Computational modelling of spoken-word recognition processes

Design choices and evaluation*

Odette Scharenborg and Lou Boves
Radboud University, Nijmegen

Computational modelling has proven to be a valuable approach in developing theories of spoken-word processing. In this paper, we focus on a particular class of theories in which it is assumed that the spoken-word recognition process consists of two consecutive stages, with an 'abstract' discrete symbolic representation at the interface between the stages. In evaluating computational models, it is important to bring in independent arguments for the cognitive plausibility of the algorithms that are selected to compute the processes in a theory. This paper discusses the relation between behavioural studies, theories, and computational models of spoken-word recognition. We explain how computational models can be assessed in terms of the goodness of fit with the behavioural data and the cognitive plausibility of the algorithms. An in-depth analysis of several models provides insights into how computational modelling has led to improved theories and to a better understanding of the human spoken-word recognition process.

Keywords: cognitive plausibility, computational modelling, computational model evaluation, modular architectures, theories of spoken-word recognition

1. Introduction

Due to the fact that the underlying neural processes are not directly accessible, theories of spoken-word processing (e.g., Johnson 1997; Gaskell and Marslen-Wilson 1997; Goldinger 1998; Klatt 1979, 1989; Luce et al. 2000; McClelland and Elman 1986; Norris 1994) tend to be quite abstract. Moreover, theories tend to focus on only particular aspects of the spoken-word recognition process, such as acoustic variability (e.g., Elman and McClelland 1986), the lexical segmentation problem (e.g., Norris et al. 1997), multiple activation of words (e.g., Allopenna et al. 1998; Gow and Gordon 1995), the lexical embedding problem (e.g., Davis et al. 2002;

Pragmatics & Cognition 18:1 (2010), **136–164**. DOI 10.1075/p&c.18.1.06sch
ISSN 0929–0907 / E-ISSN 1569–9943 © John Benjamins Publishing Company

Gow and Gordon 1995; Salverda et al. 2003; Salverda et al. 2007), and the flow of information (e.g., McClelland and Elman 1986; Norris et al. 2000). Building a computational model[1] forces theorists to be explicit about the details of the theory. Moreover, when computational models are used to combine and integrate partial theories to obtain a more comprehensive account of the processes, the representations at the interfaces between independently proposed modules must match. This increase in specificity and coverage is regarded as a significant contribution of computational modelling to theory building (Morse and Ziemke 2008; Newell 1973), also because without this there is the risk that theories lack detail (and thus can claim to predict anything; see Norris 2005). Therefore, it is unsurprising that computational modelling has proven to be a valuable line of research in the field of spoken-word processing in the past decades.[2]

Computational models of spoken-word processing may aim to test the adequacy of a particular theory with respect to behavioural data (Morse and Ziemke 2008) or to resolve debates on whether a theory really predicts what it claims to predict (Norris 2005). Computational models can also be used to investigate the influence of specific factors on spoken-word recognition, which are difficult or impossible to control in human listeners. An example of such a factor is the familiarity of a word (sequence), e.g., for some listeners the word sequence *computational modelling* might be well-known, while others do not know the meaning of the phrase. As one knows exactly what representations, processes, and parameters are present in a computational model, they are easier to control, and their role during word recognition can (more) easily be investigated. Finally, computational models can be used to make predictions about spoken-word recognition. Computational models can produce output phenomena that have not been described in existing literature; behavioural studies can then be designed to test these predictions of the computational model.[3]

The contribution of computational models in advancing theories of some process is all but straightforward. In this paper, we analyse the relation between models and theories in terms of Marr's (1982) three levels of complex information processing systems. The top level is what Marr refers to as 'abstract computational theory'. This level deals with *what* functions an information processing system must compute, and *why* those computations are required to achieve the goals of the system. The second is the algorithmic level, which addresses the question *how* the necessary functions are performed. In particular, this level specifies the input and output representations and the algorithms that must transform the input into the output. The third level is concerned with the hardware device in which the representations and algorithms are to be realised physically.

Computational models, as defined in the current paper, actually operate on the representations hypothesised in a theory and perform the computational processes hypothesised by a theory. This should result in a better understanding of the representations and the processes that form the heart of a theory.

Most spoken-word recognition processes hypothesised at the computational level (see Section 2.1) can be computed with several different algorithms; for instance, the 'best' interpretation of a speech signal in terms of a sequence of words can be computed using interactive-activation networks or using a kind of beam search (see Section 3.1.2). In turn, most algorithms can be implemented in several different ways. This means that when implementing a computational model, one is confronted with two types of design choices (or modelling assumptions). First, it is necessary to turn the functional specifications at Marr's computational level into specifications of representations and algorithms. More often than not there are multiple different representations that seem to be compatible with the theory. Once global specifications of representations and algorithms are in place, these must be converted into a complete technical specification. Also at this stage there may be several competing options. Because the behaviour of a model will depend on the design choices, the failure of a specific model need not invalidate the theory. On the other hand, the implementation of a computational model does not prove a theory to be correct either; instead, it shows that certain computational principles implemented in a certain way can account for the data. Computational models thus can only provide support for a theory. For that purpose, it is not sufficient that a computational model be able to reproduce behavioural data from one or two experiments. Almost invariably, individual data sets can be reproduced in many different ways, with many different algorithms. The task of the modeller is thus not only to simulate behavioural data, but also to explain why this particular choice of algorithm and its implementation might be (more) plausible (than others). Furthermore, a computational model becomes more convincing if it can reproduce multiple independent sets of behavioural data with the same parameter settings, and when it can accurately predict not-yet-observed behavioural data. Such a computational model provides a 'proof of principle' that the implementation of processes is sufficient for producing the observed data.

As there is hardly ever a straightforward mapping between theory and computational model, assessing the contribution of a particular computational model to the advancement of a certain theory is not trivial. The explanatory power of computational models of cognitive processes hinges on the credibility of the design choices. Therefore, it is important to bring in independent arguments for the cognitive plausibility of the representations and algorithms chosen to compute the processes in a theory as well as for the design choices that must be made in the implementations of a specific algorithm. Only if the design is cognitively plausible can the capability of some model to reproduce behavioural data be taken as support for the underlying theory.

In this paper, we show how computational modelling has contributed to improving a particular class of theories of the human spoken-word recognition process. We do this through an in-depth analysis of the evolution of the theory resulting from experiments with several computational models. To that end, this paper first discusses the relation between behavioural studies and global aspects of the theory (Section 2). We then analyse the relations between the theory and the design choices that need to be made in order to simulate behavioural data collected in experiments on spoken-word recognition (Section 3). Then we explain how different computational models of that theory can be assessed in terms of the goodness of fit with the behavioural data and the cognitive plausibility of the algorithms (Section 4). While doing so, we will show how the need to be explicit about the representation of behavioural data has spawned a new experimental paradigm that allows measuring the underlying processes more accurately. We conclude with identifying theoretical issues that have not yet been resolved by means of computational modelling and we suggest directions for further research (Section 5). It is hoped that our analysis of how computational modelling was able to advance one specific theory of one specific cognitive process, i.e., spoken-word recognition, can help researchers in other fields to analyse and understand the role of computational models.

2. From behavioural study to theory

It is impossible to directly investigate what happens in the brain of a person when recognising spoken words.[4] Therefore, psycholinguists rely instead on overt behaviour observed in experiments. The data obtained in these experiments usually consists of response times (RTs) or proportions of eye fixations, perhaps in combination with error rates. An example of behavioural data consisting of response times are the results of a series of cross-model lexical priming experiments that investigated which word meanings listeners accessed when confronted with lexically ambiguous sequences, which could either be interpreted as a single longer word or as two shorter words (Gow and Gordon 1995). Listeners heard sentences like "She placed her two lips on his cheek" that contained sequences of two short words — here *two lips*, which could be combined to form a single longer word, *tulips*. At the offset of the priming sequence *two lips*, a lexical decision probe appeared on a screen and the listeners had to determine whether the probe was a word or a nonword. Four types of probes were used: a word semantically related to the single longer word, here *flower*; a semantically unrelated word, here *grammar*; a nonword; and crucially, the single longer word itself, here *tulips*. They found faster lexical decision response times to the lexical decision probe (*tulips*) when subjects were presented with a sequence of short words (*two lips*) that comprises the same phoneme sequence as the decision probe. On the basis of the behavioural data in these experiments they concluded that listeners may simultaneously access words associated with several parses of ambiguous sequences, as well as other words than those intended by the speaker (ibid.: 352).

In general, faster response times are assumed to be associated with stronger lexical hypotheses, and thus with stronger *word activations*, than for competing lexical interpretations. Conversely, slower response times indicate an increase in the difficulty of recognising the word(-initial cohort), which in turn corresponds to a smaller difference between the lexical strength of the word(-initial) cohort and its competitors. Likewise, higher error rates indicate an increased difficulty of the recognition of the word(-initial cohort), and thus also corresponds to smaller differences of lexical strength, and *vice versa*.

Another example of behavioural data obtained in spoken-word understanding experiments are proportions of eye fixations obtained using the visual world eye-tracking paradigm. In this paradigm, subjects are asked to follow instructions to look at, pick up, or move one of a set objects presented in a well-defined visual workspace (Tanenhaus and Spivey-Knowlton 1996), usually on a computer screen, instead of making explicit meta-linguistic decisions about the speech stimuli. Using the visual world paradigm makes it possible to monitor the speech comprehension process over time. It has been demonstrated (e.g., Allopenna et al. 1998; Dahan et al. 2001a; Dahan et al. 2001b; Tanenhaus 2000) that the timing and pattern of the eye fixations to possible referents in the visual workspace provide a sensitive measure of the time course of lexical activation in continuous speech, and that a simple 'linking hypothesis' provides a good mapping of pattern and timing of eye fixations onto the underlying lexical activation. In general, the word corresponding to the picture that attracts most eye fixations is assumed to be the strongest lexical hypothesis.

Using the visual world paradigm it could be shown that listeners can make the distinction between the two interpretations of an ambiguous sequence in the case of initially embedded words, such as 'ham' in 'hamster', even before the acoustic end of the first syllable *ham* (Davis et al. 2002; Salverda et al. 2003, 2007). Salverda et al. (2003) showed that a picture of the embedded word attracted more eye fixations in an eye-tracking experiment when the ambiguous sequence 'ham' was cross-spliced from a monosyllabic word than when it was cross-spliced from the first syllable of another recording of the longer word in which it was embedded. They concluded that a phoneme sequence with a longer duration tends to be interpreted as a monosyllabic word more often than a shorter one; it follows that the word activation of the monosyllabic word 'ham' is higher than the word activation of the polysyllabic word 'hamster'.

2.1 Theory of spoken-word processing

Speech is highly variable. Two acoustic realisations of the same word are never identical, not even when they are spoken by the same person. This variability in the speech signal is due to factors such as speaker characteristics (e.g., gender, age, emotional state), speaking style and rate, phonetic context (e.g., sounds appearing at different places within a syllable or word or in different phonemic contexts can be pronounced differently), and prosody (Benzeghiba et al. 2007). Humans are thus faced with the task of mapping a highly variable speech signal onto some kind of invariant meaning, most likely by virtue of some kind of invariant lexical representations.

There are two largely antagonistic theories of spoken-word processing: the 'episodic' and the 'abstract' theory of spoken-word processing. The episodic theory assumes that each lexical unit is associated with a possibly large number of stored acoustic representations (e.g., Johnson 1997; Goldinger 1998; Klatt 1979, 1989). There are, however, very few computational models based on this theory (Johnson 1997; and models based on or derived from MINERVA2, a computational multiple-trace memory model (Hintzman 1986), e.g., Maier and Moore 2005, 2007; Wade, et al. 2002). In this paper, we focus on the 'abstract' theory, because there is a much larger number of computational models and papers which makes it possible to analyse the role of computational modelling in the development of this theory.

The abstract theory of spoken-word recognition (e.g., Gaskell and Marslen-Wilson 1997; Luce et al. 2000; McClelland and Elman 1986; Norris 1994) assumes that the speech recognition process consists of two stages. This two-step process

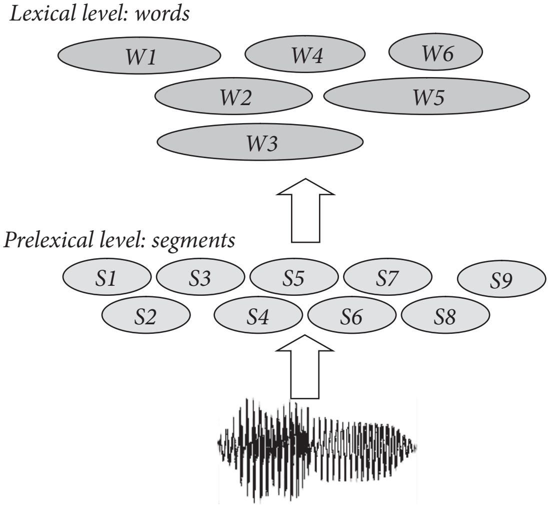

Figure 1. Graphical illustration of the abstract theory of spoken-word recognition.

is illustrated in Figure 1. First, listeners map the incoming acoustic signal onto a limited set of abstract so-called prelexical representations at the prelexical level. The details of the prelexical representations are so far unknown, and so is the process that converts continuous audio signals into discrete prelexical units. At the lexical level, all lexical representations are stored in the form of sequences of prelexical units. The core problem solved at the lexical level is the matching of the prelexical representations with the lexical representations in the lexicon. There is now considerable evidence that multiple candidate words that match the input to a sufficient degree are 'activated' simultaneously during spoken-word recognition and are processed in the lexical phase (e.g., Allopenna et al. 1998; Gow and Gordon 1995). Therefore, the abstract theory maintains that there must be processes that resolve the matching of the prelexical representations with the lexical representations by using several kinds of information that can be represented by the prelexical units. We refer to this as the 'disambiguation process'. Since word hypotheses can start and end at any time, word hypotheses that overlap in time 'compete' with each other. An example: take the phonemic representation of the phrase 'ship inquiry' /ʃɪpɪŋkwaɪəri/. This phoneme sequence contains many embedded words, such as 'ink' and 'choir' in 'inquiry', but also words that straddle the word boundary such as 'shipping' and 'pink'. While the speech signal unfolds over time, all these possible words will become activated and will be evaluated, and the best matching word candidate (sequence) is selected (we will come back to this in Section 3.1.2). This activation and selection is influenced by a large set of partially constraining cues (including acoustic, prosodic, statistical, and lexical cues). The result is a sequence of non-overlapping words, usually identical to the sequence of words actually produced by the speaker. The spoken-word recognition process thus resolves the temporary ambiguity of overlapping words, and results in the optimal segmentation of the input.

The account of the abstract theory given in the preceding paragraph only describes the representations and processes involved in spoken-word recognition that are present (in some way or another) in all versions of the theory. A controversial issue is the flow of information, which is only bottom-up in the scheme depicted in Figure 1. We return to this issue in Section 5.

3. From theory to computational model

Computational models should be able to reproduce observed behavioural data by simulating the processes stipulated by the theories. In addition, the computational model should be able to make accurate predictions of aspects of the speech recognition process that do not directly follow from the observed data or the literature

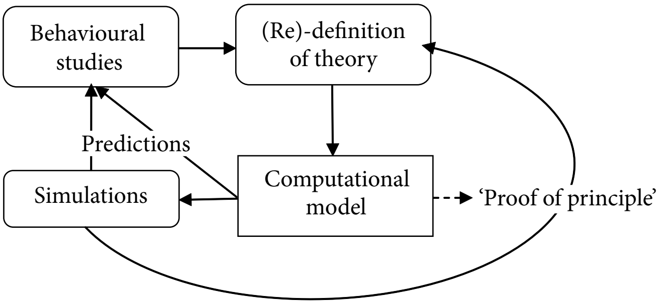

Figure 2. The relationship between behavioural studies, psycholinguistic theories, and computational modelling.

(Norris 2005; Tuller 2003). If the model indeed is able to reproduce the observed as well as predict not-yet-observed behavioural data, the computational model provides a proof of principle of the underlying theory. The predictions the model makes can be used for the design of new behavioural studies. The data obtained with these new behavioural studies can then be used to test these predictions. The outcome of this test may lead to a redefinition or adaptation of the original theory on which the computational model was based. Furthermore, the results of the simulations may warrant a redefinition of the theory. Figure 2 illustrates the relationship between behavioural studies, psycholinguistic theories, and computational modelling (see also Norris 2005; Tuller 2003).

When implementing a computational model, two different types of design choices must be made. The first type of design choice is related to the specifications of the algorithm and representation that are to perform the functions defined at the computational theory level. This is because the theory (usually) does not define the details of these representations and processes. Secondly, most representations and algorithms can be implemented in several different ways, requiring design choices regarding the actual implementation. The fact that quite different computational models seem to be compatible with the theory raises questions about the extent to which a model can indeed provide support for the theory. If a model fails to simulate behavioural data, theorists can blame this to the choice of the representations or the algorithm or to the way these were implemented. If the model does simulate the data, proponents of another theory may claim that some aspects of the model have been tweaked in such a manner that they actually violate aspects of the underlying theory, or that the actual algorithm makes assumptions that clearly violate computational constraints imposed by the human wetware.[5] It is therefore crucial that the theoretical and modelling assumptions are pulled apart in order to evaluate a model (and/or a theory). In the following, we will illustrate the design choices that need to be made to get a working computational model of spoken-word recognition based on the abstract theory of spoken-word recognition.

3.1 Design choices

In our explanation of the design choices that need to be made when building a computational model of the abstract theory of spoken-word recognition, we will go into more detail with respect to the theory as outlined in Section 2.1. The first part of this subsection deals with the design choices related to the prelexical level; the second, with the design choices related the lexical level.

3.1.1 *Prelexical level*

The first issue that needs to be addressed is the input to the model. One might wonder why this is an issue, as it might seem obvious that the input to the prelexical level should be the acoustic signal. In truth, however, virtually all computational models based on the abstract theory simply claim that some kind of prelexical units can be computed, and in reality miss an explicit prelexical level. So, instead of dealing with the complexity and variability of real speech at the prelexical level, most influential computational models (Distributed Cohort Model (DCM): Gaskell and Marslen-Wilson 1997; ARTphone: Grossberg et al. 1997; PARSYN: Luce et al. 2000; TRACE: McClelland and Elman 1986; Shortlist: Norris 1994) avoid the specification of the computational processes at the prelexical level that are needed for this function, and instead use an artificial, often hand-crafted, idealised discrete (prelexical) representation of the acoustic signal as input to the *lexical* level of the model.

Existing versions of the abstract theory only hold that there is a limited number of discrete abstract prelexical units. The question of what form the prelexical representations take and how these can be derived from speech signals is not decided (McQueen 2005). In the absence of an answer to this issue, different computational models make different design choices with respect to the form of the prelexical representations. Shortlist, ARTphone, and SpeM (Scharenborg et al. 2005), for instance, take phoneme symbols, DCM and TRACE use multi-dimensional features, which in the case of TRACE internally activate phonemes, Fine-Tracker (Scharenborg 2008, 2009) uses articulatory-acoustic features, and PARSYN (Luce et al. 2000) uses context-sensitive allophones. Shortlist B (Norris and McQueen 2008), the successor of Shortlist, uses a prelexical representation consisting of sequences of multiple phoneme probabilities over three time slices per phoneme, which are derived from the performance of listeners in a large-scale gating study. However, Shortlist B does not specify the process that listeners use to convert speech to sequences of phoneme probabilities.

To our knowledge, there are only three computational models of the abstract theory of spoken-word recognition that take the actual speech signal as input. By doing so, these models extend the theory of spoken-word recognition by specifying a computational process that converts the acoustic signal to the prelexical representations. An important side-effect of using the acoustic signal as input is that these computational models can perform simulations by being fed with exactly the same speech stimuli used in a behavioural experiment. Two of these models are SpeM (Scharenborg et al. 2005) and its successor Fine-Tracker (Scharenborg 2008, 2009). SpeM and Fine-Tracker use techniques from the field of automatic speech recognition to achieve the conversion of speech signals into prelexical units. In the case of SpeM, a phone recognition system based on hidden Markov Models is used to convert the acoustic signal into phone graphs. Like Shortlist, SpeM thus uses prelexical units that take the form of phonemes. Fine-Tracker uses a set of multi-layer perceptrons that convert the acoustic signal into vectors consisting of articulatory-acoustic features; these articulatory-acoustic features are the prelexical representations. The third model is TRACE I (Elman and McClelland 1986), which converts digitised speech to a set of real-valued phonetic features, reminiscent of the features which are converted into phoneme symbols in the model of the lexical stage TRACE II (referred to as TRACE in this paper; McClelland and Elman 1986).

To summarise, most existing computational models based on the abstract theory miss an explicit prelexical level. Furthermore, there is no general agreement with respect to the definition of the prelexical units. The status of a specific definition of prelexical unit is different between models that essentially miss the prelexical stage — and therefore are free to handcraft the prelexical units — and models that compute prelexical units from actual speech signals. Even if the latter type of models still need to choose a specific form of prelexical representation, different types can be compared on the basis of the accuracy with which they can be derived from real speech input. Additional design choices need to be made in the case of an explicit prelexical level, namely with respect to the processes needed to convert the acoustic signal into the prelexical representations. The specification of these processes is again dependent on the definition of the prelexical representations.

3.1.2 *Lexical level*

The single most important task to be accomplished at the lexical level is the disambiguation of the representations that enter from the prelexical level. Although one might think that ambiguity resolution is not necessary when recognising isolated words, temporary ambiguity is always present. For example, the word 'inquiry' will also activate embedded words such as 'ink' and 'choir' (see Section 2) and word sequences such as 'ink wire'. In the case of the recognition of a word sequence the number of different possible parses can be extremely large.

The two most influential computational models of the lexical stage of spoken-word recognition, TRACE and Shortlist, differ in their choice for the representation at the input (multi-dimensional features, which are converted into phonemes, in TRACE, only phonemes in Shortlist) but they seem to agree on the representation of the mental lexicon (a store of orthographic strings with associated phonemic transcriptions). Moreover, both chose interactive-activation networks (also referred to as connectionist models) for computing the (string of) word(s) that matches the input best. These networks suggest the concept of 'word activation', because lexical representations (words) are represented by a single node which is assigned an activation value. There thus is a link between the theoretical concept of word activation and the activation of words in the networks. Both models assume that the activation of a node increases through input that matches the word and decreases through inhibition from other words. This suggests the concept of 'competition'. Furthermore, both models assume that the activation of words occurs in parallel. In older versions of the theory (e.g., McClelland and Elman 1986; Norris 1994) the concepts of 'activation' and 'competition/inhibition', analogous to what is happening in the interactive-activation networks, were considered as essential parts of the theory.

Despite the fact that both TRACE and Shortlist opted for interactive-activation networks, the actual disambiguation proceeds differently. In Shortlist, the prelexical representations are first compared to the lexical representations using an exhaustive search through the lexicon. The degree of match between the prelexical and lexical representations is calculated as follows: for each matching phoneme the candidate word scores +1, and -3 for each mismatching phoneme. A shortlist of the most promising candidate words is then entered into an interactive-activation network. The word nodes in the network are activated in proportion to their match to the input as determined by the match/mismatch score, and words that derive their evidence from the same input phonemes are connected via inhibitory links. TRACE's input representation has featural nodes as well as phoneme nodes, and there is competition (inhibition) between phoneme nodes as well as word nodes. Furthermore, word activation in TRACE is not decreased by the presence of mismatching phonemes. Finally, in TRACE all words in (a necessarily small) lexicon are wired into the interactive-activation network. These differences between TRACE and Shortlist, however, are design choices not imposed by the theory. They thus do not play a role when comparing the theoretical implications of Shortlist and TRACE. What remains, of course, are the differences between the two models with respect to their assumptions about the input representations and the difference in flow of information.

In SpeM and Shortlist B the disambiguation process is not implemented using an interactive-activation network; instead, in these models, disambiguation is regarded as a search problem (Norris and McQueen 2008; Scharenborg et al. 2005). Scharenborg et al. (2005) showed that, given a certain (probabilistic) representation at the prelexical level, disambiguation as implemented in an interactive-activation network and as a search are similar at the computational theory level: a clever beam search process returns the word (sequence) that is most likely given the speech signal. During the processing of incoming speech, multiple paths corresponding to (sequences of) candidate words are considered simultaneously, and each candidate word (or to be more precise, each path) is assigned a likelihood score that indicates the match between the word (sequence) and the input. The path with the best score wins. Since it is not possible to compute $P(W|X)$, the likelihood of a word sequence W given the speech signal X directly, Bayes' Rule is used and the problem is transformed in computing the sequence of words W that maximises the likelihood of observing the acoustic signal X (i.e., $P(X|W)$ multiplied by the prior probability of the word sequence ($P(W)$). During the search, pruning techniques remove the most implausible paths, in order to keep the number of paths through the search space manageable. As a result, only the most plausible words are considered in the search. This search process can use arbitrarily large lexicons without the need to create a shortlist before the disambiguation starts.

The difference between Shortlist and Shortlist B illustrates how computational modelling suggested an adaptation of the theory of spoken-word recognition. When TRACE and Shortlist were developed, the meaning of 'activation' was directly related to the representation of words as a single node in a connectionist model, which was assigned an activation value. However, in Shortlist B (as well as in SpeM), the meaning of 'activation' shifted to a score representing the degree of match with the input. Likewise, originally the 'competition process' was considered to be a process in which candidate words competed with each other like athletes compete in a wrestling match, i.e., candidate words were trying to actively suppress or inhibit the other candidate words. However, in SpeM and Shortlist B, the 'competition process' is regarded as a process in which candidate words compete with each other like track and field athletes, i.e., candidate words have their own match with the input, they run their own race, and do not actively suppress or inhibit other candidate words.

To summarise, the design choices that need to be made when building the lexical level of a computational model of the abstract theory of spoken-word recognition focus on the implementation of the disambiguation process. Originally, the lexical stage was modelled as an interactive-activation network, while subsequent models regarded the disambiguation process as a search problem. Not surprisingly, there is a strong interaction between the assumptions about the prelexical representation of the input of the lexical stage and the computations that are needed to match the input to a sequence of phonemic representations of words in the lexicon. It appears that recent advances in computational modelling have affected the theory of spoken-word recognition. At the very least, modelling experiments have clarified the status and importance of the concepts of 'activation' and 'competition' that were central in older versions of the theory.

4. Computational model evaluation

The evaluation of the contributions that a computational model can make to the advancement of a cognitive theory involves at least two aspects: assessing the model's fit to the empirical data (Section 4.1), and the cognitive plausibility of its design choices (Section 4.2). Ideally, a model should also predict not-yet-observed behaviour. If a computational model fails to accurately simulate empirical data, further analyses should determine whether the problem is in the algorithms used to compute the functions stipulated by the theory or whether the theory needs revisiting. If the assessment is positive, the implications of the design choices should be investigated (Section 4.3) with respect to the underlying theory.

In cognitive science, computational models are not a goal per se, but rather a means to improve theories and therefore to advance our understanding of as yet unobservable cognitive processes. Therefore, it does not make much sense to ask which computational model is 'best' (Myung 2001). This is due to multiple reasons. First of all, the results from behavioural studies are never clear-cut; different subjects and stimuli may give different results (see, e.g., Scharenborg 2009). And perhaps even more importantly, the link between the behavioural data and the underlying processes may not always be direct. In addition, there may not even be a generally agreed formal procedure for establishing the degree to which the output of the model fits the behavioural data (cf. Section 4.1). Secondly, computational models are generally tested on only a limited number of data sets (e.g., data from two or three experiments; Pitt, Myung, Montenegro, and Pooley 2008). As a consequence, it is difficult to know whether the results are due to essential features of the computational models or rather to some features of the data sets. In addition, almost invariably individual data sets can be reproduced in many ways due to the flexibility of computational models, governed by the model's parameter set. The more parameters a computational model has, the more flexible the model is, and the better its fit with the behavioural data. However, there is always the danger of over-fitting the data. Pitt et al. (2008) present an overview of methods to compare computational models and investigating the models' flexibility. They present a framework called 'parameter space partitioning', which was originally developed for evaluating statistical computational models, but they showed that it can also be applied for the evaluation of procedural models, such as the models in focus in this paper.

4.1 The fit to the empirical data

In general, a computational model is evaluated by assessing the fit of its output with empirical/behavioural data. However, in the case of computational models of spoken-word recognition there is no one-to-one mapping between the data from behavioural studies and the output given by computational models. Therefore, either the behavioural data must be 'interpreted' in terms of the output of the model, or — perhaps more likely — the other way around: we must interpret the model output in terms of behavioural measures.

Computational models such as TRACE, Shortlist, and SpeM do not output response times or eye fixation durations, nor are error rates usually being calculated. Cutler and Robinson (1992: 190) describe a way of extracting response times out of an automatic speech recognition system in order to directly compare it to the human response times, and Scharenborg et al. (2005: 907) calculated error rates for their computational model and compared those with the error rates of listeners.

Models such as TRACE, Shortlist, ARTphone, PARSYN, and DCM were designed to compute word activations as a function of 'time'. Word activation is a continuously changing measure that indicates how strong a word hypothesis is at a specific moment in time. However, in most experimental paradigms we can only observe instantaneous behaviours (e.g., button presses) that correspond with the result of the disambiguation and decision process of a specific subject.

Shortlist B, SpeM, and Fine-Tracker were also designed to compute the equivalent of the time course of word activations. SpeM provides an estimate of the probability that a listener would identify that word given that input — an estimate which changes over time as the speech input unfolds. Word activation as used in SpeM provides a joint measure of the goodness of fit of the word to a particular stretch of a given input and the goodness of fit of the path on which that word occurs to the complete input (more specifically, the score of the best path associated with that word). Shortlist B is based on the theoretical assumptions that listeners use a "near optimal strategy" (Norris and McQueen 2008: 358) for the recognition of spoken words, in other words it is suggested that listeners behave as optimal Bayesian decision makers. Because of this perspective on spoken-word recognition the output of Shortlist B consists of time varying posterior probabilities (which by definition cover the interval [0, 1]), rather than of word activations (which cannot be expressed in a fixed scale). Fine-Tracker computes time varying word scores, which are based on the goodness of fit between articulatory-acoustic features extracted from the acoustic signal and corresponding features in the lexical representations.

However, the response times provided by the experimental paradigms used in the past do not provide the detailed patterns of the activation functions of the eventual winner, and even less so from the competitors. Consequently, it is not possible to compute a quantitative measure of the degree of fit between the output of the models and the behavioural data, leaving it to the modeller to decide if a certain degree of match is sufficient or not.

In this respect, the visual world eye-tracking paradigm and the 'linking hypothesis' are a substantial improvement (e.g., Allopenna et al. 1998; Tanenhaus et al. 2000). It is assumed that the activation of a word is reflected by the proportion of the time that the eyes are focused on the corresponding picture. As such, the visual world eye-tracking paradigm has provided a completely new type of behavioural data that must be accounted for by theories of spoken-word processing. Moreover, the new data are more similar to the output of existing computational models, thus allowing for a more direct and formal measure of the degree of fit (Dahan et al. 2001a, 2001b). It goes without saying that the new type of behavioural data will have an impact on the development of future computational models (e.g., Allopenna et al. 1998; Dahan et al. 2001a, 2001b).

In any behavioural experiment there are going to be differences (between items and between subjects) that are of interest to the experimenter and those that are not. The interesting differences are the ones that are related to the underlying theory; the 'uninteresting' differences are considered noise and it is hoped that these are removed by averaging over subjects and stimuli. Therefore, behavioural results are usually reported as averages over multiple stimuli and multiple subjects. Most computational models, therefore, are 'macroscopic' models, i.e., the overall response of the model is fitted to the averaged behavioural data.

Several sets of average data that can be used for building and testing macroscopic models are available in the literature. 'Microscopic' computational models, on the other hand, are models whose response pattern has a high probability of being part of the set of response patterns of individual participants in a behavioural study. One way of building a microscopic model is to change the parameter settings of a macroscopic model in order to model individual and item-specific differences. In order to be able to evaluate a microscopic model, the raw behavioural data is needed, i.e., the data of each stimulus and subject individually. Furthermore, the availability of raw behavioural data enables one to make useful in-depth comparisons of the simulation results and the behavioural data, including analyses at the level of error rates and the responses to individual stimuli. We would therefore strongly suggest that raw data are published or at least made available on request.

4.2 Cognitive plausibility

In order to evaluate a computational model in terms of its cognitive plausibility, the algorithm, the input and output representations, and relevant implementation details need to be assessed. For instance, as explained above, models such as TRACE, Shortlist, PARSYN, and DCM use a discrete representation of the acoustic signal as input to the lexical stage of spoken-word recognition. It is tacitly assumed that there is some (plausible) process that converts the acoustic signal into the discrete prelexical representation that the models expect. In this context it may be telling that to the best of our knowledge TRACE I (the mapping from speech to features) and TRACE (the lexical disambiguation) have not been used in tandem to model the spoken-word recognition process using the acoustic speech signal as input. (Note that the lack of a (cognitively plausible) process that can convert speech into prelexical units not only raises questions about the validity of the theory,[6] but also complicates attempts to compare different versions of the theory by means of computational modelling experiments. It is difficult to compare lexical level models that rely on handcrafted input, if only because the details of a specific set of input representations might have substantial effects on the output of the simulations).

At the present state of our understanding of speech processing it is highly unlikely that it will ever be possible to convert neither an arbitrary speech signal, nor even a carefully articulated read sentence into a feature or phoneme representation with an accuracy that comes close to the input representations needed by TRACE or Shortlist. Automatic phoneme recognition experiments using the well-known TIMIT corpus (Garofolo 1988) invariably show that only about 75% of the phonemes are recognised, irrespective of the recognition algorithm (e.g., Schwarz et al. 2004). Attempts to extract phonetic features from a carefully articulated corpus such as TIMIT do not fare any better (e.g., Scharenborg et al. 2007; Schuppler et al. 2009). Yet, the well-known computational models of spoken-word processing require (close to) perfect accuracy of the input representation. For instance, Scharenborg et al. (2003) created an automatic phone recogniser to convert the acoustic signal into a string of phones that was subsequently fed into Shortlist. Only 76.5% of the words were present in the shortlist generated by Shortlist (which subsequently entered the competition phase), and only 54.1% of the words were correctly recognised by Shortlist. Scharenborg et al. (2003: 3034) suggested that the limitations of the joint model of the phone recogniser and Shortlist could be overcome by avoiding hard phone decisions at the output side of the phone recogniser and by using a match between the input and the internal lexicon that can cope with deviations from canonical phonemic representations. This led to the development of SpeM (Scharenborg et al. 2005). It is evident that a model that can only operate successfully with an input representation that cannot be computed in a plausible manner can only claim limited cognitive plausibility.

In addition to issues with the cognitive plausibility of the input, there are also design choices related to the implementation of the processing algorithms that raise questions about the cognitive plausibility. TRACE, for example, needs to duplicate the entire lexical network many times. As a consequence, TRACE could only deal with small lexicons and it is not evident that this limitation can be lifted by a different implementation of the algorithm. One of the aims for the development of Shortlist was to tackle this limitation. In Shortlist, competition takes place in a small lexical network that only considers those word candidates that match the input best. This set-up resulted in the possibility of using a more realistically-sized lexicon. However, the cognitive plausibility of the actual implementation in Shortlist is questionable: the competition process can only start after the end of (a stretch of) the input utterance, because the creation of the shortlist and the competition are implemented as sequential processes.

It is worthwhile pointing out that "cognitive plausibility" is not a precisely defined concept. This explains why models such as TRACE and Shortlist have made major contributions to our understanding of human spoken-word recognition, despite the fact that they make unrealistic requirements for their input. These models can account for many behavioural data, and by so doing they have helped to obtain a better understanding of the processes that are needed to generate the behavioural responses observed in a range of experiments.

4.3 Implications of the computational model for the underlying theory

If the assessment of the computational model is positive in terms of goodness of fit and cognitive plausibility, then the implications of the design choices, the theoretical assumptions, and the parameter settings for the theory underlying the model should be investigated. The important question to ask is: what is it about the model that makes it able to be successful? Is it because of the computational functions specified in the theory or due to coincidental side-effects of the design choices? Answers to these questions will give us insights into the way human spoken-word recognition actually works and provide insights into how and under which conditions simulation experiments can lead to improved theories and better understanding of the human spoken-word recognition process.

An example of this enterprise is related to a study by Scharenborg (2009) on the role of durational information in spoken-word recognition: there is now considerable evidence that durational cues in the acoustic speech signal help resolve temporarily ambiguous speech input due to lexical embedding, such as 'ham' in 'hamster' (e.g., Davis et al. 2002; Salverda et al. 2003). Scharenborg (2009) presented two simulation experiments with Fine-Tracker, using the acoustic stimuli from two behavioural studies as input. The simulations showed that the model, like humans, takes benefit from durational cues during word recognition, and uses it to disambiguate the incoming speech signal. Durational information in Fine-Tracker is stored in its lexicon, which is a design choice. An analysis of the question what it is about the model that makes it successful shows that there are three aspects that are crucial to the workings of Fine-Tracker: 1) the differentiation in the lexical representations between monosyllabic words and phonemically identical syllables which are part of polysyllabic words; 2) the ability to represent durational information at the prelexical level; 3) and to use this durational information at the lexical level to distinguish between the monosyllabic and the polysyllabic word. These aspects of the model are in accordance with theoretical assumptions about the role of durational information in spoken-word recognition, and not due to, for instance, side-effects of the chosen implementation.

5. Open issues in theories of spoken-word processing

In this section, we explore three open issues in theories of spoken-word processing where computational models are an essential tool for making progress. These issues are the debate about the role of feedback from the lexical to the prelexical level that was already alluded to in relation to Figure 1; the way in which abstract theories can deal with adaptation to strong foreign or regional accents; and the relation between abstract and episodic theories.

5.1 The role of feedback in spoken-word recognition

Perhaps the best way to introduce this issue is by means of a quote from the personal research website of Dennis Norris:

> One of the most hotly debated issues in perception is whether the early stages of perceptual analysis are modular or not. In the context of spoken word recognition, the critical question is whether lexical information feeds back down to influence earlier stages of phonological or phonetic analysis. [7]

It should be clear that the abstract theory as outlined in Figure 1 is fundamentally modular: it is assumed that the prelexical process is independent of the lexical process, with the prelexical representations as the interface between the two. The question that has intrigued researchers for almost two decades is whether the representation at the interface can or cannot be affected by knowledge from the lexical level. The issue at stake is not how subjects recognise spoken words, but rather what factors affect phoneme recognition. One camp maintains that such feedback is not logically necessary, because results from behavioural experiments can be satisfactorily accounted for under the assumption that the information flow is strictly bottom-up, meaning that feedback cannot be demonstrated (e.g., Norris 1994; Norris et al. 2000). The competing camp (e.g., McClelland and Elman 1986; Samuel and Pitt 2003) maintains that experiments in which subjects must make phoneme decisions on speech that is manipulated to create ambiguous sounds do show that the feedback phenomenon can be attested, implying that the spoken-word recognition process is not strictly modular. Perhaps not surprisingly, the debate centres to a large extent around the question whether behavioural effects — if at all statistically significant — are 'real' or rather the result of biases in the experiments (McQueen, Jesse, and Norris 2009).

One would expect that computational modelling can shed light on the issue of (strict) modularity. In investigating this issue it must be taken into account that all models that subscribe to the architecture in Figure 1 assume that there are indeed two modules that operate in tandem: in all models, the prelexical processing (if implemented) precedes the lexical level. Therefore, there is only one way in which a model that implements the modular structure can lead to the conclusion that the spoken-word recognition process is not strictly modular, namely by showing that the modular architecture is not able to simulate all relevant behavioural data. In the past, simulation experiments have been conducted with feed-forward only models (e.g., Shortlist, Merge (Norris et al. 2000), SpeM, and Fine-Tracker) as well as with models that have on-line feedback (among others PARSYN and TRACE). Both types of computational models have been able to simulate relevant behavioural data; however, all models leave some behavioural results unexplained. Therefore, it might seem that computational modelling has not been able to resolve the debate.

However, the question of whether or not feedback plays a role can be approached from a somewhat different computational modelling perspective. In Section 3.1.1 we have pointed out that most models have not specified the process that should convert speech input to the prelexical representation. It was also mentioned that different models make different assumptions about the details of the prelexical representations. In the summary of Section 3 we emphasised the (seemingly obvious) fact that there is a strong interaction between the assumptions about the prelexical representations and the processes that compute these representations from the speech signal and that map prelexical to lexical representations. And in Section 4.2 we argued that it is extremely unlikely that a computational process (natural or artificial) can exist that converts speech into an unambiguous discrete prelexical representation that is good enough to allow the older models to successfully simulate spoken-word recognition. We also pointed out that the newer models (SpeM, Fine-Tracker, and Shortlist B) assume that the prelexical representations are essentially probabilistic. Shortlist B does this by design, while SpeM and Fine-Tracker do this by necessity, because they do include a full-fledged prelexical processing stage. And although SpeM and Fine-Tracker are implemented in such a way that they can export prelexical representations for inspection by the experimenter, these models share basic properties with the integrated search that is the arguably the single most important characteristic of state-of-the-art automatic speech recognition systems. The search implemented in Shortlist B shares the same property. But if it is so that prelexical representations are fundamentally ambiguous and probabilistic, the issue of phoneme recognition must be fundamentally re-thought. All results from experiments with speech signal processing strongly suggests that accurate purely bottom-up phoneme recognition is not possible. And from automatic speech recognition we know that phoneme recognition is not necessary. Therefore, it would seem that our re-analysis of computational modelling experiments strongly suggests that phoneme recognition may not be possible without invoking the notion of a lexicon in which words are represented as phoneme sequences.

5.2 Coping with unfamiliar pronunciations

A phenomenon that has attracted a lot of attention recently in the spoken-word recognition community is the ease with which listeners in communicative situations and subjects in experiments can adapt to unfamiliar (foreign or regional) pronunciation variants (e.g., Bradlow and Bent 2008; Kraljic and Samuel 2007; McQueen et al. 2006). Part of the newly learned pronunciation features generalises across speakers and words, while another part seems to be speaker-dependent. These findings raise the question how these adaptations can be made compatible with a theory that attributes an essential role to abstract invariant phonemes at the interface between the prelexical and lexical stages of spoken-word processing.

Al least since the seminal paper by Peterson and Barney (1952) — and most probably already long before that time — it has been known that speech sounds display a tremendous degree of variation and a substantial overlap in their acoustic characteristics. Knowledge about the human auditory system does not suggest procedures that could be invoked for mapping the variable acoustic features onto distinct abstract vowel points or volumes in the acoustic space. Technical procedures for normalising vowels have not been particularly effective (Adank et al. 2004). To be able to deal with between-speaker variation a theory that claims that prelexical representations can only consist of a limited number of 'abstract' units must assume that prelexical processing in some way or another involves statistical distributions (perhaps of relative distances between sounds, rather than of features of individual sounds) that are learned during language acquisition. And there are no compelling reasons to assume that learning will stop completely at some age.

While phoneticians prefer to think of vowels in terms of two or three formant frequencies (Harrington 2010), it is much more likely that speech sounds are represented in the auditory system in terms of energies in some 20–30 band pass filters, combined with the speed and acceleration of energy changes over time (Jurafsky and Martin 2009). Even if the redundancy due to correlations between adjacent frequency bands is removed, we are still facing a high-dimensional acoustic space. In such a high-dimensional space virtually all distributions tend to be underspecified. As a result, almost every observation in that space sits in the tail of the distribution in some of the dimensions. It is quite conceivable that idiosyncratic sounds produced by a foreigner or a speaker with a strong regional accent can be integrated in the tails of a sparsely specified high-dimensional distribution. This would provide a means for representing new knowledge, without interfering with previously learned representations which must be retained for recognising 'standard' speakers. Although it remains to be proven in actual simulation experiments, it is reasonable to assume that some version of Fine-Tracker should be able to combine on-line adaptation to idiosyncratic speech with unchanged performance for 'standard' speech and without interfering with the assumption that phonemes are represented as abstract invariable units.

While the idea of sparsely specified multidimensional distributions is compatible with conventional Gaussian models, we can also think in terms of nonparametric distributions or of set membership of observations. Recent computational modelling experiments in the ACORNS project have shown that general purpose structure discovery techniques such as Non-Negative Matrix Factorisation (NMF) (Lee and Seung 2001; van Segbroeck and van Hamme 2009) and DP-Ngrams (Aimetti et al. 2009) are able to learn discrete acoustic representations of speech signals which might not be identical to conventional phonemes, but that may eventually obtain phoneme-like status. Interestingly, the number of these acoustic units that will be learned can be determined by global parameters of the learning techniques. These parameters also determine the manner in which new observations will be merged into previously learned representations. It goes without saying that the basic units can be used for recognising novel input.

While the seminal computational models of spoken-word processing basically skipped the prelexical stage, newer models such as SpeM, Fine-Tracker, and the models developed in the ACORNS project feature concrete specifications and operational implementations of the prelexical stage. These models suggest ways for dealing with acoustic variation and idiosyncratic pronunciations in a framework that is still compatible with an essential role for abstract phonemic representations.

5.3 Episodic and abstract models

The discussion of the role of computational modelling as a means for better understanding human speech recognition focused on the 'abstract' theory. What motivated this choice was the large number of computational models and papers on a large number of behavioural and modelling studies that allow us to trace the history and show how computational modelling has advanced the theory. However, we cannot avoid the question whether computational modelling might come into play in the discussion between 'abstract' and 'episodic' models of spoken-word recognition. After all, one might harbour hopes that computational models might prove a class of theories right or wrong. However, from our discussion of the development of models based on the abstract theory it is obvious that we do not believe that any model comes close to providing decisive arguments. That being said, we still believe that computational modelling will be at the heart of future research and theorising in this domain.

There is a hot debate among phonologists whether phonological underlying representations should be considered as one single underlying form that is derivationally mapped onto a phonetic representation (e.g., Chomsky and Halle 1968) or whether phonological underlying representations should be considered as a 'cloud of examplars' (e.g., Bybee 2001; Pierrehumbert 2003). This issue is presented as the crucial difference between episodic and abstract theories and computational models. However, from our discussion of possible solutions for the problem of how to deal with ever new variation and occasional outliers in Section 5.2 it should be clear that we do not believe that abstract and episodic theories are impossible to reconcile. On the contrary, we believe that further investigations of the emergence during language acquisition and the eventual mental representation of 'subword units' (to use a theory-neutral term borrowed from speech technology) will show that these representations do not consist of units that all have the same 'size' and live on a single level. Rather, there are strong arguments in favour of more complex representations, some on the level of sub-phonemic detail, some on the level of what conventionally are called phonemes, yet others on the level of the syllable, and perhaps even representations on the level of frequent word sequences (cf., the discussion on the representations of the word sequence "*I don't know*" in Hawkins (2003)).

If one thing has become clear from the debate about the status of feedback in theories of spoken-word processing, it is how exceedingly difficult it is to design decisive behavioural experiments aimed at elucidating mental representations of sub-word units (McQueen et al. 2009). At the same time, computational models are able to advance our understanding of the issues related to these representations (ten Bosch et al. 2009). Representations that aim at speaker-independent acoustic characteristics (perhaps in the form of statistical distributions) require different learning strategies than representations that are based on clustering speaker-specific observations. And these two different kinds of representations make different (testable) predictions about, for instance, the impact of the number of different speakers that a baby interacts with during the first stage of language acquisition (Newman 2008). If we accept that it is not possible to directly observe 'prelexical' representations in a subject's brain, computational modelling is the next best thing. However, for this purpose we need models that take real acoustic signals as input. In addition, we believe that these models will have to learn the representations as emergent units, rather than as a fixed set of pre-defined phonemes. The demands that these representations must be able to fulfil seem daunting given the seemingly irreconcilable outcomes of behavioural experiments. Yet, we are convinced that models such as Fine-Tracker (Scharenborg 2009) and the models developed in the ACORNS project (ten Bosch et al. 2009) all point in the direction of integrating concepts from the abstract and the episodic theories in explaining the representations on what is still considered as the prelexical level in the spoken-word recognition process.

6. Discussion and conclusion

The principle *divide and conquer* has been extremely successful in the sciences, and especially in the natural sciences where it is possible to isolate phenomena from their context so that they can be analysed in conditions where a small number of possibly relevant factors can be tightly controlled. This research strategy has been extremely influential in psychology and in the cognitive sciences, to the extent that 'modularisation' became a central tenet in many theories, instead of what it originally was: a strategy to come to grips with phenomena that seemed far too complex for a holistic analysis. Modular theories seemed to be supported by early brain research that was aimed at identifying specific regions in the brain as dedicated to specific cognitive functions. However, recent advances in brain imaging strongly suggest that the idea of strong modularisation was wrong: virtually all cognitive functions seem to invoke many different areas in the brain. Therefore, it is not surprising that the cognitive science field is becoming increasingly aware of the arguably unwarranted transformation of 'modularity' from a research approach to a corner stone in many theories.

The idea that cognition is modularised has dominated linguistics for a number of decades, resulting in heated debates about (the lack of) interactions between phonology, morphology, syntax, semantics, etc. While it is of course true that language use is extremely complex and that we are far from a comprehensive theory of language as a cognitive tool (Dascal 2002; Dascal and Dror 2005) we believe that recent developments in modelling language acquisition and speech recognition as an integrated process, starting from the acoustic speech signal and working the way upward until the final understanding of the utterance, have shown that computational modelling holds the promise of coming to grips with the full complexity of language. This is the more so because the modelling approach that we promote can profit from parallel developments in eScience and the science of complexity (e.g., Weaver 1948; Corominas-Murtra, Valverde, and Solé 2009). Without doubt language is Weaver's 'organised complexity' domain where systems are characterised as having emergent, rather than predictable properties, so that they are only amenable to in-depth analysis by means of (computer) simulations.

We believe that the data and the computer power that are needed to simulate the processes involved in language use — rather than trying to reproduce some form of meta-level description of language products — are becoming available. The benefit of such an approach is clearly illustrated by the computational models that start from real speech, rather than from some hand-crafted discrete symbolic representation. The first attempts at this type of models have advanced the theory of spoken-word processing and at the same time elucidated gaps in the theories. Future research in modelling spoken language understanding should concentrate on what is the prelexical stage in current 'abstract' theories. Coming to grips with this stage is an essential prerequisite for understanding the complete picture. And we predict that this line of research will resolve the discussion about abstract versus episodic models, and that it will eventually result in an integrated model, rather than in a modular one.

Notes

* This research was supported by a Veni-grant from the Netherlands Organisation for Scientific Research (NWO) to the first author. The authors would like to thank three anonymous reviewers for their usful comments on an earlier version of this manuscript.

[1] Throughout the paper 'model' and 'computational model' are used interchangeably. Note that in the psycholinguistic literature the term 'model' is sometimes used for something that is closer to what we here refer to as 'theory'.

[2] The term 'computational model' has also been used to indicate research in which relations are investigated between behavioural data and (possibly large numbers of) independent factors. In this paper, we will limit ourselves to computational models that aim to account for the cognitive processes involved in speech comprehension.

[3] For instance, Tuller (2003) presents an example of how testing predictions of not-yet-observed behaviour can further knowledge on switching in speech categorisation. In speech categorisation, subjects listen to stimuli from a continuum, for instance the continuum 'say'-'stay',

where the length of the silent gap after the /s/ is changed in a stepwise fashion. When listeners are presented with the stimulus from one end of the continuum, e.g., 'say', many times over, the position of the perceived category boundary in an identification test of stimuli from the entire continuum will move towards the repeated stimulus, thus 'say'. When subsequently trained with stimuli from the other end of the continuum, the position of the category boundary in an identification test will again move towards the repeated stimulus. Tuller refers to this as the destabilisation of the perception of the stimulus. Tuller's model, however, predicts that when a word is perceived many times over, this destabilisation should not occur. This difference between the literature and the model's prediction motivated two experiments, which tested and confirmed the model's prediction, and showed that not only the number of stimulus repetitions is crucial for category switching, but also how the stimuli move through perceptual space.

4. Brain imaging techniques such as fMRI and ERP are able to show the areas of the brain that are most active in speech processing tasks, but the (interpretation of the) link between activated brain areas and speech recognition processes is not straightforward. So, even when using imaging techniques it is not possible to directly investigate neural processes.

5. For example, an implementation may require a larger number of sequential sub-processes than the brain can perform in the time that is available. A well-known example of such a violation is the theory of voice pitch control proposed by R. Husson (1962).

6. It should be noted that existing episodic theories of spoken-word comprehension are facing similar problems in defining input representations.

7. http://www.mrc-cbu.cam.ac.uk/people/dennis.norris/personal/ last accessed on December 22nd, 2009.

References

Adank, P., Smits, R., and van Hout, R. 2004. "A comparison of vowel normalization procedures for language variation research". *Journal of the Acoustical Society of America* 116: 3099–3107.

Aimetti, G., ten Bosch, L., and Moore, R.K. 2009. "The emergence of words: Modelling early language acquisition with a dynamic systems perspective". *Proceedings of EpiRob-09*, 17–24.

Allopenna, P.D., Magnuson, J.S., and Tanenhaus, M.K. 1998. "Tracking the time course of spoken-word recognition using eye movements: Evidence for continuous mapping models". *Journal of Memory and Language* 38: 419–439.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. 2007. "Automatic speech recognition and speech variability: A review". *Speech Communication* 49: 763–786.

ten Bosch, L., Räsänen, O., Driesen, J., Aimetti, G., Altosaar, T., Boves, L., and Corns, A. 2009. "Do multiple caregivers speed up language acquisition?". *Proceedings of Interspeech*. Brighton, UK.

ten Bosch, L., Van hamme, H., Boves, L., and Moore, R.K. 2009. "A computational model of language acquisition: The emergence of words". *Fundamenta Informaticae* 90: 229–249.

Bradlow, A.R., Bent, T. 2008. "Perceptual adaptation to nonnative speech". *Cognition* 106: 707–729.

Bybee, J. 2001. *Phonology and Language Use.* Cambridge: Cambridge University Press.

Chomsky, N. and Halle, M. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Corominas-Murtra, B., Valverde, S., and Solé, R.V. 2009. "The ontogeny of scale-free syntax networks: Phase transitions in early language acquisition". *Advances in Complex Systems* 12(3): 371–92.

Cutler, A. and Robinson, T. 1992. "Response time as a metric for comparison of speech recognition by humans and machines". *Proceedings of ICSLP*. Banff, Canada, 189–192.

Dahan, D., Magnuson, J.S., and Tanenhaus, M.K. 2001a. "Time course of frequency effects in spoken-word recognition: Evidence from eye movements". *Cognitive Psychology* 42(4): 317–367.

Dahan, D., Magnuson, J.S., Tanenhaus, M.K., and Hogan, E.M. 2001b. "Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition". *Language and Cognitive Processes* 16(5/6): 507–534.

Dascal, M. 2002. "Language as a cognitive technology". *International Journal of Cognition and Technology* 1(1): 35–89.

Dascal, M. and Dror, I. 2005. "The impact of cognitive technologies: Towards a pragmatic approach". *Pragmatics & Cognition* 13(3): 451–457.

Davis, M.H., Marslen-Wilson, W.D., and Gaskell, M.G. 2002. "Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition". *Journal of Experimental Psychology: Human Perception and Performance* 28: 218–244.

Elman, J.L. and McClelland, J.L. 1986. "Exploiting lawful variability in the speech wave." In J.S. Perkell and D.H. Klatt (eds), *Invariance and Variability of Speech Processes*. Hillsdale, NJ: Erlbaum, 360–380.

Garofolo, J.S. 1988. "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database". National Institute of Standards and Technology (NIS), Gaithersburgh, MD.

Gaskell, M.G. and Marslen-Wilson, W.D. 1997. "Integrating form and meaning: A distributed model of speech recognition". *Language and Cognitive Processes* 12: 613–656.

Goldinger, S.D. 1998. "Echoes of echoes?:An episodic theory of lexical access". *Psychological Review* 105: 251–279.

Gow, D.W. and Gordon, P.C. 1995. "Lexical and prelexical influences on word segmentation: Evidence from priming". *Journal of Experimental Psychology: Human perception and performance* 21(2): 344–359.

Grossberg, S., Boardman, I., and Cohen, M. 1997. "Neural dynamics of variable-rate speech categorization". *Journal of Experimental Psychology: Human Perception and Performance* 23(2): 483–503.

Harrington, J. 2010. "Acoustic phonetics". In W.J. Hardcastle, J. Laver, and F.E. Gibbon (eds), *The Handbook of Phonetic Sciences*. Hoboken, NJ: Wiley-Blackwell, 81–129.

Hawkins, S., 2003. "Roles and representations of systematic fine phonetic detail in speech understanding". *Journal of Phonetics* 31: 373–405.

Hintzman, D.L. 1986. "Schema-abstraction in a multiple-trace memory model". *Psychological Review* 93: 411–427.

Husson, R. 1962. *Physiologie de la phonation*. Paris: Masson.

Johnson, K. 1997. "The auditory/perceptual basis for speech segmentation". *OSU Working papers in Linguistics* 50: 101–113.

Jurafsky, D. and Martin, J.H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, 2nd edition. Upper Saddle River, NJ: Prentice Hall.

Klatt, D.H. 1979. "Speech recognition: A model of acoustic-phonetic analysis and lexical access". *Journal of Phonetics* 7: 279–312.

Klatt, D.H. 1989. "Review of selected models of speech recognition". In W.D. Marslen-Wilson (ed), *Lexical Representation and Process*. Cambridge, MA: The MIT Press, 169–226.

Kraljic T. and Samuel, A.G. 2007. "Perceptual adjustments to multiple speakers". *Journal of Memory and Language* 56: 1–15.

Lee, D. and Seung, H., 2001. "Algorithms for non-negative matrix factorization". *Advances in Neural Information Processing Systems* 13: 556–562.

Luce, P.A., Goldinger, S.D., Auer, E.T., and Vitevitch, M.S. 2000. "Phonetic priming, neighborhood activation, and PARSYN". *Recognition & Psychophysics* 62: 615–625.

Luce, R.D. 1959. *Individual Choice Behaviour*. New York: Wiley.

Maier, V. and Moore, R.K. 2005. "An investigation into a simulation of episodic memory for automatic speech recognition". *Proceedings of Interspeech*. Lisbon, Portugal, 1245–1248.

Maier, V. and Moore, R.K. 2007. "Temporal episodic memory model: An evolution of MINERVA2". *Proceedings of Interspeech*. Antwerp, Belgium, 866–869.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman & Co.

McClelland, J.L. and Elman, J.L. 1986. "The TRACE model of speech recognition". *Cognitive Psychology* 18: 1–86.

McQueen, J.M. 2005. "Speech perception". In K. Lamberts and R. Goldstone (eds), *The Handbook of Cognition*. London: Sage Publications, 255–275.

McQueen, J.M., Cutler, A., and Norris, D. 2006. "Phonological abstraction in the mental lexicon". *Cognitive Science* 30(6): 1113–1126.

McQueen, J.M., Jesse, A., and Norris, D. 2009. "No lexical–prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes?". *Journal of Memory and Language* 61(1): 1–18.

Morse, A.F. and Ziemke, T. 2008. "On the role(s) of modelling in cognitive science". *Pragmatics & Cognition* 16(1): 37–56.

Myung, I.J. 2001. "Computational approaches to model evaluation". In N.J. Smelser and P.B. Baltes (eds), *The International Encyclopedia of the Social and Behavioral Sciences.* Oxford: Elsevier, 2453–2457.

Newell, A. 1973. "You can't play 20 questions with nature and win: Projective comments on the papers of this symposium". In W.G. Chase (ed), *Visual Information Processing*. New York: Academic Press, 283–308.

Newman, R.S. 2008. "The level of detail in infants' word learning". *Current directions in Psychological Science* 17: 229–232.

Norris, D., McQueen, J.M., Cutler, A., and Butterfield, S. 1997. "The possible-word constraint in the segmentation of continuous speech". *Cognitive Psychology* 34: 191–243.

Norris, D., McQueen, J.M., and Cutler, A. 2000. "Merging information in speech recognition: Feedback is never necessary". *Behavioral and Brain Sciences* 23(3): 299–370.

Norris, D., McQueen, J.M., and Cutler, A. 2003. "Perceptual learning in speech". *Cognitive Psychology* 47(2): 204–238.

Norris, D. 2005. "How do computational models help us develop better theories?". In A. Cutler (ed), *Twenty-first Century Psycholinguistics: Four Cornerstones.* Hillsdale, NJ: Erlbaum, 331–346.

Norris, D. and McQueen, J.M. 2008. "Shortlist B: A Bayesian model of continuous speech recognition". *Psychological Revie*w 115(2): 357–395.

Peterson, G.E. and Barney, H.L. 1952. “Control methods used in a study of the vowels”. *Journal of the Acoustical Society of America* 24(1): 175–184.

Pierrehumbert, J. 2003. “Probabilistic phonology: Discrimination and robustness”. In R. Bod, J. Hay, and S. Jannedy (eds), *Probability Theory in Linguistics*. Cambridge, MA: The MIT Press, 177–228.

Pitt, M.A., Myung, J.I., Montenegro, M., and Pooley, J. 2008. “Measuring the flexibility of localist connectionist models of speech perception”. *Cognitive Science* 32: 1285–1303.

Salverda, A.P., Dahan, D., and McQueen, J.M. 2003. “The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension”. *Cognition* 90: 51–89.

Salverda, A.P., Dahan, D., Tanenhaus, M.K., Crosswhite, K., Masharov, M., and McDonough, J. 2007. “Effects of prosodically modulated sub-phonetic variation on lexical competition”. *Cognition* 105: 466–476.

Samuel, A.G., Pitt, M.A. 2003. “Lexical activation (and other factors) can mediate compensation for coarticulation”. *Journal of Memory and Language* 48: 416–434.

Scharenborg, O. 2008. “Modelling fine-phonetic detail in a computational model of word recognition”. *Proceedings of Interspeech*. Brisbane, Australia, 1473–1476.

Scharenborg, O. 2009. “Using durational cues in a computational model of spoken-word recognition”. *Proceedings of Interspeech*. Brighton, 1675–1678.

Scharenborg, O., ten Bosch, L., Boves, L., and Norris, D. 2003. “Bridging automatic speech recognition and psycholinguistics: Extending Shortlist to an end-to-end model of human speech recognition”. *Journal of the Acoustical Society of America* 114 (6): 3032–3035.

Scharenborg, O., Norris, D., ten Bosch, L., and McQueen, J.M. 2005. “How should a speech recognizer work?”. *Cognitive Science* 29(6): 867–918.

Scharenborg, O., Wan, V., and Moore, R.K. 2007. “Towards capturing fine phonetic variation in speech using articulatory features”. *Speech Communication* 49: 811–826.

Schuppler, B., van Doremalen, J., Scharenborg, O., Cranen, B., and Boves, L. 2009. “Using temporal information for improving articulatory-acoustic feature classification”. *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*. Merano, Italy.

Schwarz, P., Matějka, P., and Černocký J. 2004. “Towards lower error rates in phoneme recognition”. *7th International Text, Speech and Dialogue Conference*. Brno, Czech Republic, 465–472.

van Segbroeck, M. and van Hamme, H. 2009. “Unsupervised learning of time-frequency patches as a noise-robust representation of speech”. *Speech Communication* 51: 1124–1138.

Tanenhaus, M.K., Magnuson, J.S., Dahan, D., and Chambers, C. 2000. “Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing”. *Journal of Psycholinguistic Research* 29(6): 557–580.

Tanenhaus, M.K. and Spivey-Knowlton, M.J. 1996. “Eye-tracking”. *Language and Cognitive Processes* 11: 583–588.

Tuller, B. 2003. “Computational models in speech recognition”. *Journal of Phonetics* 31: 503–507.

Wade, T., Eakin, D.K., Webb, R., Agah, A., Brown, F., Jongman, A, Gauch, J., Schreiber, T.A., and Sereno, J. 2002. “Modeling recognition of speech sounds with Minerva2”. *Proceedings of the International Conference on Spoken Language Processing*. Denver CO, 1653–1656.

Weaver, W. 1948. “Science and complexity”. *American Scientist* 36: 536.

Authors' addresses

Odette Scharenborg
Centre for Language and Speech Technology
Radboud University Nijmegen
P.O. Box 9103
6500 HD Nijmegen
The Netherlands

O.Scharenborg@let.ru.nl
http://www.odettes.dds.nl

Lou Boves
Centre for Language and Speech Technology
Radboud University Nijmegen
P.O. Box 9103
6500 HD Nijmegen
The Netherlands

L.Boves@let.ru.nl

About the authors

Odette Scharenborg (PhD) is employed at the Radboud University Nijmegen. Her research interests focus on narrowing the gap between automatic and human word recognition, and she publishes widely in both fields. Currently, she holds a fellowship from the Netherlands Organisation for Scientific Research on the topic of the computational modelling of the role of durational information in spoken-word recognition.

Lou Boves is a professor of Language and Speech Technology at the Radboud University Nijmegen. His research is focused on cross-fertilisation of automatic speech recognition and human speech processing. He has published widely in both fields. Recently, his attention is focused on computational modelling of language acquisition, using only basic learning algorithms that avoid implausible prior knowledge in the learning agent.