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 practiced 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-Peissl 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 calls 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 phoation, 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 rhythm, 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, 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, beliefs, 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 (Levent 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 babbling. Babbling is 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 1981; 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 Pettito 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. Levent 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/featral 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 multoword 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:

**Rhetorical/semantical/syntactic system**
- Conceptional preparation
- Grammatical encoding

**Mental lexicon**
- Lemmas
- Morpho-phonological encoding

**Phonological/phonetic system**
- Articulation
- Phonetic encoding

**Model of addresses (ToM)**
- Knowledge of external and internal world
- Discourse model, etc.

**Syllabary**
- Gestural scores

*Fig. 4.1* A blueprint of the speaker.

**Conceptional 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 skillfully 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-di*. 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.

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 QToM: coordination 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 Grosse 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.