

## Spoken Word Recognition

**Helen Fraser**

University of New England  
helenbfraser@gmail.com

### Introduction

One of the most basic aspects of learning a language is learning its words, or vocabulary. Applied linguistics rightly gives a good deal of attention to understanding the processes by which word meanings are learned, in both first and second language acquisition. But of course words are not just meanings. Each word, as well as a characteristic meaning, also has a characteristic sound, or pronunciation, which is just as important for learners to understand as its meaning. However, the processes by which the sounds of words are recognized tend to be given less attention in applied linguistics.

This may be because recognizing spoken words seems so easy in everyday experience. The difficult issue, it seems, is recognizing written words, especially in languages like English, with irregular spelling. However, the reason spoken word recognition seems easy is because it is highly practised and very familiar, not because it is simple – as becomes evident in second language contexts, when learners struggle to recognize even the simplest words (such as *red* or *led*).

In fact, spoken word recognition is one of the most complex skills of human cognition, and the foundation of other crucial skills, especially (since words must be recognized before they can be reproduced) of pronunciation. Applied linguistics really needs a solid understanding of the complex nature of spoken word recognition, framed in a theory that offers practical guidance on how to respond to a variety of common problems in teaching and learning.

This article outlines some of the key findings of research on spoken word recognition and suggests how understanding this topic is relevant to applied linguistics, especially second language teaching.

## The nature of speech

Spoken word recognition is commonly assumed to be a straightforward process of identifying phonemes in the acoustic speech wave, and putting them together to form meaningful words. It may seem odd to call such an apparently obvious idea a theory, but it functions as a theory in the sense that it provides a conceptual framework which guides the understanding of many observations about spoken language.

The problem is, this apparently obvious theory rests on the assumption that words in speech are discrete (have clear boundaries), invariant (essentially the same each time they are spoken), and composed of a sequence of phonemes (individual speech sounds) that are themselves discrete and invariant – in much the same way as words on a printed page are made up of discrete, invariant letters grouped into discrete invariant words (but without the irregularity of spelling).

As was discovered in the 1950s, when attempts to build computer speech systems on the basis of this everyday theory resulted in spectacular failure, these assumptions about speech are incorrect. Despite the impression given by everyday experience, speech is not a sequence of discrete invariant phonemes, but a continuous stream of sound. This is hard to demonstrate convincingly without audio examples, but it is a very well known fact of phonetics. In reality, there are no clear boundaries systematically segmenting the speech stream into words and phonemes (Port, 2007). In addition, unlike printed letters, which have an invariant form (e.g. the letter 'r' has the same shape each time it appears on this page), each individual phoneme covers a whole category of highly variable forms (called allophones), not normally noticed by ordinary speakers. For example, the phoneme /r/ is pronounced with distinctly differently allophones in 'rain' and 'train'. In fact, phonemes show so much variation (unnoticed in everyday perception), it is fair to say they are different every time they are pronounced (Ladefoged, 2005), and in ways far more complex than suggested by the very basic examples that can be given here. The effect is that, far from being like printed letters, even the clearest speech is like very messy handwriting – except that, unlike handwriting, speech lacks gaps not just between letters, but also between words.

These characteristics of speech create what are known as the ‘segmentation’ and ‘invariance’ problems, which together create what became the defining question of spoken word recognition research for many decades (Perkell & Klatt, 1986): How do people categorize the continuous stream of speech into phonemes to allow word recognition – despite the segmentation and invariance problems?

## **Theories of spoken word recognition**

Currently several broad classes of theory of spoken word recognition are available to address this question. One, originating from early ‘motor theory’ (Lane, 1965), claims that invariant units do in fact exist in speech – not in the form of acoustically defined phonemes, but rather in their underlying physical gestures (movements of tongue, lips, etc.), which the human perceptual system is geared by evolution and experience to pick up directly. A major problem faced by this class of theories is difficulty in identifying such invariant gestures in the stream of speech. So while more sophisticated versions of this idea survive as ‘direct realist’ or ‘ecological’ theories (Fowler, 2003), motor theory as such has mostly fallen aside. It did, however, make a lasting contribution with its concept of ‘categorical perception’ – a postulated mechanism whereby acoustically different allophones (such as those in ‘rain’ and ‘train’) can be categorised as the same phoneme (/r/) at low levels of mental processing, without conscious awareness (Repp, 1984).

A second class of theories of spoken word recognition, currently the most widely accepted, tend to be called ‘cognitive’ in applied linguistics (e.g. Zuengler and Miller, 2006). A better term, however, is ‘computational’, since, following the lead of Chomsky and others, these theories model the human mind as a kind of computer (Pinker, 2003). Rather than attributing the apparent invariance of phonemes to invariant gestures, as in motor theory, these theories see spoken word recognition as a process that takes a continuous, variable speech wave as input, and computes (subconsciously) invariant representations suitable for ‘lexical access’ (the process by which a word’s meaning is ‘looked up’ in a ‘mental lexicon, or dictionary’ (Byrd & Mintz, 2010)). Categorical perception is considered an important part of this mental processing, and has been investigated extensively, with considerable development in its theoretical conceptualization.

One issue that emerged early in the development of computational theories is that they require the phonological form of a word (i.e. its pattern of phonemes, syllables and features) to be determined before lexical access – much as an orthographic form (spelling) is needed before a word's meaning can be looked up in a physical dictionary. Reasonable as this may seem from the perspective of the everyday theory of word recognition, many experiments over the past half century or more have shown that recognition of the meaning of entire words often precedes, overrides or is independent of recognition of individual phonemes (Byrd & Mintz, 2010). For example, words can be recognized even if significant portions of their phonological form are missing or altered; words containing exactly the same units of sound can be interpreted very differently depending on the context in which they are heard (not just the preceding context; the following context can retrospectively influence recognition of earlier words); parts of words, or even whole sequences of words, which sound quite clear in meaningful context are frequently unrecognizable if heard in isolation. Such phenomena are called 'real word effects', and are similar to the more widely known 'word superiority effect' of visual word recognition, which acknowledges that, in reading, it is easier for people to recognise whole words than individual letters.

Observations of real word effects, along with other considerations, have caused computational theories of word recognition to be elaborated to include processes that allow meaning and context to influence word recognition via interaction of 'top-down' information (from meaning and context), with 'bottom-up' information (from the acoustic speech signal). However, while there is general agreement about the need to incorporate top-down processing into word recognition models, there is considerable debate about exactly how this is achieved. The result is there is now a range of quite complicated theories of spoken word recognition, postulating different phonological units (some rejecting the role of the phoneme in favour of larger units like syllables, or smaller units like phonological features), and different types of computation (some rejecting the role of rules in favour of more distributed processes).

## **Spoken word recognition in applied linguistics**

Computational theories have provided a great deal of knowledge about speech in general and word recognition in particular, and are useful for predicting and explaining observations about the spoken word recognition behaviour of proficient speakers. Unfortunately they tend to be rather less useful when applied to teaching and learning situations. Originating during the post-behaviorist split between theoretical and applied linguistics, these theories were developed through computational modeling rather than through direct engagement with teaching and learning practice. The mental processes they postulate tend to be abstract, disembodied, divorced from social and cultural context, and, most importantly, beyond the reach of conscious control, which limits their relevance to classroom teachers seeking to influence the linguistic behaviour of learners. Together these factors contribute to the situation alluded to in the introduction: **spoken** word recognition is given less attention than it deserves by applied linguistics.

Another contributing factor is that, during the period computational theories were developing, language teaching practice was developing communicative methods, favouring implicit learning over explicit teaching. While this resulted in generally improved methods of teaching, it offered little guidance for teachers in how to help learners with spoken word recognition and production when explicit instruction is needed, as it often is (Derwing and Munro, 2009).

The problem is that without a good understanding of the complexities of spoken word recognition, teachers may tend, in offering explicit instruction, to fall back on the everyday theory of spoken word recognition. For example, they may try to make things simpler for learners by focusing on phonemes as if they were discrete, invariant components of words. Indeed, this was the recommended approach for some time (Baker, 1981). More recently, a number of limitations of purely phoneme-based teaching have been identified (Celce-Murcia, Brinton, Goodwin, & Griner, 2010), and teachers are increasingly recommended to focus on suprasegmentals, such as syllables, stress patterns and intonation (Gilbert, 2005). While this has been shown (Hahn, 2004) to be more successful in some situations, it is not a panacea.

After all, it is one thing to advocate teaching suprasegmentals, and another to understand exactly how to do so effectively (Fraser, 2010).

## **Socio-cognitive theories of pronunciation**

It is here that a third class of theories is proving valuable. ‘Cognitive’ or ‘socio-cognitive’ theories reject the assumption that human cognition involves the kind of subconscious mechanistic processing featured by computational theories. They take a special interest in categorization as a basic, general cognitive process (Taylor, 2003), and give a central role to meaning, embodiment, and socio-cultural context in understanding how the human mind categorizes reality in all its forms.

Socio-cognitive theories have a lot to offer applied linguistics in general, and are being actively exploited in several areas of language teaching, including vocabulary (Tyler, Takada, Kim, & Marinova, 2005). Their take-up in relation to spoken word recognition has been slower, but involves exactly the same principles: categorization of the variable acoustic speech wave into spoken words relies, like all other kinds of categorization, on meaning and context (cf. the fascinating series of experiments summarized in Cutler, 2010).

While it is difficult to give central place to meaning and context in computational theories, this is no problem in socio-cognitive theories, which recognize that the ‘simplest’ form of speech perception is not decontextualized recognition of meaningless syllables (the ‘nonsense words’ of many computational experiments), but understanding of real, meaningful words and phrases in real, meaningful social contexts. This requires rejection of the assumption shared by both computational theories and the everyday theory: that phonemes are the basic units of speech (De Knop, Boers, & De Rycker, 2010). In fact, socio-cognitive theory argues that recognition of meaningless sublexical units, such as phonemes or syllables, far from preceding word recognition, actually depends upon prior recognition of larger, meaningful units, such as morphemes, words and phrases. While this idea may seem strange at first, it offers a simple, general account for the ‘word effects’ that cause such difficulties for computational theories, and sees variability not as a ‘problem’ but as an aid to perception (Dahan & Magnuson, 2006).

## **Pedagogical implications**

Accepting the view of socio-cognitive theory that words are more basic than phonemes suggests that, when learners have trouble recognizing and reproducing words, the most useful way to help may be, not via reference to decontextualized phonemes, but via assistance with picking out similar words from different speakers in different contexts. In other words, it suggests it might be easier to teach *red* and *led* than /r/ and /l/. Once words can be recognized, they can be compared and contrasted with other words to help achieve phonological awareness (the ability to recognize, produce and manipulate sublexical units such as phonemes and syllables).

This is similar to the order in which words come to be recognized in first language acquisition (though second language learners generally have the advantage of more mature cognitive skills). The problem is that it is easy for teachers, and even theorists, to forget how much effort is needed in first language acquisition before children can achieve even the most basic task of recognizing the same word spoken in different contexts by different people. Thus adults tend to believe that each word has 'a pronunciation' (as represented in a dictionary). In fact, of course, just as each word has multiple meanings, so each also has multiple pronunciations. Think, for a simple example, of the word *probably* as pronounced by men, women and children, as a citation form and in various sentence contexts, and so on.

As discussed above, speech is more like messy handwriting lacking gaps between words than like a printed text. Reading such handwriting cannot be achieved by looking for individual letters to put together into words, but requires first picking out words and then reconstructing the letters. The same is true of speech – meaning necessarily comes first. Indeed, it is interesting to note that even in reading a printed page, where the segmentation and invariance problems do not apply, meaning and context frequently override 'bottom-up' information from the printed letters (making it difficult, for example, to spot typographical errors in proof reading).

According to socio-cognitive theories, then, the first stage of acquisition is learning to recognize meaningful words in meaningful contexts. This involves the difficult task of

learning to categorize a range of very different word-pronunciations as 'the same word'. Only when this has been achieved, can this category of pronunciations be compared and contrasted with those of other words to start the process of recognizing sublexical categories of phonemes, syllables and so on. In other words, the first 'categorical perception' that has to be mastered, according to socio-cognitive theory, is not categorization of a range of different allophones as instances of a single phoneme, but categorization of a range of different pronunciations as instances of one particular word.

A further implication of socio-cognitive theories of spoken word recognition is the use they make of observations about the stages through which phonological awareness is achieved. It is clear that in first language acquisition, children learn to categorize larger sublexical units, such as units of meaning, rhythm and rhyme, long before they gain awareness of individual phonemes (Gillon, 2007). The same is true for many second language learners. This gives a rationale for the success of teaching word recognition through suprasegmentals, but also highlights the importance of basing the teaching of suprasegmentals on larger meaningful units of sound. For example, it suggests, as a way to help learners struggling with suprasegmentals, defining syllables not as sequences of phonemes, but as parts of words, defined by patterns of rhythm and stress. Then, through comparing and contrasting those patterns, syllables can be shown to have constituents, which can be defined as phonemes.

## **Conclusion**

The facts of spoken word recognition are usually presented via computational theories which are not just hard for teachers to understand but also of limited relevance to the practical tasks of applied linguistics. Presenting them instead via socio-cognitive theory makes them more accessible, and easier to incorporate into established practices of communicative language teaching. An apparent difficulty is that understanding the socio-cognitive theory of spoken word recognition means overcoming some highly entrenched but incorrect everyday assumptions, especially the view that recognition of spoken words requires the prior recognition of phonemes or other phonological units. Once this initial hurdle is overcome, however, socio-cognitive theory offers a conceptual framework that can make teaching pronunciation exceptionally effective, interesting and enjoyable.

## Suggested Readings

- Field, J. 2004. *Psycholinguistics: The Key Concepts*. London: Routledge.
- Fraser, H. (2011). Phonetics and phonology. In J. Simpson (Ed.), *Routledge handbook of applied linguistics*. New York: Routledge.
- Gaskell, M. G. (Ed.). (2007). *The Oxford handbook of psycholinguistics*. Oxford: Oxford University Press.
- Goodman, J., & Nusbaum, H. (Eds.). (1994). *The development of speech perception: The transition from speech sounds to spoken words*. Cambridge, Mass.: MIT Press.
- Vihman, M. (1996). *Phonological development*. Oxford: Basil Blackwell.

## Cross-references

SEE ALSO: Vocabulary Learning Strategies; Approaches to Second Language Vocabulary Teaching; Lexical Access in Visual Word Recognition in Second Language Processing; Lexical and Conceptual Representations in Second Language Acquisition; Vocabulary Acquisition in Second Language Acquisition

## References

- Baker, A. (1981). *Ship or sheep?* (2nd ed.). Cambridge: Cambridge University Press.
- Byrd, D., & Mintz, T. H. (2010). *Discovering speech, words and mind* Wiley-Blackwell Publishing.
- Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide*. Cambridge: Cambridge University Press.
- Cutler, A. (2010). Abstraction-based efficiency in the lexicon. *Journal of Laboratory Phonology*, 1, 301-318.
- Dahan, D., & Magnuson, J. S. (2006). Spoken word recognition *Handbook of psycholinguistics* (pp. 249–283). Elsevier.
- De Knop, S., Boers, F., & De Rycker, T. (Eds.). (2010). *Fostering language teaching efficiency through cognitive linguistics*. Berlin: Mouton de Gruyter.
- Derwing, T., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, 42(4), 476-490.
- Fowler, C. A. (2003). Speech production and perception. In A. Healy & R. Proctor (Eds.), *Handbook of psychology, Vol. 4: Experimental Psychology* (pp. 237-266). New York:

John Wiley and Sons.

- Fraser, H. (2006). Helping teachers help students with pronunciation. *Prospect: A journal of Australian TESOL*, 21(1), 80–94.
- Fraser, H. (2010). Teaching suprasegmentals like the stars. *Speak Out!*, 43(8–12).
- Gilbert, J. (2005). *Clear speech from the start: Basic pronunciation and listening comprehension in north american English* (3rd ed.). Cambridge: Cambridge University Press.
- Gillon, G. (2007). *Phonological awareness: From research to practice*. New York: Guilford Press.
- Hahn, L. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38(2), 201–223.
- Ladefoged, P. (2005). *Vowels and consonants: An introduction to the sounds of language* (2nd ed.). Oxford: Blackwell.
- Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, 72(4), 275–309.
- Perkell, J., & Klatt, D. (Eds.). (1986). *Invariance and variability in speech processes*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Pinker, S. (2003). *How the mind works*. London: Penguin.
- Port, R. (2007). The graphical basis of phones and phonemes. In O.-S. Bohn & M. J. Munro (Eds.), *Second language speech learning*. Amsterdam: John Benjamins.
- Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice vol.10* (pp. 223-335). New York/London: Academic Press.
- Taylor, J. R. (2003). *Linguistic Categorization: Prototypes in linguistic theory*. Oxford: Oxford University Press.
- Tyler, A., Takada, M., Kim, Y., & Marinova, D. (Eds.). (2005). *Language in use: Cognitive and usage-based approaches to language and language learning*. Washington DC: Georgetown University Press.
- Zuengler, J., & Miller, E. R. (2006). Cognitive and Sociocultural Perspectives: Two Parallel SLA Worlds? *TESOL Quarterly*, 40(1), 35-58.