Complex Word Recognition Behaviour Emerges from the Richness of the Word Learning Environment

Alastair C. Smith

Max Planck Institute for Psycholinguistics,
Wundtlaan 1, Nijmegen, Netherlands.
E-mail: alastair.smith@mpi.nl

Padraic Monaghan

Department of Psychology, Lancaster University,
Lancaster LA1 4YF, United Kingdom

Falk Huettig

Max Planck Institute for Psycholinguistics,
Wundtlaan 1, Nijmegen, Netherlands

Computational models can reflect the complexity of human behaviour by implementing multiple constraints within their architecture, and/or by taking into account the variety and richness of the environment to which the human is responding. We explore the second alternative in a model of word recognition that learns to map spoken words to visual and semantic representations of the words' concepts. Critically, we employ a phonological representation utilising coarse-coding of the auditory stream, to mimic early stages of language development that are not dependent on individual phonemes to be isolated in the input, which may be a consequence of literacy development. The model was tested at different stages during training, and was able to simulate key behavioural features of word recognition in children: a developing effect of semantic information as a consequence of language learning, and a small but earlier effect of phonological information on word processing. We additionally tested the role of visual information in word processing, generating predictions for behavioural studies, showing that visual information could have a larger effect than semantics on children's performance, but that again this affects recognition later in word processing than phonological information. The model also provides further predictions for performance of a mature word recognition system in the absence of fine-coding of phonology, such as in adults who have low literacy skills. The model demonstrated that such phonological effects may be reduced but are still evident even when multiple distractors from various modalities are present in the listener's environment. The model demonstrates that complexity in word recognition can emerge from a simple associative system responding to the interactions between multiple sources of information in the language learner's environment.

Keywords: Speech comprehension; Visual attention; Multimodal processing; Visual world paradigm; Spoken word recognition.

1. Introduction

One of the key aims of computational models is to provide an abstract characterization of a task, in order to determine which cognitive processes may be required in order to simulate observed behaviour. Consideration of the richness of the environment around the cognitive system can produce surprisingly complex behavior as a result of the operation of only very simple internal cognitive processes. Thus, the complexity is in the world, rather than (or as well as) within the system itself.

Yet, decisions not only about what to include, but also what to exclude, in representing the input to the system clearly have huge implications for the validity of the computational model, and also its relevance to the actual task in hand. Language processing is one such domain where input has often been isolated to a single abstracted modality of representation – usually auditory input to the system[7,15,28]. Thus, language models take as input streams of sound, usually pre-segmented into linguistically interpretable units (such as phonemes, morphemes or words) and then test the effect of environmental exposure to this linguistic information on learning mappings to other assumed linguistic representations of language, such as the written form of the word (for a task associated with literacy)[10], or to the meaning of the word (for tasks such as auditory word comprehension)[9], or even a syntactic or discourse structure (for processing of sentences or utterances)[8].

However, in our recent modelling, we have been examining the interacting role of multimodal information sources on language processing, and we have discovered that this interaction is critical to a full appreciation of how the language processing system operates in situ. We have been exploring a model of single word processing that takes into account not only the phonological representations of words, but also the word's meaning, and visual objects present in the environment of the language learner as they acquire their language[18,24,25]. Rather than just considering mapping between pairs of representations, as in classic models of language processing[9,10,17,20,21,22], we have investigated how auditory, visual, and semantic information about words inter-relates in word processing. This has been advantageous for two reasons. First, it has uncovered the nuanced interactions between sources of information and the way in which information integration adjusts over time. The time scales have been both in terms of developmental time, as the model learns more and more words, but also in terms of how a single word processing trial unfolds. Second, the modeling has enabled us to test the role of each type of representational input in the language system's

processing without necessarily requiring those representations to be directly involved in the task.

Phonological awareness tasks have frequently been used as tools for probing the structure of representations activated by the speech signal, which interface with the broader language processing system[4]. Phonological awareness can be defined as the ability to manipulate the phonological structure of spoken words. Typical tasks used to assess phonological awareness comprise tests of participants' ability to explicitly adjust the sequence of phonemes in the word[14]. One example is phoneme isolation tasks, where participants are required, for instance, to report the third phoneme in a word. An alternative is a phoneme deletion task, where participants have to report a word or non-word phoneme sequence with one sound removed, e.g., what is "facts" without the /t/ sound.

Phoneme awareness is a major predictor of literacy outcomes, with good performance highly correlated with vocabulary size, reading level, later reading comprehension, and is also an early indicator of developmental dyslexia[2,5,6,14]. Phoneme awareness is also associated with the effect of literacy[30] on the language processing system[23]. Literates in alphabetic languages demonstrate more accomplished phoneme awareness performance than those who have low-literacy or no literacy skills[1,19]. Thus, phoneme awareness, though a predictor of literacy outcomes, is also shown to be at least partially a consequence of literacy development itself. One problem with phoneme awareness tasks is that the phonological representation has to be explicitly manipulated. Thus, it is not only a test of the phonological representation but also the speaker's meta-cognitive ability to manipulate this representation.

An alternative way to assess changes in phonological representation without also requiring direct explicit access to that representation, is to test the effect of phonological processing on other language processing tasks. Huettig, Singh, and Mishra[13] did precisely this, by determining the effect of literacy on processing spoken word input on activation of words with overlapping phonological representations. The extent of the overlap clarifies the granularity of the listener's phonological representation during word processing, where if there is overlap of a single phoneme then sensitivity to this overlap indicates a phoneme-level representation of the input. As an outcome measure, Huettig *et al.*[13] presented the listener with a set of objects, one of which was named by a word similar in sound to the auditory input, one of which was named by a word with similar meaning to the auditory input, and two distractor objects, which were unrelated in sound or meaning to the heard word. The time-course of eye-movements to the phonologically-related, or the semantically-related word indicated the extent to which phonology[3] and semantics[11,29] were affecting word processing at different

points in the unfolding of the speech signal. They found that both high-literacy and low-literacy participants looked to the semantically-related object more than to the distractor objects, but that only the high-literacy participants looked more to the phonologically-related object. Furthermore, this phonological effect was observed at an earlier point in the word processing than the semantic effects for the high-literacy individuals. In a second experiment low-literates were shown to display a phonological effect when scenes contained a single phonologically related item accompanied by unrelated distractors but unlike literates this effect was weak, late and pro-longed. This suggested that the low-literacy participants were less able to access the individual phonemes within the word during auditory word processing, thus providing an indirect measure of phonological awareness.

Smith, Monaghan, and Huettig[25] constructed a multimodal language processing model that aimed to simulate the cognitive effects of literacy on phonological representations for visual world language processing tasks. The model is illustrated in Figure 1. The model has a central processing resource, which is connected to and from a variety of modalities that can be inputs and/or outputs to the language processing system. In Figure 1, visual, phonological, and semantic information is indicated, and the output of the model is a set of units representing the fixation position of the eye. Two versions of the model were compared. The first was able to process the individual phonemes in the word as the auditory input unfolded over time. The second was able only to process at a coarser granularity, with syllables, or whole words, rather than individual phonemes processed from the input. The model with phoneme-level representations was able to simulate both the later semantically-related and the early phonologically-related influences on eye movements displayed by literates. The model with syllable-level representations simulated the later semantically-related effects and weaker, prolonged phonological effects displayed by illiterates. Thus, the model showed that the effects of literacy on visual world language processing were consistent with changes from coarse- to fine-coding representations of the word's phonology.

The developmental trajectory of the same multimodal model's processing was also investigated[27]. In this simulation, we investigated the simultaneous involvement of phonological, semantic, and visual information in auditory word processing. The same model as shown in Figure 1 was trained to learn to link auditory representations, that were akin to the fine-grained phoneme representations in the literacy simulations, to visual and semantic representations of words. After training, the model was tested on its eye-movement simulated behaviour when presented with visual scenes comprising a phonologically-related object, a semantically-related object, a visually-similar object, and an unrelated

object. The model was able to simulate the temporal order of all three effects in adult word processing: an effect of phonology early in word processing, and then a later effect of semantic and visual relatedness. Analysis of semantic activity in the model indicated that early phonological effects were partly driven by activation of the semantic properties of the phonologically-related object at levels similar to that of the target during periods in the unfolding of the spoken word in which the phonology overlapped.

The development of these effects as the model learned gradually to map between each of the individual representations, in a simulation of vocabulary development, was also consistent with patterns of phonological, semantic, and visual processing in early and later childhood. Mani and Huettig[16] tested children aged 2, 4, 6, and 8 years old on word processing when viewing a phonologically-related, a semantically-related and two unrelated distractor pictures. They found that semantic effects increased with age, whereas phonologically-related effects became more stable as age increased, altering from a long-lasting effect after the word had been spoken, to resemble more closely adult performance, with a more focused phonological processing effect occurring closer in time to the word's spoken production.

However, these developmental simulations were, as mentioned, conducted with a fine-coding phonological representation where speech was represented as a sequence of individuated phonemes, whereas this is unlikely to occur at earlier stages of word learning, such as ages 2 and 4 years old, with children only beginning to transition to phoneme grain-size representations at or after onset of formal literacy (e.g., ages 6 and 8 years old).

In this chapter we report simulations to test whether it is possible to simulate the developmental effects of multimodal processing during word recognition when a more realistic, coarse-coding representation of phonology is initially available to the model. A further issue explored by the simulations in this chapter is to generate predictions about how the effects of phonological, semantic, and visual-relatedness unfold over a single trial in low-literacy participants who do not have fine-grained phoneme level representations of phonology. Previously, effects of literacy in visual world language processing tasks have not also included visually-related objects, and these have been shown to dramatically reduce observable effects of phonological relatedness[26]. Thus far, participants with low-literacy have not been tested on the relative role of visual processing in the visual world paradigm and so the current simulations provide also predictions for future behavioural investigations.

2. Modelling multimodal language development

2.1. *Method*

2.1.1. *Architecture*

The model is similar to that employed in Smith *et al.*[27], and illustrated in Figure 1. The model comprised several modalities connecting to a set of 400 hidden units in an integrative layer: a set of 80 visual units, with 20 units each associated with one of four object positions in the visual environment; a set of 60 phonological units representing the spoken form of a monosyllabic word; and 200 semantic units, to represent the word's meaning. The integrative, hidden layer was self-connected, and also projected to the semantic layer. This was in order that the model could also generate semantic representations, and not only use semantic information as an input source. As output from the model, there was a set of four units, each representing one of four visual object positions, to which the eye could be directed. This eye layer was also connected back to the integrative layer. The eye layer enabled a read-off of the model's behaviour regarding simulating eye-movement behavioural experiments using the visual world paradigm.

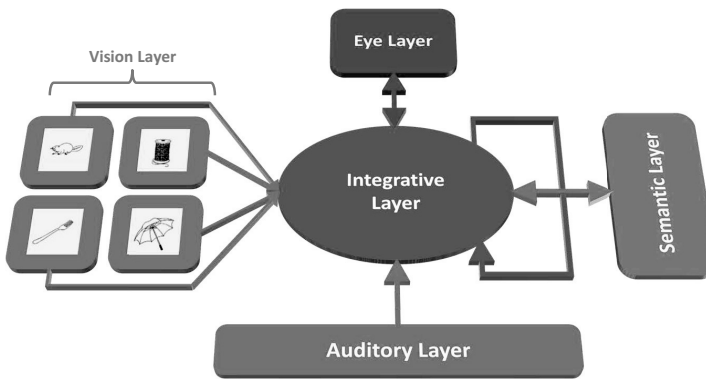

Fig. 1. Architecture of the Multimodal Integration Model of language processing (MIM).

2.1.2. *Training and testing*

There were 200 items altogether, each with a unique visual, semantic, and phonological representation. Each visual object representation was composed of 20 units, 10 representing low spatial frequency and 10 representing high spatial frequency visual features. Each unit was active with probability of 0.5. For the semantic representation, a randomly selected unique set of eight units were active

for each different word. For the phonological representation, this was encoded at the syllable level, so each word was represented as a unique pattern over a set of 60 units, with each unit active with probability 0.5. This differed from simulations of the model where phonology was encoded at the phoneme level, with a unique pattern distributed over 10 units representing each phoneme and each word encoded by a unique combination of 6 phonemes in the input[27]. Thus, the current simulations were a coarser-coding of the auditory input than fine-grained, phoneme-level representations that have been used to simulate word processing in participants who have phoneme-level awareness of speech stimuli.

To construct stimuli to mimic the visual world paradigm tasks, we generated representations for a subset of words that were related in terms of their visual, semantic, or phonological properties. Within the 200 items were embedded 80 competitor and unrelated distractor items that were controlled for their relationships to 20 target items. This included twenty visual competitors (which each shared their low-spatial frequency visual features with a target item), twenty semantic competitors (which shared 4 of their 8 active semantic features with a target), and 20 phonological competitors, (each of whose phonological pattern was identical to 50% of the yoked target's phonological pattern, while the remainder overlapped randomly). Twenty unrelated distractors were also constructed that were designed to have minimal overlap in visual, semantic and phonological dimensions (see Smith *et al*.[25], for more details).

The model was trained to learn to map between each pair of modalities, in order to simulate the model's development of learning relationships between representations. Thus, for a vision to meaning mapping, simulating an object recognition task, the model was given a visual object representation at one of the four possible visual positions (the location of the object in the visual field was randomised across training trials) and the eye layer unit corresponding to its location was fully activated. The system was then required to learn to activate the semantic pattern corresponding to that object. Similarly, for phonology to semantic mapping a phonological input was provided and random noise as input to the visual layer, with the system required to activate its corresponding semantic pattern. For a phonology to vision mapping, in an object naming task, the model was presented with the phonological input and four different objects in each of the four visual positions randomly selected from the training set (which included all competitors and unrelated distractors), and was required to activate the eye position associated with the target object. The same procedure but with a semantic input rather than a phonological input, was used to train mapping from semantics to vision. For each trial there were 15 time steps, where the input was presented at time step 1, then the model had a further 14 steps in order to activate the other

representations associated with the word, and to generate an activation in the eye movement units. The relative activation of each eye position, associated with one of the locations in the visual field of containing a distractor, enabled an assessment of the model's activation of phonological, visual, and semantic information during word recognition.

The model was trained for 1,000,000 separate training trials, and the model's performance was assessed at 250,000 intervals to simulate developmental stages of word learning. The learning rate was 0.05, and weights in the model were adjusted using recurrent backpropagation after each trial, with error computed as the mean squared difference between the actual and target output at the last time step of the trial. Results were averaged over 8 instantiations of the model.

During testing, the model was presented with a simulation of the visual world paradigm, whereby the phonology of a word was presented to the model (time steps 5 - 15), along with a set of objects presented at the visual input (time steps 1 - 15). The objects in the visual display represented the phonologically-related, visually-related, and semantically-related patterns, along with an unrelated distractor item. There were in total 480 testing trials, with 24 different arrangements of object location for each of the 20 words that had an overlapping phonological, semantic, and visual pattern.

The model's performance was assessed by determining the eye-movement over the 15 time steps of the training trial with an additional 15 time steps added to discern further the time-course of activation of the different representational modalities. The influence of overlap in each representational modality was appraised by determining the strength of activation of the eye unit corresponding to the phonologically, semantically, or visually related object, relative to activation of the eye unit corresponding to the unrelated object. We additionally recorded the cosine-distance between activation in the semantic layer and the semantic representations corresponding to each item in the display and the spoken target word in order to examine its relationship to fixation behavior.

In this simulation, we explored two broad aspects of the data. First, we wanted to examine the extent to which the model was able to simulate developmental effects of a growing contribution of semantic information with experience, and also to determine whether, in the model, the emerging early effects of the phonological distractor were evident as a consequence of literacy – so a change from coarse-coding to fine-coding. Discovering that the early phonological distractor effect is not evident in the model at the end of training would be an indication that literacy is required before phonological distractors can exert an effect. However, if the model can demonstrate an early effect of the phonological distractor then this indicates that it is experience rather than a

literacy-related qualitative change that can induce the observed behavioural effects. Furthermore, previous developmental behavioural data is based on tests including just a phonological and semantic distractor, so the model enables a further prediction of the role of visual distractors in processing.

Second, we wanted to examine whether the model could reproduce previous findings of a semantic distractor effect, with only a very small phonological distractor effect, in low-literates[13,25], i.e., with proposed coarse-coding of phonological input, when a visual distractor was also present. This, too, was a novel investigation of the task as previous simulations and behavioural studies had not tested the effect of also including visually-related distractors in the display. In this case, we also wanted to explore the time-course of fixation to the visual distractor, to generate predictions about how such multiple modalities might influence word recognition performance in a low-literacy population.

2.2. *Results*

The model's performance was assessed at each of the four developmental stages, with fixation of phonologically related, semantically related, visually related and unrelated objects tracked as the trial unfolded. The results are shown in Figure 2.

Early in development, after 250,000 training trials, the model demonstrated no significant eye movement activation to any of the related distractors relative to the unrelated distractor, all $|t| \leq 1.130$, all $p \geq .296$.

However, from 500,000 trials onwards, the model demonstrated a growing semantic distractor effect. For 500,000 trials, the semantic distractor resulted in a higher activation than the unrelated distractor from time step 15 through to the end of the trial at time step 30, all $|t| \geq 2.641$, all $p \leq .033$, with a maximum ratio difference of 1.745, meaning that the semantic distractor position resulted in 1.745 times the activation than the unrelated distractor. For 750,000 trials, the effect was larger and earlier, beginning at time step 14, all $|t| \geq 2.693$, all $p \leq .031$, with maximum ratio of 2.285, It was earlier still, from time step 13, after 1,000,000 training trials, all $|t| \geq 2.648$, all $p \leq .033$, but the maximum ratio of 2.054 was no greater than after 750,000 trials.

In terms of activation of phonology, the model showed an effect of the phonological distractor from 500,000 trials, with activation of the eye position associated with the phonologically-related object greater than that of the unrelated distractor from time step 13, all $|t| \geq 2.434$, all $p \leq .045$, maximum ratio = 1.410. After 750,000 trials, the effect was earlier, from time step 12, all $|t| \geq 2.513$, all $p \leq .040$, maximum ratio = 1.667, and after 1,000,000 trials, it continued to be significant from time step 12, all $|t| \geq 2.630$, all $p \leq .034$, with maximum ratio =

1.525. Interestingly, as for the simulations in Smith et al.[27] using a fine-coding of the phonological input, the model demonstrated a prolonged effect of phonological relatedness, which was consistent with the developmental data, but not with adult performance for such visual world tasks. Furthermore, as in the developmental behavioural data, the phonological effect was substantially smaller than that of the semantic effect (as seen in Figure 2, panels B, C and D).

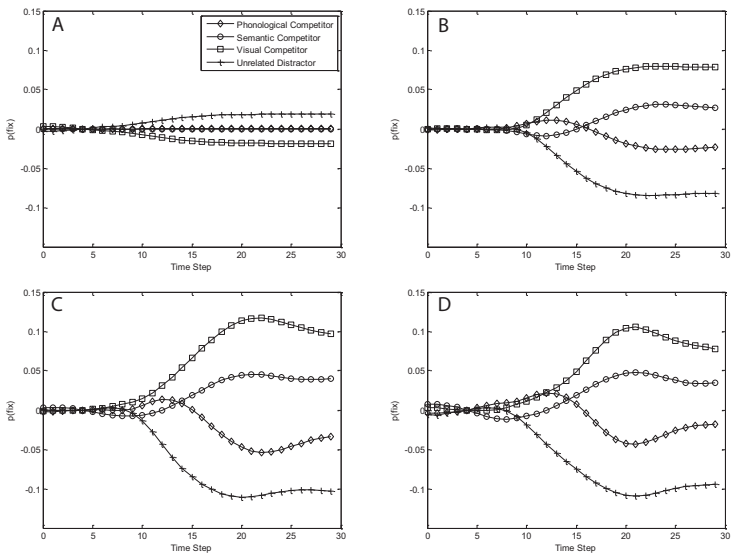

Fig. 2. The MIM model's performance on phonological, semantic, and visual distractor stimuli, compared against an unrelated distractor, across 30 time steps of a single trial. A after 250,000 training patterns; B after 500,000 training patterns; C after 750,000 training patterns; and D after 1,000,000 training patterns.

For the novel predictions of the visual representation effects, these were also observable from mid- to late-stages of development of the model. After 500,000 trials, there was a significant effect of the visual distractor from time step 14 onwards, all $|t| \geq 3.045$ all $p \leq .019$, maximum ratio = 2.051. After 750,000 trials, the effect was earlier and larger, from time step 13, all $|t| \geq 2.891$, all $p \leq .023$, maximum ratio = 2.966. Further training resulted in an earlier and more reliable difference, from time step 12, all $|t| \geq 3.996$, all $p \leq .005$, but not a larger effect: maximum ratio = 2.539.

The mean cosine similarity between semantic layer activity and the semantic representations of each of the visually displayed objects and the spoken target word was calculated at each time step within test trials performed by networks

after 1,000,000 training trials. The difference from this measure calculated at word onset (time step 5) for all subsequent time steps is displayed in Figure 3 in addition to the level of activation of eye layer units corresponding to each category of object in the visual display. These data indicated strong activation of the semantic properties of both the semantically related object and the spoken target word but little activation of the semantic properties of the visually related object, the phonologically related object and the unrelated object. This contrasts with simulations in Smith et al.[27] using a fine-coding of the phonological input which showed early strong activation of the semantic properties of the phonological competitor.

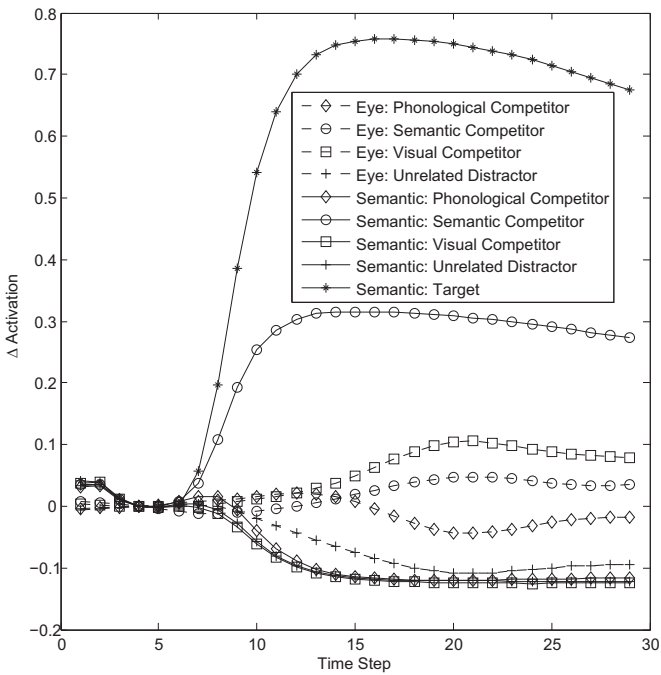

Fig. 3. Change in semantic layer and eye layer activity from word onset in the MIM model after 1,000,000 training patterns across 30 time steps of a single trial simulating visual world conditions with scenes containing a phonological, a semantic, a visual and an unrelated distractor. [Eye = activation of eye layer unit corresponding to location occupied by given category of object; Semantic = difference from word onset in cosine similarity calculated between current semantic layer activation and semantic representation of given item (i.e. target, unrelated distractor, visual competitor, semantic competitor, phonological competitor)].

3. Discussion

The response of the model at the end of training demonstrated that the model was able to replicate low-literacy behaviour for phonologically-related and semantically-related distractors, even when including also a visual distractor. The model's predictions, at all stages of development, indeed, suggested that coarse-coding resulted in a growing effect of semantic and visual representations influencing eye-gaze behaviour, and a slight but evident effect of phonological distractors, even in the presence of visual distractors and with coarse-coding of the phonology.

Thus, the model was able to reproduce previous computational investigations of adult behaviour with fine-coding of phonological representations[27], as well as reproducing key behavioural effects from the visual world paradigm[12]. Indeed, data from a visual world paradigm with phonological, visual, and semantic distractors indicates behaviour that is very similar to that of the model unfolding over time[12]. Such data shows an early deviation of eye fixations to the phonological distractor which then recede, and are later overwhelmed by a growing effect of the semantic and the visual distractors, with these remaining in influence until the end of the trial. As with our previous simulations[25], the model's temporal dynamics are extremely close to those of human behaviour.

So, what are the reasons for this nuanced time-course of the different representational modalities during word recognition? In the model, we can precisely define the constraints that result in the temporal patterns of behaviour. Initially, the phonological form of the word becomes active in the model's integrative layer, then this is aligned with information from the visual display resulting in early but weak fixations to phonologically related objects. At a later stage, information from activated semantic properties of the spoken word in the semantic layer flows back into the integrative layer causing eye movements to begin to move to the semantically related object. The visually-related distractor is also slower in activation because it requires the phonology to generate a code for the visual properties that are to be observed in the object array, and so this too is a later effect than the initial phonological relatedness. However, though the visual and the semantic effects are indirectly triggered by the phonological input, they are longer-lasting and more robust than the phonologically related distractor. This may be because as the phonological signal is produced, the model eventually receives perfect information that the phonologically related object is not a match to the actual phonological input. However, the indirect generation of semantic and visual information from phonology never provides an explicit signal that the mismatch is suboptimal, which may be the reason why negative information from

the phonology exerts a smaller effect in suppressing looks to these other distractor modalities.

Regarding predictions of the model for testing with children and low-literacy populations, the model's results are consistent with previous studies of different phonological encoding affecting activation of phonological information. As in Smith et al.[25], we found that the size of the phonological effect was much reduced compared to semantic distractor effects, and also as compared to a model with fine-coding of phonology[27], where the auditory stream is parsed into individual phonemes, rather than, as in the current simulation, into a syllable, or word-level representation. In contrast to the fine-coding results, coarse-coding did not result in early activation of the semantic properties of the phonological competitor. Therefore, at early stages of processing, semantic information was not available to drive fixation of the phonologically related distractor, hence early phonological effects were much reduced.

Furthermore, the model's results suggest that visual information would be encoded in a similar manner in low-literacy populations as in high-literacy groups. For visual and semantic effects there was no qualitative difference between the model's performance and that of behavioural data with all distractor types present[12], even with coarse-coded input. Thus, the effect of literacy affects only the nature of the effect of phonological distractors, and not the contribution of both semantic and visual information in word recognition. Similarly, for the developmental data, the modeling here makes predictions that addition of a visual distractor will not affect the developmental profile of the phonological distractor effects that increase as the learner acquires more experience of the language. The visual distractor may reduce the apparent size of the effect of a phonologically-related distractor, but not its qualitative contribution to behaviour.

Thus, the model provides an illustration of the benefits of considering the confluence of a rich, multimodal environment converging on the cognitive system. The cognitive structure as implemented in the model, then, is merely an associative network, thus subtle behaviour is entirely a consequence of constraints from the representations themselves, and the requirements of tasks that directly or indirectly activate individual representations. The value of such a multimodal approach to modeling language means that observations of phenomena, such as phoneme awareness ability, can be grounded in processes that make sense of the role of such skills in the language processing system. Fine-coding of phonological input, as demonstrated in comparisons between previous simulations using such phoneme-level segmented representations[27] versus the coarse-coding of the current model, enhances the role of phonology in word recognition, both by

increasing the size of the influence of phonology and by introducing its influence earlier in the process of recognising a spoken word.

The modeling here presents an enterprise that draws together multiple sources of information to simulate language processing in the presence of various, sometimes competing, sources of information regarding the meaning and referent of a spoken word. The behavioural effects of the model, in terms of the role of different modalities, and the time-course of their influence, are emergent properties of the multiple, co-occurring sources of information that surround the listener, both in measures of competent adult performance as well as affecting development of the language processing system. Such apparently complex behaviour is thus the result of a very simple system reacting to multiple, complex representations for words.

Acknowledgments

The second author was supported by the Economic and Social Research Council (UK) (grant number ES/L006936/1).

References

1. Adrián, J. A., Alegria, J., & Morais, J. (1995). Metaphonological abilities of Spanish illiterate adults. *International Journal of Psychology*, **30(3)**, 329–351.
2. Alcock, K. J., Ngorosho, D., Deus, C., & Jukes, M. C. H. (2010). We don't have language at our house: Disentangling the relationship between phonological awareness, schooling, and literacy. *British Journal of Educational Psychology*, **80(1)**, 55–76.
3. Allopenna, P. D., Magnuson, J. S., & 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(4)**, 419–439.
4. Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science*, **14(5)**, 255–259.
5. Burnham, D. (2003). Language specific speech perception and the onset of reading. *Reading and Writing*, **16(6)**, 573–609.
6. Caravolas, M., Volin, J., & Hulme, C. (2005). Phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies: Evidence from Czech and English children. *Journal of Experimental Child Psychology*, **92**, 107–139.

7. Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, **10(7)**, 335–344.
8. Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, **6(2)**, 78–84.
9. Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, **12(5–6)**, 613–656.
10. Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review*, **106(3)**, 491–528.
11. Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, **96(1)**, B23–B32.
12. Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language- mediated visual search. *Journal of Memory and Language*, **57(4)**, 460–482.
13. Huettig, F., Singh, N., & Mishra, R. K. (2011). Language-mediated visual orienting behavior in low and high literates. *Frontiers in Psychology*, **2**, 285.
14. Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The causal role of phoneme awareness and letter- sound knowledge in learning to read combining intervention studies with mediation analyses. *Psychological Science*, **23(6)**, 572–577.
15. Levelt, W. J. (1999). Models of word production. *Trends in Cognitive Sciences*, **3(6)**, 223–232.
16. Mani, N., & Huettig, F. (submitted). The changing dynamics of word-referent mapping across development. *Manuscript submitted for publication*.
17. McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18(1)**, 1–86.
18. Monaghan, P. & Nazir, T. (2009). Modelling sensory integration and embodied cognition in a model of word recognition. In J. Mayor, N. Ruh, & K. Plunkett (Eds.), *Connectionist models of behaviour and cognition II*, pp.337–348. Singapore: World Scientific.
19. Morais, J., Content, A., Cary, L., Mehler, J., & Segui, J. (1989). Syllabic segmentation and literacy. *Language and Cognitive Processes*, **4(1)**, 57–67.
20. Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, **115(2)**, 357–395.

21. Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, **103**, 56–115.
22. Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, **96(4)**, 523–568.
23. Serniclaes, W., Ventura, P., Morais, J., & Kolinsky, R. (2005). Categorical perception of speech sounds in illiterate adults. *Cognition*, **98(2)**, B35–B44.
24. Smith, A. C., Monaghan, P., & Huettig, F. (2013). An amodal shared resource model of language-mediated visual attention. *Frontiers in Psychology*, **4**.
25. Smith, A.C., Monaghan, P., & Huettig, F. (2014). Literacy effects on language and vision: Emergent effects from an amodal shared resource (ASR) computational model. *Cognitive Psychology*, **75**, 28–54.
26. Smith, A.C., Monaghan, P., & Huettig, F. (submitted). Constraints on multimodal integration during spoken word processing: Testing predictions of a parallel integration model of language mediated visual attention. *Manuscript submitted for publication*.
27. Smith, A. C., Monaghan, P., & Huettig, F. (submitted). The multimodal nature of spoken word processing in the visual world: Language acquisition in a multimodal parallel integration model. *Manuscript submitted for publication.*
28. Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. Wiley Interdisciplinary Reviews: *Cognitive Science*, **3(3)**, 387–401.
29. Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **32(1)**, 1–14.
30. Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin*, **131(1)**, 3–29.