largest animal?* The tone starts high up from the preceding tune and drops all the way to the base level, just to stay there without further boundary movement. This is a very common tone. It expresses seriousness in a matter-of-fact way. It is typically used for declarative statements. The two parts of the tone fulfil different functions. The nuclear pitch movement is a focusing device, drawing the listener's attention to the one focused word in the IP. In addition it expresses illocutionary force, such as matter-of-factness, reassurement, opposition, etc. The boundary tone has a different kind of illocutionary function. It either rounds up, or it signals non-finality. Non-finality is probably universally expressed in a rising boundary tone. This is clearest in rising question intonation. But it also works in the Fig. 4.6a and b examples. The listener is invited to correct his opinion, and indeed a connecting move like *Oh, I didn't know that* would be appropriate. In these cases, rising intonation

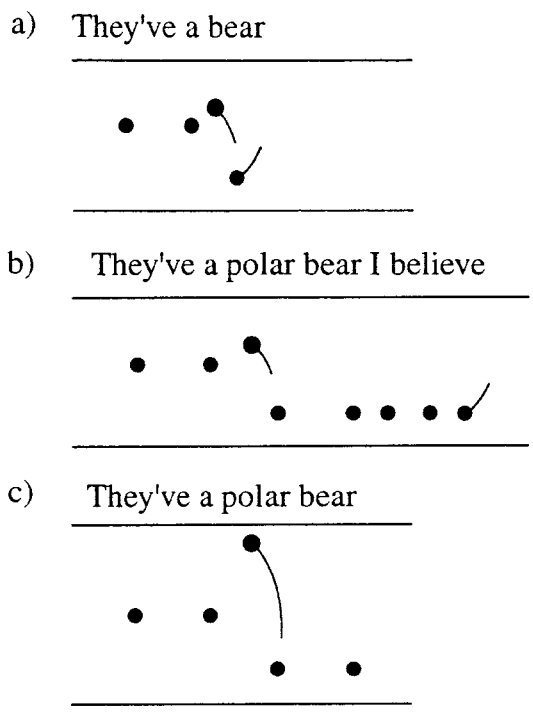


Fig. 4.6 Three nuclear tones: (a) fall–rise with pitch accent and boundary tone on the same syllable; (b) fall–rise with pitch accent and boundary tone on different syllables; (c) high–fall. (Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* MIT Press, Cambridge, MA, reproduced by permission.)

invites the interlocutor to make some move. But a speaker can also use a rising, non-final boundary tone to signal that there is more to come, that the utterance is not yet finished. This is particularly clear in so-called 'listing intonation', for instance when the speaker instructs: *I want you to buy a box of beer* (rise), *a bag of ice* (rise), *and a bottle of Chardonay* (fall). The first two IPs here signal that there is more to come, only the third and last one rounds up the instruction. A steady or falling boundary tone (see also Fig. 4.6c) signals completeness; the case is closed. Each language provides a small number of basic tones, each with its own illocutionary meaning. We discussed two English tones here ('fall–rise' and 'high–fall'), but there are more of them (see Levelt 1989; Cruttenden 1986).

The generation of the nuclear tone doesn't require much look-ahead on the part of the listener. Nuclear pitch movement is made on the stressed syllable of the focused element. That lemma is marked as such in the surface structure. That information can be used as soon as it comes up for phonological encoding. The boundary tone is always at the IP's final syllable, which needs a one-syllable look-ahead to be spotted by the speaker. However, intonation can become much more euphonious when the speaker is early aware of upcoming foci. In euphonious speech, successive focusing pitch movements are often connected by a single pitch contour, for instance in an utterance like *I hópe to be présent at your birthday*. Here the pitch accents on *hope* and *present* can be made as two rise–falls, but also as a rise on *hope* and a subsequent fall on *present*.

The latter so-called 'hat-pattern' sounds a lot better. Empirical evidence for such larger stretches of intonational planning can be found in Blaauw (1995).

4.7 Phonetic encoding and articulation

The output of phonological encoding is a *phonological score* (see Fig. 4.1). The phonological score is an incremental pattern of phonological syllables, metrically grouped and marked for the tones they are participating in. This score must be phonetically realized. Remember that the purpose of the phonological/phonetic system is to prepare a sequence of articulatory gestures. These are patterns of syllabic gestures with their roots in the syllabary that began to develop by the end of the first year of life. These gestural scores have to be addressed by phonological codes and they have various free parameters to be set, both local and global ones, such as duration, amplitude, pitch movement, key, and register (see below).

How are gestural syllable scores addressed or constructed? Languages differ substantially in the number of syllables they use. Chinese and Japanese have no more than a few hundred syllables. Speakers of languages such as these have intensive experience with the articulation of all of their language's syllables. Hence, it is not far-fetched to assume that all of these syllables are stored as gestural scores, maybe in the supplementary motor area, which is the repository of frequently used motor routines (Rizzolatti and Gentilucci 1988). But what about languages such as English or Dutch that use more than 12 000 different syllables? Would the native speaker have them all

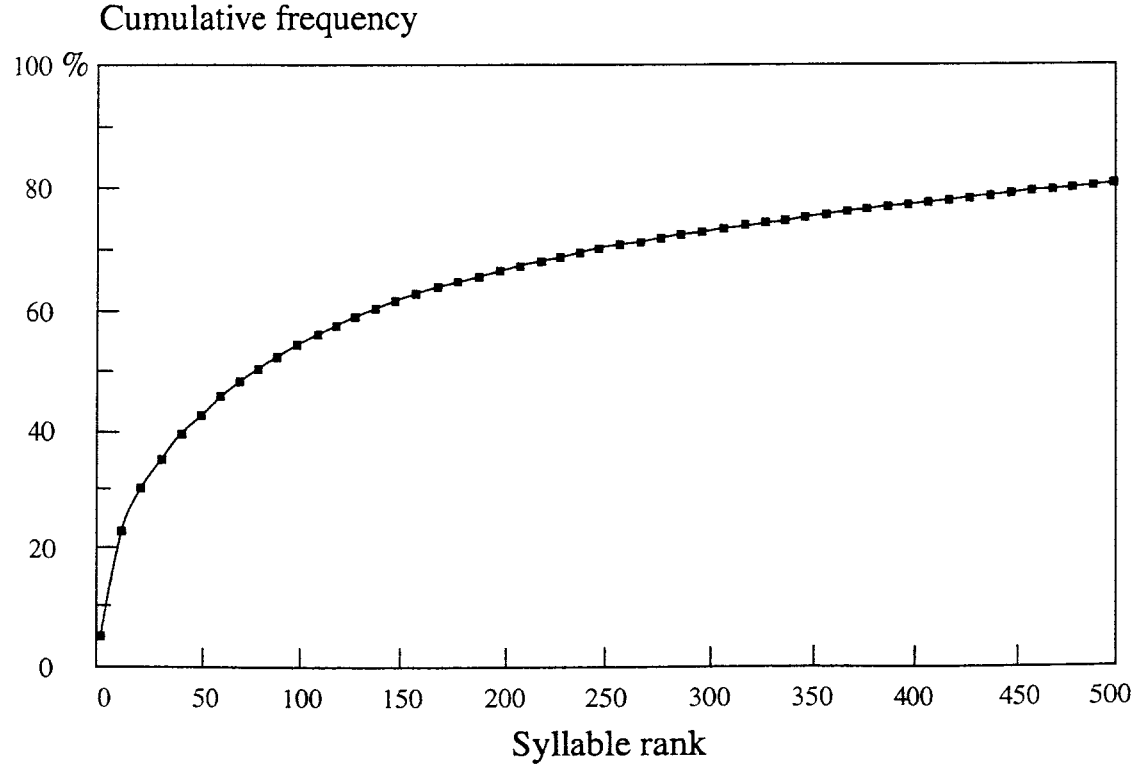


Fig. 4.7 Cumulative frequency distribution of the 500 most frequent syllables of English.

stored as motor routines? We don't know, but there are some relevant statistics that can qualify the question. Figure 4.7 shows the cumulative statistics of syllables used in running English texts. The figure shows that an English speaker produces 50 per cent of his speech with no more than 80 very high-frequent syllables (such as /ə/, /rɪ/, /ðə/, /tu:/, /lI/) and no less than 80 per cent with just 500 different syllables. The question need not be whether all odd 12 000 syllables are stored as gestures, but whether the small number of frequently used syllables are. That would do most of the work. The other syllables may or may not be stored. In the latter case one would need a mechanism that can compute new syllabic gestures.

Addressing a gesture in the syllabary begins during segmental spell-out (see section 4.5.1). Each spelled-out segment, for instance the first /l/ of *select* (see Fig. 4.4), activates all syllables in the syllabary that contain the segment. For instance, /l/ not only activates the phonetic gestural score [lɛk], but also [lɛkt], [lɪst], [bɔ:l], [əʊld], etc. Similarly, /ɛ/ activates not only [lɛk], but also such syllables as [lɛkt], [tɛn], etc. Because the three spelled-out segments /l/, /ɛ/, and /k/ will all activate [lɛk], it will accumulate a relatively high degree of activation. Still, its selection must wait till the phonological syllable /lɛk/ has been created in the incremental syllabification process. Only then the system can establish whether the highly activated syllable score [lɛk] is indeed the correct target. This will be so when the phrase *select us* is constructed, as we have seen, but not when *Whom do we select?* is being generated (see Roelofs 1997 for more details). It was noted in section 4.5.1 that the production of a word can be facilitated by presenting the speaker with an auditory prime, any segment or syllable of the target word. This speeds up the spell-out of the corresponding segments. But that, in turn, speeds up the activation of the target gestural scores in the syllabary. It is, eventually, at this level of phonetic encoding that the priming of segmental spell-out is 'cashed in'. The word's phonetic syllables come faster available, speeding up the spoken response.

A syllable's stored gestural score is still a rather abstract entity. It specifies which articulatory goals have to be successively achieved (Browman and Goldstein 1992), such as a tongue tip closure of the oral cavity at the onset of [lɛk]. There are also parameters to be set, such as for amplitude and pitch movement. They, in turn, depend on the metrical and intonational properties of the larger phrase. In *They'll select us*, for instance, [lɛk] will carry the main accent in the phonological word. Hence it will be stressed, which will be realized in the settings for the amplitude and duration of the vocal part of the gesture. Also, [lɛk] will be the nucleus of the intonational phrase. Depending on the nuclear tone, parameters for its pitch movement will be set (for instance high start, full fall).

Apart from such local settings of gestural parameters, there are also global settings, in particular for *key* and *register*. Key is the range of movement in a phonological phrase. The same pitch movement can be made with a relatively small pitch excursion or with sweeping, full-octave range of movement. This choice of key not only depends on whether the speaker wants to foreground or background the phrase's information for the interlocutor, it is also under emotional control, larger keys expressing more ego-involvement in what is said. Register is the pitch level of the baseline of intonation,

the 'fall-back pitch' of intonational phrases. Whether desired or not, a high register universally expresses vulnerability, helplessness, or special deference. The origin of that impression may be the child's very high speech register. The articulatory score for an utterance is complete when all of these free parameters have been set.

In this chapter I will not go into the intricacies of articulation. A major theoretical and empirical issue is how the abstract gestural tasks are executed by the laryngeal and supralaryngeal systems. The same articulatory task can be performed in many different ways. Producing [l] in [lɛk] can be entirely realized by tongue tip movement. But the oral closure can, in part, be brought about by lifting the jaw, thereby pushing the tongue upward. Similarly lip closure in pronouncing [pɪt] can be realized by moving the upper lip, the lower lip, the jaw, or all three to some extent. In other words, there are many more degrees of freedom than there are articulatory tasks in the execution of a syllabic gesture. Theories differ in how they handle this reduction problem. Usually there is some kind of economy principle involved; how can the task be performed with a minimum of effort? For reviews of these matters, see Levelt (1989) and especially Kent *et al.* (1996).

4.8 Self-monitoring

There is no more complex cognitive–motor activity than speaking. The semantic/syntactic system has to map states of affairs in various modalities onto syntactically organized strings of lemmas. These come at a speed of two to three per second in normal conversation. The phonological/phonetic system must map this abstract surface structure onto the high-speed articulatory movements (10–15 consonants and vowels per second) that generate overt speech. Much can go wrong here, as appears from various kinds of speech errors that we make. But most surprising is how little goes wrong. Although error statistics differ, most of us do not make many more errors than about one per thousand words. The effort of keeping control is more apparent from hesitations, dysfluencies, and fresh starts that abound in normal speech. We are continuously monitoring what we produce or what we are about to produce. How is this monitoring system organized?

Let us return once more to the two systems that underlie our speech production, the semantic/syntactic system and the phonological/phonetic system. Both systems are subject to self-monitoring, but probably in different ways. Here is an example of self-monitoring within the former system, resulting in self-repair (from Schegloff 1979):

(6) Tell me, uh what—d'you need a hot sauce?

The speaker probably started out saying *what do you need?*, but then decided to rather issue a yes/no question. This led to interruption of the original utterance and a fresh start. As Levelt (1989, p. 460) puts it: 'The speaker can directly monitor the messages he prepares for expression, and he may reject a message before or after its formulation has started'. There has never been serious doubt that the conceptual system is as much involved in the production as the perception of speech; it is a shared system. Nobody ever proposed that the message is first conceptually produced and then (more or less incrementally) conceptually parsed or perceived as a way of self-monitoring. But the

story is less obvious for the control of syntactic operations, as in the following case (from Levelt and Cutler 1983):

(7) What things are this kid—is this kid going to say incorrectly?

Opinions differ about the question whether our systems for generating syntax and for syntactic parsing are shared or different. Levelt *et al.* (1999) opted for sharing: 'the perceptual and production networks coincide from the lemma level upwards'. That paper dealt with the generation of words, but consistency may require to extend the claim to all processing within the concept/lemma domain. Kempen (1997) provided further arguments in support of the shared system claim. Still, the issue is at present unsettled.

Almost certainly not shared is the phonological/phonetic system. It simply cannot be the case that the neuromotor system that generates spoken-word gestures is identical to the neuroacoustic system that parses the auditory speech signal. Certainly, there will exist important connections between these systems, as Liberman has time and again claimed. It may in particular be the case that it is the listener's business to detect the articulatory gestures that produced the acoustic wave (Liberman 1996). But the neural substrate for acoustic/phonetic analysis is primarily the left temporal lobe (Démonet *et al.* 1992), whereas the phonetic generation of speech is largely controlled by the motor and premotor areas of the frontal lobe and the left central gyrus of the insula (Dronkers, 1996). Hence, for the speaker who made the error *unut* in (8):

(8) A unut—a unit from the yellow dot.

and corrected it, the most likely route of self-monitoring was through the speech perception system. The speaker either heard the error in listening to his own overt speech or there was some internal representation, let us call it 'internal speech', that the speaker had been attending to. Levelt (1983) called this feedback mechanism the 'perceptual loop', which has an external branch (via overt speech) and an internal one (via internal speech). McGuire *et al.* (1996) provided support for this perceptual loop hypothesis by showing in a PET study that the monitoring of self-generated speech involves the temporal cortices, engaging areas concerned with the processing of externally presented speech.

Still, the notion of 'internal speech' is notoriously vague. What kind of representation is the inner voice that we can attend to in self-monitoring? The choice is essentially given in the previous sections. There are three alternatives. The internal representation monitored for in the internal loop can be (i) the spelled-out phonological code—see Fig. 4.5, step 2; (ii) the phonological score—see Fig. 4.5, step 3; or (iii) the articulatory score—see section 4.7. Wheeldon and Levelt (1995) developed a self-monitoring task for distinguishing between these three alternatives. In the task the (Dutch) subject was given a target segment (or string), for instance the consonant /l/. The subject would then hear an English word (example: *hitch hiker*) whose translation equivalent in Dutch was known to the subject (for the example, the translation equivalent is *lifter*). The task was not to overtly produce the Dutch word in response to

the English word stimulus, but merely to check whether the Dutch translation equivalent contains the target segment and, if so, to push the response button. In this task, the subject supposedly checks the internally generated Dutch response word for the presence of the target segment. Figure 4.8, left upper panel, presents average response latencies when monitoring for segments that can be in syllable-onset position of the first or second syllable in bisyllabic words. The detection latencies for targets in these two positions differ by about 110 ms.

Is the speaker monitoring whatever there is in the articulatory buffer, that is the articulatory score (as proposed by Levelt 1989)? This can be tested by filling the buffer with different materials during the execution of the experimental task. Following Baddeley *et al.* (1984), we had the subject count aloud during execution of their detection task. The right upper panel of Fig. 4.8 gives the results for this condition. Response latencies were, of course, somewhat longer and the difference between monitoring for the two target segment positions was somewhat reduced, but a substantial, highly significant effect remained. Hence, the internal speech monitored must involve a representation different from the articulatory score. How to distinguish between the two remaining alternatives? Remember that the spelled-out phonological code is not yet syllabified; the phonological word, however, is syllabified. Hence, we tested whether internal speech monitoring is syllable-sensitive. Subjects were now given consonant/vowel (CV) or consonant/vowel/consonant (CVC) targets to monitor. For instance, the target could be /ma:/ for one block of trials and /ma:x/ for another block. For the Dutch internal response word *ma-gen* ('stomachs') the former target coincides with the first syllable, whereas for the word *maag-den* ('virgins') the latter target coincides with the first syllable. Still both targets occur in both words. The subjects' reaction times are presented in the lower panel of Fig. 4.8. It shows a full cross-over effect. Subjects are much faster when the target coincides with the first syllable than when it doesn't. In other words, the subjects monitored a syllabified representation. This excludes alternative (i) above, the spelled-out phonological code, and the conclusion is that internal self-monitoring runs on the phonological score, the string of syllabified phonological words.

Given the evidence for an internal perceptual loop in phonetic/phonological self-monitoring, it is hard to imagine that the system would use this loop exclusively for detecting word form trouble. It would always have to stop the perceptual parse at some phonological level of processing. But speech perception is reflex-like and it will rush forth into the syntactic/semantic domain. So even if there exists a self-contained monitoring facility within the semantic/syntactic system of the type Kempen (1999) proposes, then the feedback loop presented in the blueprint of Fig. 4.1 will keep contributing to semantic/syntactic self-monitoring as well.

4.9 Conclusion: relevance for brain-imaging research

Nothing is more useful for functional brain-imaging than an explicit processing theory. The subtraction method, and variants thereof, require the theoretical isolation

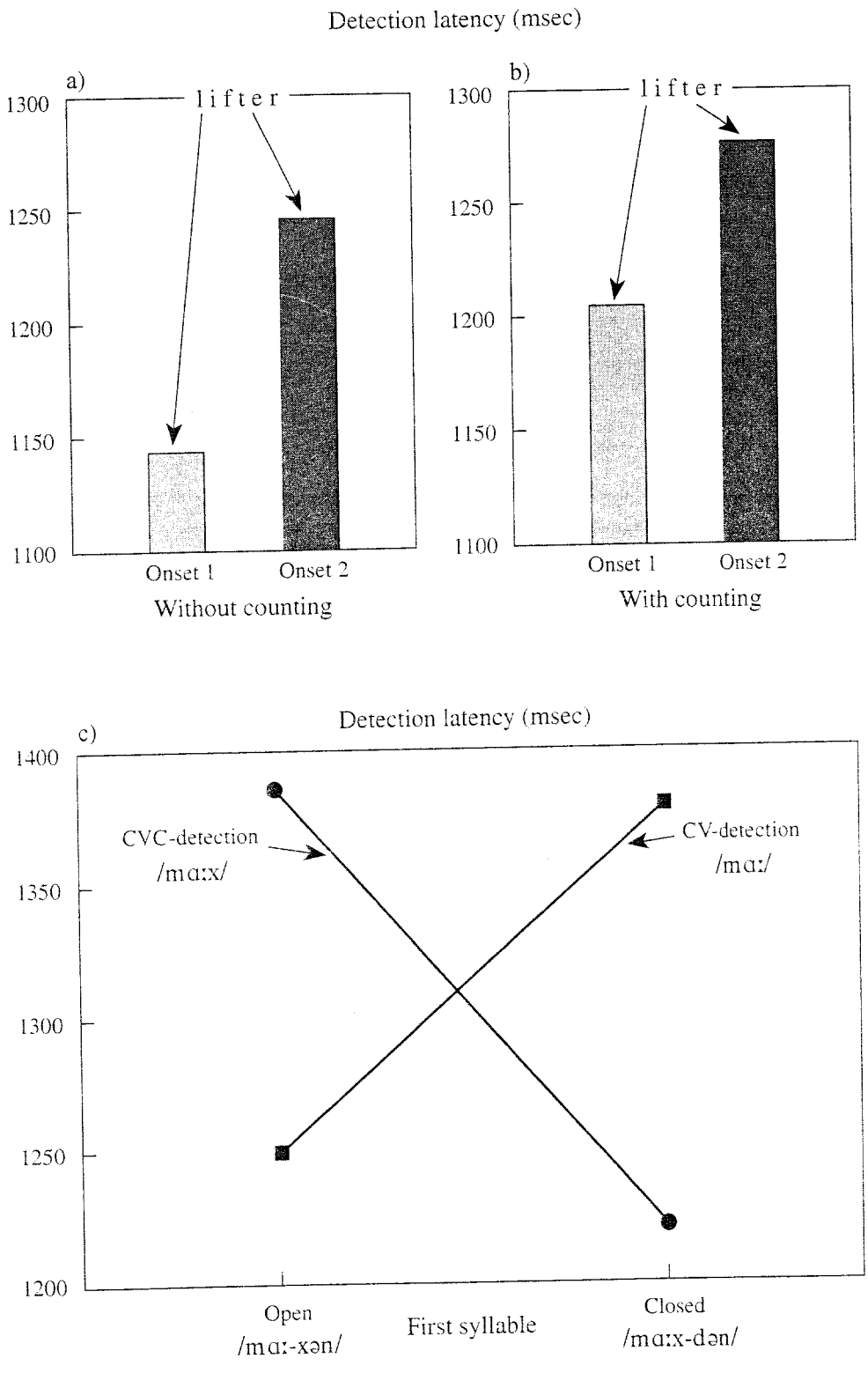


Fig. 4.8 Latencies for monitoring targets in internal speech: (a) syllable-initial consonant targets in bisyllabic words; (b) same, but with concurrent counting aloud; (c) CV and CVC targets that either do or do not correspond to the word-initial syllable.

of particular processing components that are involved in the execution of a task. But intuition doesn't suffice here. The componential processing analysis of any complex task is a research programme in itself. The present chapter exemplifies this for speaking tasks. A quarter century ago only a small minority of the components discussed in the previous sections had been recognized at all, let alone been analysed as a modular process. Now, there is little disagreement about the major building blocks of the functional architecture of speaking and there is a wealth of detail about the component processes involved in the generation of words. Word production tasks used in brain-imaging research vary greatly, ranging from overt picture naming to silent word reading. All of these tasks incorporate some subset of the core processing components in word production that have been discussed in the present chapter, but they differ from task to task. This allowed Indefrey and Levelt (in press) to perform a meta-analysis of the word production findings in the imaging literature, which converged on a surprisingly consistent picture of the cerebral network subserving the generation of words. In addition, we now have reasonable timing estimates of successive functional steps in the process of word generation. These timing windows can with profit be used in the analysis of MEG or EEG activation patterns during the execution of word production tasks, as demonstrated by Levelt *et al.* (1998).

Notes

1. The present blueprint differs from the one in Levelt (1989, p. 9) rather in the partitioning of components than in their character and order of processing. In the original blueprint, grammatical and phonological encoding were grouped together as components of the formulator. That partitioning still makes sense from a linguistic perspective; they are the two components involved with purely linguistic representations. In the present version I rather stressed the evolutionary, developmental, and processing distinction between the symbolic and the form processors. The roots of phonetic encoding, such as the repository of frequently used syllabic gestures (now called the syllabary), were all there in the 1989 theory, but it does make sense to partition the many newly discovered phenomena of what was originally called phonological encoding under two separate rubrics: phonological and phonetic encoding, which involve rather different kinds of processes.
2. The formalism used here is a much slimmed down version of Kempen's (1999) proposal.
3. Again, I am roughly following Kempen's (1999) 'Performance Grammar', but there is no room here to do justice to the extent and detail of his treatment.

References

Baddeley, A., Lewis, V., and Vallar, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology*, **36A**, 233–52.

Badecker, W., Miozzo, M., and Zanuttini, R. (1995). The two-stage model of lexical retrieval: Evidence from a case of anomia with selective preservation of grammatical gender. *Cognition*, **57**, 193–216.

Béland, R., Caplan, D., and Nespoulous, J.-L. (1990). The role of abstract phonological representations in word production: Evidence from phonemic paraphasias. *Journal of Neurolinguistics*, **5**, 125–64.

Berg, T. (1989). Intersegmental cohesiveness. *Folia Linguistica*, **23**, 245–80.

Blaauw, E. (1995). *On the perceptual classification of spontaneous and read speech*. OTS Dissertation Series.

Bock, J. K. (1982). Towards a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, **89**, 1–47.

Bock, K. and Levelt, W. (1994). Language production. Grammatical encoding. In *Handbook of psycholinguistics* (ed. M. A. Gernsbacher), pp. 945–84. Academic Press, New York.

Bogdan, R. (1997). *Interpreting minds: The evolution of a practice*. MIT Press, Cambridge, MA.

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, **49**, 155–80.

Brysbaert, M. (1996). Word frequency affects naming latency in Dutch when age of acquisition is controlled. *The European Journal of Cognitive Psychology*, **8**, 185–94.

Caramazza, A. C. and Miozzo, M. (1997). The relation between syntactic and phonological knowledge in lexical access: Evidence from the 'tip-of-the-tongue' phenomenon. *Cognition*, **64**, 309–43.

Carroll, J. B. and White, M. N. (1973). Word frequency and age-of-acquisition as determiners of picture naming latency. *Quarterly Journal of Experimental Psychology*, **25**, 85–95.

Clark, E. V. (1997). Conceptual perspective and lexical choice in language acquisition. *Cognition*, **64**, 309–43.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Cruttenden, A. (1986). *Intonation*. Cambridge University Press.

De Boysson-Bardies, B. and Vihman, M. M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language*, **67**, 297–318.

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

Démonet, J.-F., Chollet, F., Ramsay, C., Cardebat, D., Nespoulos, J. L., and Wise, R. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain*, **115**, 1753–68.

Dronkers, N. F. (1996). A new brain region for coordinating speech articulation. *Nature*, **384**, 159–61.

Dunbar, R. (1996). *Grooming, gossip and the evolution of language*. Faber and Faber London.

Elbers, L. (1982). Operating principles in repetitive babbling: A cognitive continuity approach. *Cognition*, **12**, 45–63.

Everaert, M., Van der Linden, E.-J., Schenk, A. and Schreuder, R. (eds) (1995). *Idioms: Structural and psychological perspectives*. Erlbaum, NJ.

Fry, D. (1969). The linguistic evidence of speech errors. *BRNO Studies of English*, **8**, 69–74.

Garrett, M. F. (1976). Syntactic processes in sentence production. In *Psychology of learning and motivation*, Vol. 9, (ed. G. Bower), pp. 231–56. Academic Press, New York.

Garrett, M. F. (1982). Production of speech: Observations from normal and pathological language use. In *Normality and pathology in linguistic performance: Slips of the tongue, ear, pen, and hand* (ed. A. W. Ellis), pp. 19–76. Academic Press, New York.

Glaser, W. R. and Düngelhoff, F.-J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 640–54.

Hovy, E. H. (1994). Automated discourse generation using discourse structure relations. In *Natural language processing* (eds C. N. Pereira and B. J. Grosz), pp. 341–86. MIT Press, Cambridge, MA.

Indefrey, P. and Levelt, W. J. M. (2000). Language production. In *The cognitive neurosciences*, 2nd edn (ed. M. Gazzaniga). MIT Press, Cambridge, MA.

Jackendoff, R. (1997). *The architecture of the language faculty*. MIT Press, Cambridge, MA.

Janssen, D.P., Roelofs, A. and Levelt, W.J.M. (submitted). Inflectional frames in language production.

Jescheniak, J. D. and Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20**, 824–43.

Kent, R. D., Adams, S. G., and Turner, G. S. (1996). Models of speech production. In *Principles of experimental phonetics* (ed. N. J. Las), pp. 3–45. Mosby, St. Louis.

Kempen, G. (1999). Grammatical performance in human sentence production and comprehension. (Book manuscript.)

Kempen, G. and Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science*, **11**, 201–58.

Leslie, A. and Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, **25**, *267–87*.

Levelt, C. C. (1994). *The acquisition of place*. Holland Institute of Generative Linguistics Publications.

Levelt, W. J. M. (1982). Linearization in describing spatial networks. In *Processes, beliefs, and questions* (eds S. Peters and E. Saarinen), pp. 199–220. Reidel, Dordrecht.

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*. MIT Press, Cambridge, MA.

Levelt, W. J. M. (1996). Perspective taking and ellipsis in spatial descriptions. In *Language and space* (eds P. Bloom, M. A. Peterson, L. Nadel, and M. F. Garrett), pp. 77–108. MIT Press, Cambridge, MA.

Levelt, W. J. M. (1998). The genetic perspective in psycholinguistics. Or where do spoken words come from? *Journal of Psycholinguistic Research*, **27**, 167–80.

Levelt, W. J. M. and Cutler, A. (1983). Prosodic marking in speech repair. *Journal of Semantics*, **2**, 205–17.

Levelt, W. J. M. and Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, **50**, 239–69.

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, Th., and Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, **98**, 122–42.

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Science*, **22**, 1–38.

Levelt, W. J. M., Praamstra, P., Meyer, A. S., Helenius, P., and Salmelin, R. (1998). An MEG study of picture naming. *Journal of Cognitive Neuroscience*, **10**, 553–67.

Liberman, A. (1996). *Speech: A special code*. MIT Press, Cambridge, MA.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, **21**, 499–511.

McCoy, K. F. and Cheng, J. (1991). Focus of attention: Constraining what can be said next. In *Natural language generation in artificial intelligence and computational linguistics* (eds C. L. Paris, W. R. Swartout, and W. C. Mann), pp. 103–24. Kluwer, Dordrecht.

McGuire, P. K., Silbersweig, D. A., and Frith, C. D. (1996). Functional neuroanatomy of verbal self-monitoring. *Brain*, **119**, 101–11.

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**, 524–45.

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.

Meyer, A. S. and Schriefers, H. (1991). Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **17**, 1146–60.

Morrison, C. M., Ellis, A. W., and Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory and Cognition*, **20**, 705–14.

Müller-Preuss, P. and Ploog, D. (1983). Central control of sound production in mammals. In *Bioacoustics—A comparative study* (ed. B. Lewis), pp. 125–46. Academic Press, London.

Oldfield, R. C. and Wingfield, A. (1965). Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology*, **17**, 273–81.

Pereira, F. C. N. and Grosz, B. J. (eds) (1994). *Natural language processing*. MIT Press, Cambridge, MA.

Petitto, L. A. and Marentette, P. F. (1991). Babbling in the manual mode: Evidence for the ontogeny of language. *Science*, **251**, 1493–6.

Ploog, D. (1990). Neuroethological foundations of human speech. In *From neuron to action* (eds L. Deecke, J. C. Eccles, and V. B. Mountcastle), pp. 365–74. Springer, Berlin.

Premack, D. and Premack, A. J. (1995). Origins of human social competence. In *The cognitive neurosciences* (ed. M. S. Gazzaniga), pp. 205–18. MIT Press, Cambridge, MA.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, **1**, 515–26.

Rizzolatti, G. and Gentilucci, M. (1988). Motor and visual-motor functions of the premotor cortex. In *Neurobiology of neocortex* (ed. P. Rakic and W. Singer), pp. 269–84. Wiley, Chichester.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, **42**, 107–42.

Roelofs, A. (1993). Testing a non-decompositional theory of lemma retrieval in speaking: Retrieval of verbs. *Cognition*, **47**, 59–87.

Roelofs, A. (1996*a*). Serial order in planning the production of successive morphemes of a word. *Journal of Memory and Language*, **35**, 854–76.

Roelofs, A. (1996*b*). Morpheme frequency in speech production: Testing WEAVER. In *Yearbook of morphology* (eds G. E. Booij and J. van Marle), pp. 135–54. Kluwer, Dordrecht.

Roelofs, A. (1997*a*). Syllabification in speech production: Evaluation of WEAVER. *Language and Cognitive Processes*, **12**, 657–93.

Roelofs, A. (1997*b*). The WEAVER model of word-form encoding in speech production. *Cognition*, **64**, 249–84.

Roelofs, A. and Meyer, A. S. (1997). Metrical structure in planning the production of spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **24**, 1–18.

Schegloff, E. (1979). The relevance of repair to syntax-for-conversation. In *Syntax and semantics*, Vol. 12 (ed. T. Givón), pp. 261–88. Academic Press, New York.

Schmitt, B. (1997). Lexical access in the production of ellipsis and pronouns. Unpublished Ph.D. thesis, Nijmegen University.

Schriefers, H., Meyer, A. S., and Levelt, W. J. M. (1990). Exploring the time course of lexical access in speech production: Picture-word interference studies. *Journal of Memory and Language*, **29**, 86–102.

Selkirk, E. (1984). *Phonology and syntax*. MIT Press, Cambridge, MA.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial order mechanism in sentence production. In *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. (eds W. E. Cooper and E. C. T. Walker), pp. 295–342. Lawrence Erlbaum, Hillsdale.

Slobin, D. (1987). Thinking for speaking. In *Berkeley Linguistics Society: Proceedings of the Thirteenth Annual Meeting* (eds J. Aske, N. Beery, L. Michaelis, and H. Filip), pp. 435–45. Berkeley Linguistics Society.

Snodgrass, J. G. and Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavioral Research Methods, Instruments, and Computers*, **28**, 516–36.

Stemberger, J. P. (1983). *Speech errors and theoretical phonology: A review*. Indiana University Linguistics Club.

Stemberger, J. P. (1991*a*). Radical underspecification in language production. *Phonology*, **8**, 73–112.

Stemberger, J. P. (1991*b*). Apparent anti-frequency effects in language production: The addition bias and phonological underspecification. *Journal of Memory and Language*, **30**, 161–85.

Sutton, D., Larson, C., and Lindemann, R. C. (1974). Neocortical and limbic lesion effects on primate phonation. *Brain Research*, **71**, 61–75.

Van Turennout, M., Hagoort, P., and Brown, C. M. (1998). Brain activity during speaking: From syntax to phonology in 40 milliseconds. *Science*, **280**, 572–4.

Van Wijk, C. (1987). The PSY behind PHI: A psycholinguistic model for performance structures. *Journal of Psycholinguistic Research*, **16**, 185–99.

Vigliocco, G., Antonini, T., and Garrett, M. F. (1997). Grammatical gender is on the tip of Italian tongues. *Psychological Science*, **8**, 314–7.

Wheeldon, L. R. and Levelt, W. J. M. (1995). Monitoring the time course of phonological encoding. *Journal of Memory and Language*, **34**, 311–34.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, **13**, 103–28.

Wingfield, A. (1968). Effects of frequency on identification and naming of objects. *American Journal of Psychology*, **81**, 226–34.

Zock, M. (1997). Sentence generation by pattern matching: The problem of syntactic choice. In *Recent advances in natural language processing*. (eds R. Mitkov and N. Nicolov), pp. 317–52. Amsterdam, Benjamins.