Speech Comprehension and Speech Production

As language is defined as a system of symbolic representations of meaning, the term does not restrict itself to a particular means of communication but applies to speech as well as to several other forms such as writing or deaf people’s signing language, but also for example logically based computer languages. Nevertheless, the core of our everyday understanding of language is speech. It is the form in which human language evolved and even today only about 200 of the estimated 6000 to 7000 spoken languages also exist in a written form. This chapter deals with the cognitive efforts people take each time they are engaged in a conversation: Speech production and comprehension are two complementary processes that mediate between the meaning of what is said (believed to be represented independent of language) and the acoustic signal that speakers exchange. Both ways of transformation include many steps of processing the different units of speech, like phonologic features, phonemes, syllables, words, phrases and sentences. Putting it simple, those steps are performed in top-down order in speech production and in bottom-up order in speech comprehension. In spite of scientific consensus about the overall structure of speech comprehension and production, there are many competing models. The chapter will present some of these models along with evidence from experimental psychology. The comprehension section of this chapter starts with an account of how sound waves turn into phonemes, the smallest units of meaning, and goes on to address processing on the word level. Sentence comprehension – which follows word processing in the hierarchy of language – has not yet been addressed in this chapter and the Wiki-community is requested to add a section dealing with this issue. In the section treating speech production, the reader is introduced to the planning involved in transforming a message into a sequence of words, the problem of lexicalization (i.e., finding the right word) and the final steps up to the motor generation of the requested sounds.

## Speech Comprehension

There is no doubt that speech comprehension is a highly developed ability in humans because despite the high complexity of the speech signal, it happens almost automatically and notably fast: We can understand speech at the rate of 20 phonemes per second, while in a sequence of non-speech sounds, the order of sounds can only be distinguished if they are presented slower than 1,5 sounds per second (Clark & Clark, 1977). This is a first hint that there must be mechanisms that utilise the additional information speech contains, compared to other sounds, in order to facilitate processing.

### Phoneme perception

The starting point of our way to comprehend an utterance is the acoustic sequence reaching our auditory system. As a first step, we have to separate the speech signal from other auditory input. This can be done because speech is continuous, which background noise normally is not, and because throughout our live our auditory system learns to utilise acoustic properties like frequency to assign sounds to a possible source (Cutler & Clifton, 1999). Next, we have to identify as segments the individual sounds that form the sequential speech signal, so that we can relate them to meaning. This early part of speech comprehension is also referred to as **decoding**. The smallest unit of meaning is the **phoneme**, which is a distinguishable single sound that in most cases corresponds to a particular letter of the alphabet. However, letters can represent more than one phoneme like “u” does in “hut” and “put”. Linguists define phonemes by the particular way of articulation, that is, the parts of the articulatory tract and the movements involved in producing them. For example, the English phoneme /k/ is defined as a “velar voiceless stop consonant” (Harley, 2008). Phonemes that share certain articulatory and therefore phonological features appear more similar and are more often confused (like /k/ and /g/ that sound more alike than /k/ and /d/ or /e/).

**Decoding**  Translating the acoustic speech signal into a linguistic representation.

**Phoneme** The smallest linguistic unit that has an impact on meaning and at the same time the smallest phonological unit perceived as a distinct sound. Often linked to a particular letter of the alphabet.

***Categorical perception of phonemes***

Although we perceive speech sounds as phonemes, it is wrong to assume that the acoustic and the phonemic level of language are identical; rather, by acquisition of our first language, phonemes shape our perception. Phonemes are categories that comprise sounds with varying acoustic properties and they differ between languages. Maybe the best-known example is that Japanese native speakers initially have difficulties to tell apart the European phonemes /r/ and /l/ which constitute one phoneme in Japanese. Likewise, Hindi speakers distinguish between two sounds which Europeans both perceive as /k/. The learned pattern of distinction is applied strictly, causing categorical perception of phonemes: People usually perceive a sound as one phoneme or another, not as something in between. This effect was demonstrated when artificial syllables that varied acoustically along a continuum between two phonemes were presented to participants. At a certain point there was a “boundary” between the ranges in which either one of the two phonemes was perceived (Liberman, Harris, Hoffman & Griffith, 1957). Nevertheless, it was found that if requested to do so, we are able to detect minimal differences between sounds that belong to the same phoneme category (Pisoni & Tash, 1974). Accordingly it seems that categorical perception is not a necessity of an early level of processing but a habit employed to simplify comprehension. There is still controversy about how categorical perception actually happens. It seems unlikely that perceived sounds are compared with a “perfect exemplar” of the respective phoneme because a large number of “perfect exemplars” of each phoneme would be necessary to apply to speakers of different age, gender, dialect and so on (Harley, 2008).

***Coarticulation***

Anyway, theories that assume phoneme perception can rely on phonologic properties alone are confronted with two basic problems: The invariance and the segmentation problem. The invariance problem is that the same phoneme can sound differently depending on the syllable context in which it occurs. This is due to co-articulation, which means that while one sound is articulated, the vocal tract is already adapting for the position required for the next sound. It has been argued therefore that syllables are more invariant than phonemes. Maybe this can partly account for the finding that participants take longer to respond to a single phoneme than they take to respond to a syllable (Savin & Bever, 1970), which lead the authors to believe syllables were processed first. Whether this is the case or not, it is without doubt that listeners make use of the information each phoneme contains about the surrounding ones, as experimentally produced mismatches in co-articulatory information (that is, phonemes were pronounced in a way that did not fit to the phoneme that came next) lead to longer reaction times for phoneme recognition (Martin & Brunell, 1981). The invariance problem therefore points to syllables as units that intercede between the meaning bearing role of phonemes and their varying acoustic form.

***Segmentation***

The **segmentation** problem refers to the necessary step of dividing the continuous speech signal into its component parts (i.e., phonemes, syllables, and words). Segmentation cannot be done by use of the signal’s physical features alone, because sounds slur together and cannot easily be separated, which is not only the case within words but also across words (Harley, 2008). If we look at the spectrographic display of an acoustic speech signal, we can hardly tell the boundaries between phonemes, syllables, or words (Figure 1). Maybe the best starting point for segmentation we find in the speech signal is its **prosody**, especially the rhythm. It differs between languages, inducing different strategies of segmentation: French has a very regular rhythm with syllables that do not contract or expand much, which allows for syllable-based segmentation. English, in contrast, strongly distinguishes between stressed and unstressed syllables and those can be expanded or contracted to fit them to the rhythm. Therefore, rhythmic segmentation in English is stress based, yielding units also called **phonological words** that consist of one stressed syllable and associated unstressed syllables, and that do not necessarily correspond with **lexical words** (for example, “select us” is one phonological word; Harley, 2008). The importance of stress structure for speech recognition was shown when participants had to detect a certain string within a non-speech signal. Their reaction was faster when the target was within one phonological word than when it crossed the boundaries of phonological words (Cutler & Norris, 1988). In the same task, the use of syllable knowledge could also be demonstrated because strings that corresponded to a normal syllable were detected faster than those that were shorter or longer (Mehler, Dommergues, Frauenfelder & Segui, 1981). It seems to be another aim of segmentation to assign all of the perceived sounds to words. Thus, detecting words embedded into non-words was more difficult (as it happened more slowly) when the remaining sounds did not have a word-like phonological structure (like f*egg*, as opposed to maff*egg*; Norris, McQueen, Cutler & Butterfield, 1997).

**Segmentation** Dividing the continuous speech signal into its underlying linguistic units like phonemes, syllables and words.

**Prosody** Tonal and rhythmic features of speech such as stress pattern and intonation. Prosody may contain meaning in spoken language; it is absent in written language.

**Phonological word** A unit of prosody that contains one stressed syllable. It may consist of one lexical word, but can be longer or shorter.

**Lexical word** A word as a unit of meaning; corresponds to the everyday understanding of “word”.

***Top-down feedback***

The processes just described operate on higher-level units, but there is evidence for (and yet strong controversy about) top-down feedback from those units onto phoneme identification. One line of evidence is the so-called lexical identification shift that occurs in research designs that examine categorical perception of phonemes with sounds that vary on a continuum between two phonemes. If these are presented in a word context, participants shift their judgement in favour of the phoneme that will create a meaningful word (for example towards /k/ and not /g/ within the word \*iss) (Ganong, 1980). *Phoneme restoration* can be observed if a phoneme of a word spoken within a sentence is cut out and replaced with a cough or a tone. Participants typically do not report that there was a phoneme missing – they perceive the phoneme expected in the word even if they are told it has been replaced. If the same word with a cough replacing a phoneme is inserted into different sentences, each giving contextual support for a different word (for example “The \*eel was on the orange.” vs “The \*eel was on the axle”), participants report to have perceived the phoneme needed for the contextually expected word (Warren & Warren, 1970). Note that the words carrying the content information appear after the phoneme to be restored. Therefore it has been questioned whether it is actually phoneme perception or some later stage of processing that is responsible for the restoration effect. It seems that while there is truly perceptual phoneme restoration (as participants cannot distinguish between words with restored and actually heard phonemes), the context effects must be accounted for by a stage of processing after retrieval of word meaning (Samuel, 1981). The *dual-code-theory* of phoneme identification (Foss and Blank, 1980) states that there are two different sources of information we can use in order to identify individual phonemes: The **prelexical code**, which is computed from the acoustic information, and the **postlexical code** that comes from processing of higher-level units like words and creates top-down feedback. Outcomes of different study designs are interpreted as resulting from use of either one of the information sources. The recognition of a certain phoneme in a non-word is as fast as in a word, suggesting use of the prelexical code, while when presented in a sentence, the phoneme is recognized faster if it is part of a word expected from sentence context than if it is part of an unexpected word, which can be seen as a result of postlexical processing. However, evidence for use of the postlexical code is too limited to support the idea that it is a generally applied strategy.

**Prelexical code** The way speech is encoded before the words are identified. It is based merely on phonological information.

**Postlexical code** The way speech is encoded after the words are identified. It contains semantic and syntactic information.

### Word identification

The identification of words can be seen as a turning point in speech comprehension, because it is the word level at which the semantic and syntactic information is represented which we need to decipher the meaning of the utterance. In the terminology just introduced in the last paragraph, this word-level information is the postlexical code. Here, the symbolic character of language comes into play: Contrary to the prelexical code, the postlexical code is not derived from the acoustic features of the speech signal but from the listener’s mental representation of words (including meaning, grammatical properties etc). Most models propose a mental dictionary, the **lexicon**. The point when a phonological string is successfully mapped onto an entry in the lexicon (and the postlexical code becomes available) is called lexical access (Harley, 2008; Cutler & Clifton, 1999). It is controversial how much word identification overlaps with processing on other levels of the speech comprehension hierarchy. Phoneme identification and word recognition could take place at the same time, at least research has shown that phoneme identification does not have to be completed before word recognition can begin (Marslen-Wilson & Warren, 1994). Concerning the role of context for word identification, theories can be located in between two extreme positions: The autonomous position states that context can only interact with the postlexical code, but does not influence word identification as such. Most specifically, there should be no feedback from later levels of processing (that is, phrase or sentence level) to earlier stages. Exactly this structural context, however, is used for word recognition according to the interactive view. Out of many models of word identification, two shall be introduced in some detail here: The cohort model and TRACE.

**Lexicon** A mental dictionary from which the meaning and syntactic properties of each word is recalled once the word is identified.

***The cohort model***

The cohort model (original version: Marslen-Wilson & Welsh, 1978; later version: Marslen-Wilson, 1990) proposes three phases of word recognition: The *access stage* starts once we hear the beginning of a word and draw from our lexicon a number of candidates that could match, the so-called cohort. Therefore the beginning of a word is especially important for understanding. The initial cohort is formed in a bottom-up manner unaffected by context. With more parts of the word being heard, the *selection stage* follows, in which the activation levels of candidates that no longer fit decay until the one best fitting word is selected. Not only phonological evidence but also syntactic and semantic context is used for this selection process, especially in its late phase. These first two phases of word recognition are prelexical. The recognition point of the word can, but often does not, coincide with its uniqueness point, which is the point at which the initial sequence is unique to one word. If context information can be utilised to dismiss candidates, the recognition point might occur before the uniqueness point, while if there is no helping context information and the acoustic signal is unclear, it might occur after the uniqueness point. With the recognition point, the third and postlexical phase, the *integration stage* begins, in which the semantic and syntactic properties of the chosen word are utilized, for example to integrate it into the representation of the sentence. Consistent with the model, experimentally produced mistakes are more often ignored by participants who have to repeat a phrase if they appear in the last part of the word and if there is strong contextual information, while they cause confusion if they appear in the beginning and if the context is ambiguous (Marslen-Wilson & Welsh, 1978).

***The TRACE model***

TRACE (McClelland & Elman, 1986) is a connectionist computer model of speech recognition, which means it consists of many connected processing units. Those units operate on different levels, representing phonological features, phonemes and words. Activation spreads bidirectionally between units, allowing for both bottom-up and top-down processing. Between units of the same processing level, there are inhibitory connections that make those units compete with each other and simulate phenomena such as the categorical perception of phonemes. Also at the word level there is evidence for competition (that is, mutual inhibition) between candidates: Words embedded in non-word strings take longer time to be detected if the non-word part bears similarity with some other existing word than if it does not. This effect was shown to co-occur with and to be independent of the influence of stress-based segmentation discussed earlier (McQueen, Norris & Cutler, 1994). TRACE is good in simulating some features of human speech perception, especially context effects, while in other points it differs from human perception, such as tolerance for mistakes: Words that have been varied in phonemic details (such as “smob” derived from “smog”) are identified as the related words by TRACE while to humans they appear as non-words (Harley, 2008). Other researchers have criticised the supposed high amount of top-down feedback as a version of TRACE that did not contain top-down-feedback simulated speech perception equally well as the original version (Clifton & Cutler, 1999).

## Speech Production

The act of speaking involves steps of processing similar to the act of listening, but those steps are performed in reverse order, from sentence meaning to phonological features. Speaking can also be seen as bringing ideas into a linear form (as a sentence is a one-dimensional sequence of words). According to Levelt (1989), speakers deal with three main issues. The first is conceptualisation, that is determining what to say and selecting relevant information to construct a preverbal message. The next is formulating this preverbal message in a linguistic form, including selection of individual words, syntactic planning and encoding of the words as sounds. The third issue is execution, which means implementing the linguistic representation on the motor articulatory system. Figure 2 gives an overview of Levelt’s model, with its particular features being addressed in the following sections. There is evidence that speech production is an **incremental** process, which means that planning and articulation take place at the same time and early steps of processing “run ahead” of later ones in the verbal sequence we prepare. For example, if a short sentence with two nouns shall be formed to describe a picture, an auditive distractor delays the onset of speaking if it is semantically related to either of the two nouns or if it is phonologically related to the first, but not to the second noun (Meyer, 1996). This supports the notion that before speaking starts all nouns of the sentence are prepared semantically but that only the first one is already encoded phonetically. It also seems that planning takes place in periodically returning phases, as it was found pauses occur after every five to eight words in normal conversations. Periods of fluent speech alternate with more dysfluent periods, both connected with different patterns of gestures and eye contact. This has been interpreted as ‘cognitive cycles’ that structure speaking (Henderson, Goldman-Eisler & Skarbeck (1966). There is much less literature on speech production than there is on comprehension; most research on speech production has focused on collecting speech errors (and asking the speaker what the intended utterance was) in order to find out how we organise speech production. Experimental studies, utilizing for example picture naming tasks, are a rather new field (Harley, 2008). Therefore there is a paragraph about speech errors before the steps of conceptual and syntactic planning, lexicalization and articulation are treated.

**Incremental processing** Sequential steps of processing take place at the same time in a fashion that while material that has already undergone step 1 is going through step 2, new material goes through step 1.

### Speech errors

All kinds of linguistic units (i. e., phonological feature, phoneme, syllable, morpheme, word, phrase, sentence) can be subjects to speech errors that happen in everyday life as well as in laboratory tasks. Those errors involve different mechanisms like blend, substitution, exchange, addition or deletion of linguistic units (Harley, 2008). To give a more plastic idea of the classification of speech errors, here are some (self-created) examples:

| **Table 1: Examples of Speech Errors.** | | |
| --- | --- | --- |
| **Error** | **Example** | **Intended** |
| Phonologic feature substitution | telete | delete |
| Phoneme deletion | probaby | probably |
| Phononeme exchange (here: also spoonerism) | nigh hoon | high noon |
| Phoneme addition (here: also perseveration) | cream crake | cream cake |
| Syllable deletion | unployment | unemployment |
| Affix substitution | hungerty and pover | hunger and poverty |
| Word deletion | The bar is next the library. | The bar is next to the library. |
| Word substitution | The sky is very hot today. | The sky is very blue today. |
| Word blend | The sky is very blot today. | The sky is very blue today. |
| Phrase blend | I saw him tell it was not a good idea. | I saw him climb up the roof. I told him it was not a good idea. |

The finding that speech errors involve particular linguistic units has been interpreted as an argument that those units are not only descriptive categories of linguists but also subject to actual cognitive steps of speech processing (Fromkin, 1971). Research has shown that errors do not happen at random: If people are confronted with materials that make mistakes likely (for example if they are asked to read out quickly texts that include tongue twisters), errors that form lexically correct words happen more often than errors that do not. Errors that form taboo words are less likely than other possible errors. Nevertheless materials that include the possibility to accidentally form a taboo word cause elevated galvanic skin responses, as though speakers monitor those possible errors internally (Motley, Camden & Baars, 1982).

***Garrett´s model of speech production***

A general model of speech production based on speech error analysis was proposed by Garrett (1975, 1992). His basic assumption is that processing is serial, and different stages of processing do not interact with each other. Phrase planning takes place in two steps: At the *functional level*, in which the content and main syntactic roles like subject and object are determined, and at the positional level that includes determining the final word order and phonological specification of the words used. Content words (nouns, verbs and adjectives) are selected at the first level, function words (like determiners and prepositions) only at the second level. Phonological specification of content word stems therefore takes place before phonological specification of function words or grammatical forms (like plural or past forms of verbs). According to the theory, word exchanges occur at the first level and are therefore influenced by semantic relations, but much less by the distance between the words in the completed sentence. Contrary, sound exchanges as products of phonological encoding occur at a later stage in which the word order already is determined, which makes them constrained by distance. Also in accordance with the theory, sounds typically exchange across short distances, whereas words can exchange across the whole phrase. Garret’s theory also predicts that elements should only exchange if they are part of the same processing level. This is supported by the robust finding that content words and function words almost never exchange with each other (Harley, 2008). Other speech errors are more difficult to explain with Garret’s model: Word blends, like “quizzle” from “quiz” and “puzzle” make it seem probable that two words have been drawn from the lexicon at the same time, contrary to Garret’s idea that language production is serial and not a parallel process. Even more problematic, word blends and even blends of whole phrases seem to be facilitated by phonological similarity. That is, intruding contents and intended contents merge more often at points where they share phonemes or syllables than would be expected by chance. This should not be the case if planning at the functional level and phonological processing are indeed separated stages with no interaction between them (Harley, 1984).

### Conceptual planning

It has already been mentioned that speaking involves linearising ideas. That is because even if what we want to say involves concepts related to each other in complex ways (for example like a network), we have to address them one by one. This is a main object of **conceptual preparation**, a step which – according to Levelt (1999) – takes place even before the ideas are transferred into words, thus yielding a preverbal message. *Macroplanning* is the part of conceptual preparation that can be described as management of the topic. The speaker has to ensure the listener can keep track when he leads his attention from one item to the next. When people go through a set of items in a conversation, they normally pick items directly related to the previous one; if this is not possible they go back to a central item that they can relate to the next one or they start with a simple item and advance to more difficult ones. Ideas that we express in sentences normally include relations between **referents**. To get those relations in the linear form required in speech, we have to assign the referents to grammatical roles like subject and object that are in most languages related to certain positions in the sentence. This is called *microplanning*. Often it is possible to express the same relation through various syntactic constructions, resembling different angles of view and we have to choose one before we can begin to speak. For example, if a cat and a dog sit next to each other we could say “The cat sits right to the dog” as well as “The dog sits left to the cat” (Levelt, 1999). It has been proposed that the overall structure of a sentence (like active versus passive or adverbial at the beginning or at the end) is determined somewhat independent of the content, maybe with the help of a “syntactic module”. Evidence comes from syntactic priming which occurs for example when participants have to describe a picture after reading an unrelated sentence. They choose a syntactic structure resembling the previously read sentence more often than expected by chance. Other aspects like choice of words and their grammatical forms do not interact with this priming (Bock, 1986).

**Conceptual preparation** Transferring relations between concepts into a sequence of syntactic relations.

**Referent** The person, object or concept a word refers to.

### Lexicalisation

The concepts selected during conceptual planning have to be turned into words with defined grammatical and phonological features so we can construct the sentence that we finally encode phonologically for articulation. This “word selection” is called **lexicalisation** and Levelt (1999) postulates it is a two-step process: First a semantically and syntactically specified representation of the word, the so-called **lemma** is drawn, which does not contain phonological information, then the lemma is linked to its phonological form, the **lexeme**. The tip-of-the-tongue state could be an everyday life example for successful lemma selection but a disrupted phonological processing: The phonological form of a word can not be found, even though its meaning and even grammatical or phonological details are known to the speaker. The model of separate semantic and phonologic processing in lexicalisation is supported by evidence from picture-naming tasks with distractors: The time window in which auditory stimuli that were phonologically related to the item had to be presented to slow down naming differed from the time window in which semantically related stimuli interfered with naming (note that both could, presented in other time windows, also speed processing). According to these findings it takes roughly 150 ms to process the picture and activate a concept, about 125 ms to select the lemma and another 250 ms for phonological processing (Levelt et al., 1991). Other researchers argue that there is overlap between these phases, allowing for *processing in cascade*: Information from semantic processing can be utilised for phonological processing even before lemma selection is completed. Peterson and Savoy (1998) found mediated priming in picture naming tasks, which means that presentation of a phonological relative of a semantic relative of the target word (like “soda” related to the target “couch” via “sofa”) at a certain moment in time, facilitated processing. Another finding in favour of processing in cascade is that word substitution errors in which the inserted word is related both semantically and phonologically to the intended one (like catalogue to calendar) occur above chance level (Harley, 2008). Controversy goes even further, questioning the existence of lemmas. As an alternative model Caramazza (1997) proposes a lexical-semantic, a syntactic and a phonological network between which information is exchanged during lexicalisation.

**Lexicalisation** Word selection in speech production.

**Lemma** A representation of the meaning and syntactic properties of a word that does not contain its phonological features.

**Lexeme** The representation of a word’s phonological form.

### Grammatical planning

For each word, grammatical features become accessible by lemma selection (or activation of relevant elements in the syntactic network according to Caramazza’s model), constraining opportunities of integrating it into the sentence. Each word can be conceptualised as a knot in a syntactic network, and to complete the sentence structure, a pathway that connects all those knots has to be found. Idiomatic terms are a special case as they are connected to very strong constraints. Therefore it is presumed that they are stored as separate entries (in addition to the entries for the single words they are constituted of) in our mental lexicon (Levelt, 1999). In many languages, also the **morphological** form of a word has to be defined to integrate it into the sentence, taking into account its syntactic relations and additional information the word contains (like tense and number). Morphological transformations can be implemented by addition of affixes to the word stem (like ‘speculat*ed*’ or ‘plant*s*’) or by changes of the word stem (like ‘swim-swam’ or ‘mouse-mice’). The amount and complexity of morphological transformations in English is moderate compared to languages like German, Russian or Arabian, while in other languages like Chinese there are no morphological transformations at all.

**Morphology** The capability of words to adapt different grammatical forms with distinct phonological forms.

### Articulation

When the phonological information about the words in their appropriate morphological form is available and word order has been determined, articulation can begin. Keep in mind that these processes are incremental so the sentence need not be prepared as a whole before its beginning is articulated. The issue is to produce the required speech sounds in the right order with the right prosody. There are different models of how this is achieved. *Scan-copier models* (Shattuck-Hufnagel, 1979) constitute a classic approach, proposing that a framework of syllable structure and stress pattern is prepared. Phonemes are inserted into this framework by a ‘copier’ module and the progress is instantly checked. Speech errors like phoneme exchange, phoneme deletion or perseveration can be explained by failures in certain points of the copy- and checking processes. According to the *competitive queuing model* (Hartley & Houghton, 1996), which adopts the idea of a framework and a copier, the phonemes that are to be inserted form a queue, and the order of insertion is controlled by activating and inhibiting connections between them and particular units that mark the beginning and the end of a word. Thus, the phoneme with the strongest connection to the beginning unit will be inserted in the first position.

***The role of syllables for articulation***

WEAVER++ (Levelt, 2001) is a two-step model that assumes that by lexeme identification, a sequence of phonemes representing the whole word is drawn simultaneously. This is supported by the finding that in naming tasks, auditorily presented distractors that prime parts of the target word speed naming, no matter which position the primed part has in the target word (Meyer & Schriefers, 1991). As the next step, syllables, which are not a part of the lexicon representation, are formed sequentially. Because of co-articulation, syllables are required as input for the articulatory process. The formation of syllables is believed to be facilitated by a reservoir of frequent syllables, the **syllabary**. Even in languages like English with a high number of different syllables (more than 12.000) a much smaller number accounts for most syllables in a given utterance. Those syllables (and in languages with only a few hundred different syllables, like Chinese or Japanese, maybe all syllables) form highly automatised motor sequences that could (according to Rizzolatti & Gentilucci, 1988) be stored in the supplementary motor area. A finding in favour of the existence of a syllabary is that pseudo-words (constructed from normal Dutch syllables) with high-frequency syllables were processed faster than pseudo-words with low-frequency syllables in an associative learning task (Cholin, Levelt & Schiller, 2006). Syllable forming can also depend on prosody. In stress-assigning languages like English phonological words are formed by associating unstressed syllables with neighbouring stressed syllables. These phonological words seem to be prepared before speaking begins, as for sentences including more phonological words the onset of speaking takes longer. In articulation, syllables bind together only within, but not across phonological words. For example, in the sentence “Give me a beer, if the beer is cold” the ‘r’ of beer is bound to the following ‘i’ only in the second part of the sentence (“bee-ris cold”), as the comma marks a boundary between phonological words (Harley, 2008). The example also shows that syllables are not determined by lexical words, as phonemes can change from the syllable they are part of when the lexical word stands alone to a syllable belonging to another lexical word.

**Syllabary** A “dictionary” of frequent syllables that is used for syllable preparation in speech production.

***Acoustic speech parameters***

During articulation, we not do only manipulate the phonematic properties of the sounds we produce, but also parameters like volume, pitch, and speed. Those parameters depend on the overall prosody of the utterance, respective the position of a given syllable within it. While prosody can be regulated directly in order to transport meaning independent of the words used (consider that different stress can make the same sentence sound like a statement or like a question), some acoustic parameters can give hints about the speakers emotional state: *Key* is the variation of pitch within a phrase and is influenced by the relevance of the phrase to the speaker and his emotional involvement. *Register* is the basic pitch and is influenced by the speakers current self-esteem (use of the lower chest register indicating higher self esteem than use of the head register), (Levelt, 1999).

***Monitoring of speech production***

According to the standard model of speech production (Levelt, 1999), monitoring takes place throughout all phases of speech production. Levelt assumes that for monitoring of syntactic arrangement we utilise the same ‘parsing’ mechanisms we employ to analyse the syntax of a heard sentence. Although speech production and speech comprehension involve different brain areas (there is activation in temporal auditory regions during listening and activation in motor areas during speaking; see the chapter about the biological basis of language), monitoring of one’s own speech also seems to involve temporal areas included in listening to other persons. Therefore a ‘perceptual loop’ for phonetic monitoring has been proposed (Levelt, 1999), although it is not clear yet whether this loop processes the auditory signal we produce or some earlier phonetic representation, a kind of ‘inner’ speech.

## Summary

Speech comprehension starts with the identification of the speech signal against an auditory background and its transformation to an abstract representation, also called decoding. Speech sounds are perceived as phonemes, which form the smallest unit of meaning. Phoneme perception is not only influenced by the acoustic features, but also by the word and sentence context. To analyse its meaning, it is necessary to segment the continuous speech signal. This is done with the help of the rhythmic pattern of speech. In the following processing step of word identification, the prelexical code that only contains phonetic information of a word is complemented with the postlexical code, that is, the semantic and syntactic properties of the word. It is proposed that a mental dictionary exists, the lexicon, from which candidates for the heard word are singled out. With the integration of the postlexical code of the single words, the meaning of the sentence can be deciphered. The endpoint of speech comprehension – the conceptual message – is the starting point of speech production. Ideas have to be organised in a linear form, as speech is a one-dimensional sequence, and have to be expressed in syntactic relations. Words for the selected concepts have be chosen, a process called lexicalisation, which is the reverse of word identification because here the semantic and syntactic representation of the word (the lemma) is selected first and has to be linked to the phonemic representation (the lexeme). The syntactic properties of the single words can be seen as constraints for their integration into the sentence, so a syntactic structure has to be constructed that meets all constraints. Also the morphological forms of the words have to be specified before the sentence can be encoded phonologically for articulation. To plan articulation, syllables are constructed from the lexical words in tune with the phonological words that result from the sentence’s stress pattern. Generally, speech production is an incremental process, which means that articulation and different stages of preparation for the following phrases take place simultaneously.