Chapter 9: Reading and speech perception

Humanity excels in its command of language. Indeed, language is of such enormous importance to us that this chapter and the following two are devoted to it. In this chapter, we consider the basic processes involved in reading and speech perception. It often does not matter whether a message is presented to our ears or to our eyes. Many comprehension processes are very similar whether we are reading a text or listening to someone talking.

However, reading and speech perception differ in important ways. In reading, each word can be seen as a whole, whereas a spoken word is spread out in time and is transitory. More importantly, it is much harder to tell where one word ends and the next starts.

The fact that reading and listening to speech are quite different in some ways can be shown by considering children and brain-damaged patients. Young children often have good comprehension of spoken language, but struggle to read even simple stories. Some adult brain-damaged patients can understand spoken language but cannot read, and others can read perfectly well but cannot understand the spoken word.

- **Reading: introduction**

Learning English is slower than other languages because it is not consistent (Caravolas et al., 2013). There are many methods designed to study reading:

- The *lexical decision task* (deciding whether a string of letters forms a word).

  **WEblink:** *Lexical decision task*

- The *naming task* (saying a word out loud as rapidly as possible). It is not clear precisely what processes are reflected in these tasks.

- A generally useful method involves recording eye movements, which provides a detailed record of attention-related processes and is unobtrusive.

- In the *priming task*, a prime word is present shortly before the target word. The prime is related to the target word (e.g. in spelling, meaning or sound).

- Monitoring brain activity.

Balota et al. (1999) argued that reading involves several kinds of processing:

- orthography (spelling);
- phonology (sounds);
- semantics (meaning);
- syntax;
- and higher-level discourse integration.

A common view is that *phonological processing* of visual words is relatively slow and inessential for word identification (Coltheart et al., 2001). There is support for the assumption that phonological processing is important when identifying words. Van Orden (1987) used *homophones*. He found that participants made more errors on questions involving homophones, demonstrating that they engaged in phonological processing of words. Words with many phonological neighbours have a processing advantage in word recognition (Yates et al., 2008). Words are processed faster when preceded by masked primes that are similar in phonology than by primes that are similar in spelling (Rastle & Brysbaert, 2006).

**CASE STUDY:** Are syllables phonological units in visual word recognition?

Several methods have been used to study reading. The lexical decision and naming tasks have been used to
assess word identification. These tasks ensure that certain processes have occurred, but have the disadvantage that normal reading processes are disrupted. A generally useful method involves eye movements, because it provides detailed online information and is unobtrusive. Priming tasks and neuroimaging studies have also contributed to our understanding of reading. There is support for the view that phonological processing plays an important role in visual word recognition. However, the strong phonological model is probably too strong and visual word recognition can occur in the absence of phonological processing.

- **Word recognition**

College students typically read at about 300 words a minute, thus averaging only 200 ms per word. Rayner and Sereno (1994) argued that word identification is fairly automatic, which makes sense considering that a student reads 20–70 million words in his or her life!

McClelland and Rumelhart (1981) proposed an influential *interactive activation model* of visual word recognition. The key assumption is that bottom-up and top-down processes interact. Bottom-up processes stemming directly from the written word proceed from the feature level, through the letter level, to the word level by means of activation and inhibition processes going from word to letter level.

The *word superiority effect* (Martin et al., 2006) suggests information about the word presented can facilitate identification of letters in the word. The word superiority effect occurs because activation at the word level (e.g., SEAT) will increase activation of the letter (e.g., A).

**WEBLINK:** [Examples of the word superiority effect in reading](#)

Time to identify a word depends in part on its *orthographic neighbours*. Orthographic neighbours typically *facilitate* word recognition if they are less frequent in the language than the word itself (Chen & Mirman, 2012). However, they have an *inhibitory* effect if they are more frequent than the target word.

The interactive activation model has been very influential. It provides an interesting example of how a connectionist processing system can be applied to visual word recognition, and can account for various phenomena including word superiority. However, the model does not provide a comprehensive account of word recognition, and does not address the role of factors such as phonological processing, meaning or relevant context.

*Semantic priming* is when a word is recognised or identified more rapidly if immediately preceded by a semantically related word (Meyer & Schvaneveldt, 1971). Neely (1977) found two priming or context effects:

- a rapid, automatic effect based on semantic relatedness;
- a slower, attentional effect based on expectations.

Readers use *context* to predict the next word and so context affects the *early* processing of that word (Kutas et al., 2011). Context can affect lexical access to stored information in the lexicon. In an ERP study, Penolazzi et al. (2007) found evidence that context affects processing of the target word within 200 ms – suggesting context affects lexical access to the target word.

Word identification is fairly automatic and is not constrained by letter identification. According to the interactive activation model, bottom-up and top-down processes are both involved in letter identification and word recognition. This model has been very influential and accounts for various phenomena such as the word superiority effect. However, it is not a comprehensive account of visual word recognition. The semantic context can play an important role in word recognition.

- **Reading aloud**
Coltheart et al. (2001) proposed a dual-route model designed to account for reading aloud and silent reading. It is a cascade model because activation at one level is passed on to the next before processing at the first level is complete. There are two main routes between printed word and speech, both starting with orthographic analysis:

- **Route 1** (grapheme–phoneme conversion):
  Converting letters into sounds.
  Patients with surface dyslexia have problems reading irregular words.
  McCarthy and Warrington (1984) studied KT, a patient with surface dyslexia who displayed regularisation over irregular words.

- **Route 2** (lexicon+semantic knowledge) and **Route 3** (lexicon only):
  Dictionary look-up. Visual word presentation leads to:
  - activation in the input lexicon;
  - obtaining its meaning from the semantic system;
  - generation of the sound pattern by the phonological output lexicons.
  Route 3 also involves orthographic input and phonological output lexicons, but bypasses the semantic system.

Patients using Route 2 or Route 3 can pronounce familiar words, whether regular or irregular, but find it hard to pronounce unfamiliar words and non-words. Patients with phonological dyslexia fit this pattern.

**INTERACTIVE EXERCISE:** Dual-route cascaded model

*Deep dyslexia* is a condition where there are problems with reading unfamiliar words, an inability to read non-words, and semantic reading errors (e.g., “ship” is read as “boat”). It occurs as a result of brain damage to the left-hemisphere brain areas involved in language (e.g. areas involved in grapheme–phoneme conversion and semantic systems).

Neuroimaging studies have also revealed individual differences in reading aloud. Jobard et al. (2011) found fewer brain areas were activated during reading by those with high working memory capacity than those with low capacity because of their more efficient processing. Of most interest, only those with low capacity had activation in areas associated with grapheme–phoneme conversion. Seghier et al. (2008) found the left anterior occipito-temporal region was associated with reading irregular words. The left posterior occipito-temporal region was associated with reading pseudowords.

The model’s limitations are:

- The assumption that time taken to pronounce a word depends on regularity rather than consistency is incorrect.
- The absence of learning.
- The assumption that only the non-lexical route is involved in pronouncing non-words.
- The assumption that phonological processing of visually presented words occurs slowly.
- The assumption that the semantic system can play an important role in reading aloud.
- It is not applicable to Chinese, Japanese and Korean writing systems.

The *connectionist triangle model* (Plaut et al., 1996) is a single-route connectionist approach, which assumes that the pronunciation of words and non-words is based on a system that is highly interactive during learning. The three sides of the triangle are orthography, phonology and semantics. Two routes are proposed from spelling to sound:

- a direct pathway from orthography to phonology;
- an indirect pathway from orthography to phonology via word meaning.

Plaut et al. (1996) argued that words vary in consistency. Highly consistent words and non-words are pronounced faster and more accurately because more available knowledge supports correct pronunciation.
Semantic knowledge is most likely to have an impact for inconsistent words since they take longer to name. According to the model, semantic factors are especially important for words that are irregular or inconsistent. McKay et al. (2008) found that, for inconsistent non-words, reading aloud was faster in the semantic condition than in the non-semantic condition. According to the model, inconsistent words will take longer to pronounce. There is clear experimental support for this assumption, supporting the triangle model over the dual-route model. The connectionist model is also more successful in accounting for consistency effects with non-words, and includes an explicit mechanism to simulate how we learn to pronounce words. There are various limitations:

- Connectionist models tend to focus on the processes involved in reading simple single-syllable words.
- There is limited description of semantic processing.
- Explanations of phonological and surface dyslexia are oversimplified.

**INTERACTIVE EXERCISE: Dual-route reading**

According to the dual-route cascaded model, lexical and non-lexical routes are used in naming words and non-words. Patients with surface dyslexia rely mainly on the lexical route, whereas patients with phonological dyslexia use mostly the non-lexical route. The dual-route model reads words and non-words very accurately, and provides a good account of surface and phonological dyslexia. However, the model fares less well when predicting human naming times, and makes several assumptions that are unsupported. Plaut et al. (1996) proposed a single-route connectionist model in which words and non-words are processed by the same highly interactive system. The model accounts for most findings, particularly the effect of word consistency on word naming speed, and provides a good account of deep dyslexia. However, it provides sketchy accounts of some important issues (e.g., phonological dyslexia; the nature of the semantic system).

- **Reading: eye-movement research**

Our eyes move in rapid jerks (*saccades*). Saccades are ballistic (once initiated, direction cannot be changed). When reading there are frequent regressions during which eyes move back over text. Saccades take 20–30 milliseconds and are separated by fixations of 200–250 ms. Rayner et al (2012) studied the amount of text scanned in each fixation using the “moving window” technique. The *perceptual span* (effective field of view) is affected by the difficulty of the text and print size. Perceptual span often extends three to four letters to the left of fixation and up to 15 to the right. The asymmetry is clearly learned because readers of languages that read right to left (e.g., Arabic) have the opposite pattern (Pollatsek et al. 1981). Reading time on a target word is less when the preview word is visually or phonologically similar to the target word, suggesting visual and phonological information can be extracted from parafoveal processing. Readers typically fixate about 80% of content words and 20% of function words. Fixation time on a word is longer when it is preceded by a rare word – this is the *spillover effect*.

Reichle et al. (1998) and Rayner et al. (2012) have accounted for patterns of eye movements in reading in versions of their *E-Z Reader model*. According to the E-Z Reader model, the next eye movement is programmed after only part of the processing of the currently fixated word has occurred. This reduces time between completion of processing on the current word and eye movement to the next word. The spillover effect arises because there is less spare time available with rare words. According to the model, readers can attend to two words during a single fixation. However, given serial processing, only one word is processed at a time for meaning. The model has several key assumptions:

- Readers check the familiarity of the current fixated word.
- Completion of frequency checking of a word is the signal to initiate an eye-movement programme.
- The second stage is *lexical access*, which involves accessing a word’s semantic and phonological forms. This task takes longer to complete.
- Completion of the second stage is the signal to shift covert attention to the next word.
- Frequency checking and lexical access are completed faster for common and predictable words.

The model has been successful in specifying major factors determining eye movements in reading. The main
limitations are:

- It emphasises early processes involved in reading and has little to say about higher processes.
- It fails to explain parafoveal-on-foveal effects.
- Doubts have been raised concerning assumptions about serial processing and processing words in the “correct” order.
- The model fails to account for context effects.
- The emphasis of the model is on explaining eye movement data rather than findings on reading.
- The model places importance on word frequency as a determinant of eye fixation time. Word predictability may be more important than word frequency.

WEBLINK: Eye movements when reading

The least obtrusive way of studying reading is by eye-movement recordings. According to the E-Z Reader model, the next eye-movement is programmed when only part of the processing of the currently fixated word has occurred. Completion of frequency checking of a word is the signal to initiate an eye-movement programme, and completion of lexical access is the signal for a shift of covert attention to the next word. This model takes insufficient account of the impact of higher-level cognitive processes on fixation times, exaggerates the importance of word frequency and has not been applied systematically to findings on reading using techniques other than eye-movement recordings.

- Speech perception: introduction

Speech perception is easily the most important form of auditory perception. Spoken language can be considered a special type of music (Brandt et al., 2012). Speech perception differs from other kinds of auditory perception in that there is left-hemisphere dominance. Speech stimuli intermediate between two phonemes are typically categorised as one or the other with an abrupt boundary (categorical perception). For example, /l/ and /r/ belong to the same category in Japanese and listeners cannot distinguish between them. Raizada and Poldrack (2007) found that differences in brain activation for /ba/ and /da/ stimuli along a continuum were amplified when they were on opposite sides of the category boundary. This suggests categories are important in speech perception. Categorical perception also exists in music perception (Locke & Kellar, 1973).

CASE STUDY: Categorical perception in American Sign Language

Husain et al. (2006) found various speech and non-speech sounds activated similar regions in the primary and non-primary areas of the temporal cortex, intraparietal cortex and frontal lobe. In addition, there were only minor differences in patterns of brain activity when speech and non-speech sounds were processed. Using a more complex task, however, Rogalsky et al. (2011b) obtained rather different findings.

The processing stages for speech perception are:

1. decoding speech signals;
2. identification of phonemes or syllables;
3. word identification;
4. interpretation;
5. integration with context.

- Listening to speech

Understanding speech is much less straightforward than one might imagine. Listeners in everyday life also have to cope with perceiving speech under various adverse conditions (Mattys et al., 2012) such as energetic masking (background noise) and informational marking. There are additional problems faced by listeners:

- The segmentation problem is the difficulty of separating out words from the pattern of speech sounds. This arises because speech is a continuously changing pattern of sound.
Coarticulation is the overlapping of adjacent articulations, i.e., the way a phoneme is pronounced depends on phonemes before and after it. This also provides information about the surrounding phonemes.

Listeners have to contend with significant individual differences in the rate of speaking, and in everyday life listeners have to contend with degraded speech due to distractions of background noise.

Language is spoken at a rate of about ten phonemes per second and requires rapid processing.

Dividing the speech they hear into its constituent words (i.e., segmentation) is a crucial task for listeners. Segmentation involves using several cues. Some are acoustic-phonetic (e.g., coarticulation, stress), whereas others depend on the listener’s knowledge (e.g., of words) and the immediate context (Mattys et al., 2012). Mattys et al. (2005) proposed a hierarchical approach, with three main categories of cues:

- lexical cues;
- segmental cues such as coarticulation and allophony;
- metrical prosody cues.

Listeners (even with normal hearing) often make extensive use of lip-reading. McGurk and MacDonald (1976) played a video of someone saying “ba” when the audio was of a voice saying “ga”. Participants reported hearing “da”, a blending of visual and auditory information. This is the McGurk effect. Both bottom-up and top-down processes (e.g., expectations) are important for the effect. Windmann (2004) found more participants produced the McGurk effect when the crucial word was presented in a semantically congruent sentence, indicating a role for top-down processes.

WEBLINK: McGurk effect
WEBLINK: Demonstrations (including the McGurk effect)

Listeners make use of various prosodic cues. They also use lip-reading, even when the visual information conflicts with the sound presented at the same time (the McGurk effect). Among the difficulties faced by listeners are the speed of spoken language, the segmentation problem, coarticulation and individual differences in speech patterns. Listeners cope by taking account of possible-word constraints, stress patterns within words and the fact that coarticulation is generally greater within than between words. There is categorical perception of phonemes, but we can discriminate between sounds categorised as the same phoneme. Ambiguous phonemes are more likely to be assigned to a given phoneme category when that produces a word than when it does not, and this lexical identification shift seems to be perceptual.

- Context effects

Context typically influences spoken word recognition (Samuel, 2011). Context can affect spoken word recognition according to two positions (Harley, 2013):

- The interactionist account claims contextual information can influence processing at an early stage.
- The autonomous account claims context has its effects late in processing.

When the sentence context makes a word improbable it is fixated upon more than when it is probable and this difference occurs early on in processing (Brock & Nation, 2014). This indicates that sentence context has almost immediate effects on word processing.

Warren and Warren (1970) studied the phonemic restoration effect. Participants heard a sentence in which a small portion had been removed. Warren and Warren found that perception of the missing element was influenced by sentence context. Shahin et al. (2009, 2012) found two processes were involved in this effect:

- unconscious sensory repair;
- subjective continuity (using top-down expectations).

Ganong (1980) wondered whether categorical perception of phonemes would be influenced by context. He
found an ambiguous initial phoneme was more likely to be assigned to a given phoneme category when that produced a word – this is the Ganong effect.

Sohoglu et al. (2014) found the perceived clarity of a degraded spoken word was greater when preceded by written text containing that word, but the perceived clarity was unaffected when a written text was presented after the word. This effect lends support for the interactionist account.

- **Theories of speech perception**

Perre and Ziegler (2008) gave listeners a lexical decision task. The words varied in the consistency between their orthography or spelling and their phonology. Listeners performed the lexical decision task slower when the words were inconsistent. Thus, spoken word processing is disrupted when there is a mismatch between phonological and orthographic information. This suggests orthographical influences exist in spoken word recognition.

Liberman et al. (1967) proposed the *motor theory* of speech perception. Listeners mimic the articulatory movements of the speaker. The motor signal produced is claimed to produce less variable and inconsistent information than the speech signal itself. Pulvermüller and Fadiga (2010) argued along similar lines that language processing (including speech perception) involves action-perception circuits. TMS impaired auditory processing of changes in speech sounds as indicated by ERPs (Möttönen et al., 2013), but only when the speech signal is impoverished (D’Ausilio et al., 2012).

There has been accumulating evidence supporting the motor theory of speech perception in recent years. However, the importance of motor processes should not be exaggerated. The underlying processes are not spelled out. Individuals with severely impaired speech production nonetheless have reasonable speech perception. Infants who have limited expertise in speech articulation still perform reasonably well on syllable detection tasks.

McClelland and Elman (1986) and McClelland (1991) produced a connectionist network of speech perception called the *TRACE model*. The TRACE model assumes that bottom-up and top-down processes interact flexibly throughout speech perception:

- Feature nodes are connected to phoneme nodes, and phoneme nodes are connected to word nodes.
- Connections between levels operate in both directions and are always facilitatory.
- There are connections among units or nodes at the same level; these connections are inhibitory.
- Nodes influence each other in proportion to their activation levels and the strengths of their interconnections.
- As excitation and inhibition spread among nodes, a pattern of activation or trace develops.
- The word recognised or identified by the listener is determined by the activation level of the possible candidate words.

Mirman et al. (2008) found the magnitude of the word superiority effect was greater when 80% of the stimuli were words than when only 20% were. This provides evidence for top-down processes in speech perception. High-frequency words (those often encountered) are generally recognised faster than low-frequency ones (Harley, 2013). The TRACE model can easily explain lexical identification shift (Ganong, 1980). Norris et al. (2003) found evidence that phoneme identification can be directly influenced by top-down processing. According to the model, high-frequency words are processed faster than low-frequency words (e.g., due to higher resting activation levels). Top-down effects seem to be less important than assumed within the model (Frauenfelder et al., 1990) and they also depend more on stimulus degradation than predicted by the model (McQueen, 1991). In an ERP study, Perre and Ziegler (2008) found inconsistency between phonology and orthography affected speech perception.

The TRACE model has various successes. It provides reasonable accounts of phenomena such as categorical
speech perception and the word superiority effect in phoneme monitoring. The assumption that top-down and bottom-up processes both contribute is a strength. The model predicts the effects of word frequency on auditory word processing and correctly predicts that words with many lexical neighbours will be harder to recognise. However, the model exaggerates the importance of top-down effects. Evidence suggests speech processing is less flexible and unstructured than implied by the model. Tests of the model have relied on computer simulations involving a small number of one-syllable words. It is not clear whether the model would perform satisfactorily if applied to larger vocabularies. The model ignores some factors influencing auditory word recognition such as orthographic information and non-phonemic information.

Marslen-Wilson and Tyler (1980) proposed the original version of the cohort model. According to the model, various knowledge sources (e.g., lexical, semantic, syntactic) are processed in parallel and interact in complex ways to produce efficient analysis of spoken language. There are main assumptions:

- Early in the auditory presentation of a word, all words conforming to the sound sequence heard so far become active; this set of words is the cohort. There is competition among these words to be selected.
- Words within the cohort are eliminated if they cease to match further information from the presented word or because they are inconsistent with the semantic or other context. For example, crocodile and crockery might both belong to the initial cohort with the latter word being excluded when the sound /d/ is heard.
- Processing continues until information from the word itself and contextual information are sufficient to permit elimination of all but one of the cohort words. The uniqueness point is the point at which only one word is consistent with the acoustic signal.
- It is assumed words vary in their level of activation and so membership of the word cohort is a matter of degree.
- Marslen-Wilson and Tyler assumed the word-initial cohort may contain words having similar initial phonemes rather than only words having the same initial phoneme as the presented word.

Marslen-Wilson and Tyler tested their theory in a word-monitoring task. The results conformed closely to the predictions of the model. They found it was only necessary to listen to the entire word when the sentence contained no useful syntactic or semantic information. Further support comes from O’Rourke and Holcomb’s (2002) ERP study in which the N400 (a measure of word processing) was earlier for words having an early uniqueness point. However, Radeau et al. (2000) cast doubt over the importance of the uniqueness point. The uniqueness point influenced performance only when the nouns were presented at a rate slower than the typical conversational rate. Van Petten et al. (1999) found very strong context can influence spoken word processing earlier than expected within the revised cohort model.

The cohort model has proved to be an influential approach to spoken word recognition. There are various limitations with the revised model: Modifications made to the original model have made it less precise and harder to test. Context sometimes influences relatively early processing. It is not clear how the starting points of individual words are identified.

According to the motor theory of speech perception, recruitment of the motor system facilitates speech perception. There has been accumulating evidence supporting this theory in recent years. According to the original version of cohort theory, the initial sound of a word is used to construct a word-initial cohort, which is reduced to one word by using additional information from the presented word and from contextual information. Cohort theory has been revised to make it more flexible and in line with the evidence. According to the TRACE model, bottom-up and top-down processes interact during speech perception. This assumption that these processes interact is probably incorrect, and the importance of top-down processes is exaggerated in the TRACE model. Non-phonemic information may be more important than assumed within the TRACE model.

**Cognitive neuropsychology**

Many brain-damaged patients experience difficulties with word repetition. Ellis and Young (1988) used
information from brain-damaged patients to propose a model of the processing of spoken words. The model has five components:

- auditory analysis system;
- auditory input lexicon;
- semantic system;
- speech output lexicon;
- phoneme response buffer.

These components can be used in various combinations to make three different routes between hearing a spoken word and saying it.

If a patient only had damage to the **auditory analysis system** there would be a deficit in phonemic processing. This patient would have impaired speech perception for words and non-words, but intact speech production, reading and writing, and normal perception of non-verbal sounds. Several patients apparently conform to this pattern, which is known as **pure word deafness**. Supporting evidence was reported by Slevc et al. (2011). NL, a patient with pure word deafness, had particular difficulties in discriminating sounds (speech or non-speech) differing in rapid temporal changes.

An assumption of the **three-route framework** is that there are three routes that individuals can use to process and repeat words. All routes involve the auditory analysis system and phonemic response buffer. Route 1 also involves the other three components (auditory input lexicon, semantic system, speech output lexicon). Route 2 involves two additional components (auditory input lexicon, speech output lexicon) and Route 3 involves an additional rule-based system that converts acoustic information into words that can be spoken.

If patients could only use Route 2:

- They should be able to repeat familiar words but would often not understand their meaning.
- They would also have problems with unfamiliar words and non-words.
- However, they should be able to distinguish between words and non-words.
- This condition is **word meaning deafness**.
- However, the notion is controversial and relatively few patients have been identified.
- Franklin et al. (1996) studied Dr O, who was a clear case of word-meaning deafness.

Patients who make use primarily or exclusively of Route 3:

- would be reasonably good at repeating words and non-words but have poor comprehension of these words;
- this condition is known as **transcortical sensory aphasia** (e.g., Kim et al., 2009).

Some brain-damaged patients have extensive problems with speech perception and production. Patients with **deep dysphasia** make semantic errors when asked to repeat spoken words (i.e., saying words related in meaning to those spoken). They also find it harder to repeat abstract than concrete words and have poor ability to repeat non-words. It could be argued that none of the three routes is intact. Jefferies et al. (2007) argued that the central problem in deep dysphasia is a general phonological impairment.

**INTERACTIVE EXERCISE:** Ellis and Young’s (1988) three-route model

Limitations of the framework are that it is difficult to relate deep dysphasia to the framework and the status of some rare conditions is controversial.

There are three routes between sound and speech. Patients with pure word deafness have problems with speech perception, which may be due to impaired phonemic processing or may reflect a more general impairment. Patients with word meaning deafness can repeat familiar words without understanding their meaning, but have problems with non-words. Deep dysphasia may reflect damage to all three routes, or a general phonological deficit.
**Additional references**


