

Modeling reading development:

Cumulative, incremental learning in a computational model of word naming

Padraic Monaghan

*Lancaster University, Lancaster, England*

Andrew W. Ellis

*University of York, York, England*

August 2010

Running head: A developmental model of reading

Please address correspondence to Padraic Monaghan, Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK. (Tel: +44 1524 593813 Fax: +44 1524 593314 Email: [p.monaghan@lancaster.ac.uk](mailto:p.monaghan@lancaster.ac.uk))

## Abstract

Natural reading development gradually builds up to the adult vocabulary over a period of years. This has an effect on lexical processing: early acquired words are processed more quickly and more accurately than later acquired words. We present a connectionist model of reading, learning to map orthography onto phonology to simulate this natural reading development. The model learned early words more robustly than late words, and also showed interactions between age of acquisition and spelling-sound consistency that have been reported for skilled adult readers. In additional simulations, we demonstrated that age of acquisition effects are a consequence of incremental exposure to words in concert with changes in plasticity as learning proceeds, and are not due to uncontrolled differences in ease of reading between early and late acquired words. Models which do not learn through cumulative training are unable to explain age of acquisition and related effects.

Key words: reading, word naming, age of acquisition, consistency, frequency, neighbors, computational modelling.

Computational models of reading map orthographic input representations (letters or letter sequences) onto phonological (sound-based) and/or semantic (meaning-based) representations. In the model of Harm and Seidenberg (1999), for example, the input units represent the letters in written words while the output units represent phonological features of phonemes (individual speech sounds) in spoken words. The model is trained over time to generate the appropriate phonological output for each of the different inputs that correspond to the written words it learns to “read”. As Harm and Seidenberg (1999) observed, their model built on several previous computational models of reading (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989) and has since been extended by others (Harm & Seidenberg, 2004; Perry, Ziegler, & Zorzi, 2007). A feature of all of these attempts to simulate visual word recognition is that the words that the model is required to learn are all entered into training together from the outset. But as Harm and Seidenberg (1999) also noted, natural reading development is not like that: young children embarking upon the task of learning to read will typically begin with a first reading book that contains 5 or 10 words, repeated several times. The next book will repeat those words and add a few more, and so on. For a reader of English it can take 10 to 15 years to acquire a full, adult reading vocabulary (Nation, 2009). This process of gradual, cumulative, incremental training that starts with a few words and builds a vocabulary over time is quite unlike the training implemented in existing computational models of visual word recognition.

The present study aimed to explore the consequences of training a computational model of reading in a more natural way. Our goal was to discover whether the final state of a developmentally-trained model would show an influence of the point at which words were entered into training and, if it did, whether that influence would be greater for words with irregular or exceptional spelling-sound correspondences than for words with regular, consistent correspondences. The reasons for investigating these possibilities are embedded in the literature on effects of age of acquisition (AoA) in adult reading. The next section will summarise the relevant

findings.

### **Age of Acquisition Effects in Adult Visual Word Recognition**

The claim that the age at which words are learned might affect the speed with which they can be processed in adulthood was first made in relation to object naming. Carroll and White (1973) analysed naming latencies for 94 object pictures and found that they were better predicted by a measure of the AoA of the object names than by a measure of their word frequency. The effect of AoA on object naming speed has been replicated many times, with several more recent studies reporting effects of both AoA and frequency (e.g., *English*: Barry, Morrison, & Ellis, 1997; Ellis & Morrison, 1998; Meschyan & Hernandez, 2002; *French*: Alario et al., 2004; Bonin, Chalard, Meót, & Fayol, 2002; *Dutch*: Ghyselinck, Lewis, & Brysbaert, 2004a; *Italian*: Bates, Burani, D'Amico, & Barca, 2001; *Spanish*: Cuetos, Ellis, & Alvarez, 1999).

Carroll and White's (1973) report of AoA effects in object naming stimulated a quest for comparable effects in visual word recognition. Whaley (1978) and Butler and Hains (1979) obtained effects of both AoA and word frequency on lexical decision speed in English, with fastest recognition of early acquired, high frequency words and slowest recognition of late acquired, low frequency words. The effect of AoA on visual lexical decision has again been replicated many times (e.g., *English*: Cortese & Khanna, 2007; Gerhand & Barry, 1999; Morrison & Ellis, 1995; 1999; Nagy, Anderson, Schommer, Scott, & Stallman, 1989; *French*: Bonin, Chalard, Méot, & Fayol, 2002; *Dutch*: Brysbaert, Lange, & Van Wijnendaele, 2000a; *Italian*: Colombo & Burani, 2002). Juhasz and Rayner (2003) found effects of both AoA and word frequency on eye fixation durations in reading text. The age (or order) of acquisition of words in both first and second languages have been shown to affect the speed with which written words can be translated from one language into another and the speed with which pairs of written words can be judged to have the same meaning in two different languages (Izura & Ellis, 2002; 2004; Murray, 1986). The age at which words were acquired in both Spanish (L1) and English (L2) affected translation judgement

RTs in the study of Izura and Ellis (2004), implying that such effects do not depend on the early words having been learned during some putative “critical period” of early childhood; rather it would appear that each time a new vocabulary is acquired, a new AoA effect is initiated.

Finally, AoA has been reported on word naming speed. Multiple regression analyses of English word naming latencies by Gilhooly and Logie (1981), Brown and Watson (1987), Morrison and Ellis (1999), and Ellis and J. Monaghan (2002)<sup>1</sup> all found effects of AoA in analyses that also included measures of word frequency, imageability and other psycholinguistic variables as predictors. Factorial studies which compared naming speeds for sets of early and late word sets have also reported effects of AoA (V. Colheart, Laxon, & Keating, 1988; Gerhand & Barry, 1998; J. Monaghan & Ellis, 2002a; b; Morrison & Ellis, 1995). Note, however, that these studies are generally limited to monosyllabic words, and the psycholinguistic variables may contribute differently for polysyllabic words (New, Ferrand, Pallier, & Brysbaert, 2006).

In a large-scale study of effects of AoA on naming and lexical decision, Cortese and Khanna (2007) obtained AoA ratings for 2,342 words and investigated whether AoA predicted variance after N, orthographic length, frequency, and phonological properties of the onset had all been entered into a hierarchical regression analysis in previous steps. They found that AoA predicted unique variance for both naming and lexical decision, but that this effect was greater for the lexical decision task. Imageability predicted lexical decision latencies but not word naming speed, which Cortese and Khanna interpreted as evidence for greater semantic involvement in making lexical decision responses than in (English) word naming, and this was taken by them to indicate that AoA may contribute via semantic representations of words. Comprehensive reviews of AoA effects in lexical processing and other situations can be found in Hernandez and Li (2007), Juhasz (2005) and Johnston and Barry (2006).

---

<sup>1</sup> Note that the J. Monaghan of Monaghan and Ellis (2002a; b) and Ellis and Monaghan (2002) is not the same person as the first author of this paper. She is, in fact, the first author's sister.

### **Criticisms of Age of Acquisition Effects in Visual Word Recognition**

The claim that the AoA of words (or the order in which they are acquired) affects their speed of processing in adulthood has proved controversial (Zevin & Seidenberg, 2002; 2004; see also Strain, Patterson, and Seidenberg, 2002, and the response by Ellis and J. Monaghan, 2002). All of the studies of AoA effects in visual word recognition have sought to match their early and late acquired word sets on the frequency with which they occur in adult language. Zevin and Seidenberg (2002) demonstrated, however, that sets matched on one frequency measure are not necessarily matched on another. Thus, Morrison and Ellis (1995) compared naming and lexical decision latencies for a set of early and late acquired words that were matched on Kucera and Francis (1967) frequencies, but those sets proved to be unmatched on WFG word frequency (Zeno et al., 1995). This is a valid criticism, though Zevin and Seidenberg (2002) never actually demonstrated that the residual differences in frequency on one measure when another is controlled would be sufficient to generate differences in RTs of the magnitude observed for early and late acquired word sets in the studies in question (Johnston & Barry, 2006). We note also that the effect of AoA remained significant in a regression analysis of English word naming RTs by Ellis and J. Monaghan (2002) when both Kucera and Francis (1967) and CELEX (Baayen, Pipenbrock, & Gulikers, 1995) measures of word frequency were included as predictors.

Zevin and Seidenberg (2002) also observed that if two words have equal frequencies of occurrence in adult language, the one that is early acquired is likely to have been encountered more often throughout life than the one learned later. An apparent effect of AoA could therefore reflect differences in the total, *cumulative* frequency with which words have been experienced (see also Lewis, 1999a; b; Lewis, Gerhand, & Ellis, 2001). The question of whether cumulative frequency can account for apparent AoA effects was considered in studies of French and Dutch word naming by Bonin et al. (2004) and Ghyselinck et al. (2004a), respectively. Bonin et al. (2004) analysed word naming latencies for a set of 190 French words for objects, using objective AoA values based on children's ability to name the corresponding object pictures (Chalard, Bonin, Méot, Boyer, &

Fayol, 2003) along with measures of word frequency, familiarity, imageability, number of orthographic neighbors, word length and initial phoneme characteristics. Objective AoA made a significant independent contribution to predicting word naming latency even when a measure of cumulative word frequency was also included as a predictor. When a measure of “frequency trajectory” based on a comparison of word frequencies in material written for adults and children was also introduced into the analyses, the frequency trajectory predictor was not significant and its inclusion did not affect the ability of objective AoA and cumulative frequency to predict word naming speed.

Ghyselinck et al. (2004a) compared the effects of AoA and adult word frequency across a variety of lexical processing tasks. In general the two effects were correlated, so that tasks which showed large effects of AoA tended also to show large effects of frequency (and vice versa). Ghyselinck et al. (2004a) argued that if AoA effects reduce to effects of lifespan (cumulative) frequency, then frequency and the time a word has been known (participant age minus AoA) should predict reaction times such that the regression coefficients for the two factors would be equal. In contrast, across a range of tasks, the contribution of the time-known factor (AoA-based) was significantly greater than the contribution of the frequency factor, with the impact of time known being about 10 times the size of the impact of word frequency. Ghyselinck et al. (2004a) also noted that if AoA effects were really cumulative frequency effects in disguise, then the observed effect of AoA should be less in older participants compared to younger participants, which does not seem to be the case (Barry, Johnston, & Wood, 2006; Morrison, Hirsh, Chappell, & Ellis, 2002).

We now turn to the simulations that relate to AoA effects which have been presented by Zevin and Seidenberg (2002) and others.

### **Simulating Age of Acquisition Effects in Connectionist Networks: The Role of the Predictability of Input-Output Mappings**

There is a growing body of modeling literature which suggests that graded exposure or learning

of words results in distinctions in representational structure for early versus late acquired items. Steyvers and Tenenbaum (2005) considered early- versus late-acquisition of words in the vocabulary in terms of associations between words. They found that early-acquired words tended to have more associations with other words than did later-acquired words, and proposed this was because early-acquired words functioned as “hubs” via which other words are associated. Computational simulations of the construction of these associations showed how such networks can be built up gradually and iteratively, and how they consequently result in a greater number of connections to early-acquired nodes in the model (see also Li, Farkas, & MacWhinney, 2004; Richardson & Thomas, 2008). Lambon Ralph and Ehsan (2006) compared a backpropagation model’s learning of input-output mappings when patterns were presented either early or late in training. They varied the extent to which the mappings were systematic or arbitrary, and found that arbitrary mappings, such as those between object pictures and object names, demonstrated a large effect of AoA, whereas this was greatly reduced for systematic mappings, such as between written and spoken words. These computational studies illustrate that the locus of the AoA effect is plausibly present in the semantic representations of models (Brysbaert, Van Wijnendaele, & De Deyne, 2000b), or in the mappings between phonological and semantic representations (Lambon Ralph & Ehsan, 2006).

But is there evidence that for learning consistent, compositional mappings, such as between orthography and phonology, such order of exposure effects can also be observed? Anderson and Cottrell (2001), Smith, Cottrell, and Anderson (2001), and Lake and Cottrell (2005) trained neural networks to associate random input and output patterns using the backpropagation learning algorithm. They noted that some items were learned faster than others, and that when the network had reached asymptote, the items learned earliest also had the lowest error scores. Ellis and Lambon Ralph (2000) trained connectionist models using backpropagation with three layers of units that mapped random, but correlated, input-output patterns. They varied the point at which their model was exposed to patterns, to more closely simulate gradual, incremental learning. Some patterns

were presented at the start of training the network (the “early” patterns) while others (the “late” patterns) were introduced into training only after the network had spent time learning the early patterns. The network was then trained further on both the early and late patterns until performance reached asymptote. Analysis of the quality of the representations in the mature network showed an advantage for the early items. The effect of point of entry (AoA) could not be explained in terms of differences in cumulative training frequency for early and late items.

Ellis and Lambon Ralph (2000) argued that the neural network model demonstrated early plasticity of learning such that the first patterns that the model encountered had a greater impact on the structure of the mappings between inputs and outputs than later acquired patterns were able to exert. This arose partly because the rate of change of the weights on connections between units reduces as training proceeds. In addition, a network that has already learned a set of early items must then fit the representations of new, late patterns into the solution space that has already been generated to store the early patterns. That solution space is not necessarily optimal for representing the late items which may seek to reconfigure the network in the direction of a different solution. The attempts of the late items to re-structure the weight space are resisted by the early items which continue to be trained alongside them, thereby avoiding the problem of *catastrophic interference* where, if one set of items entirely replaces another in training, the first set may be overlaid and lost (McCloskey & Cohen, 1989; Ratcliff, 1990). Ellis and Lambon Ralph (2000) showed that the same network trained on the same patterns can show catastrophic interference (forgetting) when a late set entirely replaces an early set in training (simulation 2) but an abiding advantage for the early set when it continues to be trained alongside sets entered into training at later points (e.g., simulations 1, 3 and 5; see Richardson & Thomas, 2008, for convergent evidence).

In some circumstances, a network structure established in response to early items will be helpful to the assimilation and representation of later items. In other situations that structure will be unhelpful. The latter situation will happen particularly when the mappings between inputs and outputs are arbitrary and unpredictable, because under such circumstances the knowledge of input-

output mappings developed as a result of training on the early items will not assist the assimilation of the later items. For example, when the input-output mappings are regular and predictable, late items should be assimilated into the network with relative ease and should incur little in the way of processing cost thereafter. English spelling-to-sound mappings are partly but not wholly regular and predictable. Learning to read NAIL and FAIL will help you to read HAIL when you first encounter it, but learning to read HINT and MINT will not ensure that you will read the word PINT correctly. That is because the -INT family of words has an inconsistent pronunciation, with HINT and MINT receiving the more regular pronunciation while PINT is an exception. J. Monaghan and Ellis (2002a; 2002b) tested the prediction that words with exceptional (irregular) pronunciations should incur more of a cost to being late acquired than words with regular, consistent pronunciations. The word sets were matched on two measures of word frequency (Kucera & Francis, 1967, and Celex: Baayen et al., 1995); also on imageability, length and N. As predicted, regular / consistent words showed no cost to being late acquired but irregular / exception words did. Regression analyses by Ellis and J. Monaghan (2002) found an effect of AoA on word naming latencies for irregular / exception words but not for regular / consistent words.

*Zevin and Seidenberg's (2002) Simulations Involving Manipulations of Frequency Trajectory*

Zevin and Seidenberg (2002) acknowledged that AoA effects should occur when input-output mappings are arbitrary and unpredictable, as in naming objects, reading aloud Japanese kanji or comprehending words in English. But they took the view that the statistical regularities between orthography and phonology in English are such that there should be no cost to being late acquired, and the speed of translating between orthography and phonology should reflect only the cumulative frequency with which different words have been experienced. Zevin and Seidenberg (2002) reported a set of simulations using a computational model of reading based on that of Harm and Seidenberg (1999).

The training corpus for the Zevin and Seidenberg (2002) simulations consisted of 2,891

monosyllabic, monomorphemic English words. In their first simulation, 96.3% of the words were trained over 10 epochs of 100,000 trials with training frequencies derived from the *Wall Street Journal* (Marcus, Santorini, & Marcinkiewicz, 1993). Two additional, small sets of 54 words (3.7% of the total) were selected to form the “early” and “late” acquired items, which were controlled for cumulative frequency but differed in the point during training when they were most frequently presented to the model. All the early and late words incorporated consistent spelling-sound correspondences. The 54 “late”, or low-to-high, words were set for training at a low frequency of 1 for the first three epochs. Their frequency then rose over the next four epochs, reaching 1,000 at epoch 8. This high frequency was maintained for the last two epochs. The 54 “early”, or high-to-low, words were trained at a frequency of 1,000 for the first three epochs then declined progressively to a frequency of 1 for the last three epochs. The training frequencies were compressed by means of a square root transformation and items were sampled probabilistically. By the end of training the model produced appropriate outputs for 98% of the training set. There was no difference between performance on the early (high-to-low) and late (low-to-high) word sets on the model’s error, which is an index of the effectiveness of the mapping for that word.

Zevin and Seidenberg’s (2002) simulation 2 was broadly similar to simulation 1, except that the early and late sets were made up of 48 orthographically “strange” words like *hymn* and *choir* rather than the consistent words as in simulation 1. Once again, the error scores showed no difference by the end of training between early and late words. An additional simulation that trained only small sets of high-to-low and low-to-high frequency trajectory words, with no large set of background training patterns, demonstrated an “AoA effect” only when the word sets had little overlap in terms of their orthography or phonology.

A number of points can be made in relation to these simulations. The first is that when a child starts learning to read, the frequency of many late acquired words is 0, not 1. By giving their “late” words a frequency of 1 in the initial epochs, Zevin and Seidenberg (2002) allowed such words to exert an influence on the model’s solution to the mapping of orthography to phonology that they

would not have had with a more natural starting frequency of 0.

A second point is that most words that are acquired early in the course of reading do *not* subsequently decline in frequency; that is, they do not follow a high-to-low frequency trajectory. The WFG lists 5,273 words as occurring in reading material suitable for grade 1 readers with a frequency of 2 per million or more. The correlation between grade 1 frequency and the frequency of the same words in grade 13+ (College grade) reading material is .812. That is, words tend to maintain their relative frequencies from childhood into adulthood.

A third point to be made in relation to the Zevin and Seidenberg (2002) simulations is that there is no natural counterpart to the background words which constituted over 95% of the training items in their simulations 1 and 2, which will have obscured possible AoA effects for the small training sets, even for “strange” words like those in simulation 2. This is because, as Zevin and Seidenberg note, even strange words are not entirely unsupported by other words in the language. To cite Zevin and Seidenberg’s example, BEIGE is not pronounced “glorp”. BEIGE overlaps with BINGE, BARGE, WEIGH and other more distant neighbors among the background stimuli (Zevin & Seidenberg, 2002, p. 17). The reconfiguration that a network might require in order to learn BEIGE as a late acquired word will be reduced by the presence of words like BINGE, BARGE and WEIGH among the background words. This is reflected in that fact that in Zevin and Seidenberg’s (2002) simulation 2, fewer trials were needed to learn late strange words than early strange words, because the background words that had been learned before the late strange words were introduced assisted the learning of those words. A psychologically plausible developmental model of reading cannot include huge numbers of background items trained from the outset, as this runs quite contrary to the developing reader’s experience.

### **Developmental and Non-Developmental Models of Reading**

These reflections led us to conclude that it is premature to claim that AoA effects in word naming are artefactual, or that they are incompatible with existing computational models of reading.

We considered it still possible that AoA effects might emerge if a computational model was trained in a manner that respected and reflected the way that reading vocabularies are actually acquired. The simulations presented below are all based on Harm and Seidenberg's (1999) computational model of reading, just trained and/or analysed in a different way.

We compare the behavior of three models which differ in the way that words were entered into training. We use these models to accomplish two goals. First, we demonstrate that age of acquisition effects are a consequence of dynamic learning systems that are exposed incrementally to information they are required to acquire. Thus, age of acquisition is to be expected as an emergent feature of models of naturalistic, gradual vocabulary learning. Second, we demonstrate that a model of reading trained on realistic, age-appropriate reading materials reflects detailed behavioural data on age of acquisition effects in reading studies. Hence, a model trained incrementally can simulate a broader range of reading phenomena than previous computational models of reading that do not take reading development into account. In summary, we show that AoA is a general principle of learning, and that incorporating this fact into a model of reading provides a computational explanation for AoA effects in word naming studies.

The first and principal model we examine is termed the *Developmental Model*. This model was trained cumulatively on a realistic corpus of age-appropriate reading materials. This approach to training acknowledges the force of Harm and Seidenberg's (1999, p. 503) comment that although past computational models have based their training on adult word frequencies, "Reading instruction is quite different: Children initially learn to read small vocabularies that expand over time." Harm and Seidenberg accepted that their own work was subject to this shortcoming but added that "there is nothing in our approach that precludes structuring the training procedure in more realistic ways". That is precisely what our Developmental Model set out to do, entering words into training in accordance with the reading ages at which they first appear in the *Educator's Word Frequency Guide* (WFG: Zeno, Ivens, Hillard, & Devurri, 1995) and modifying their subsequent training frequencies in accordance with the frequency trajectories that the words follow across

reading ages in the WFG. We refer to this objective measure of age of presentation of words in age-appropriate reading materials as WFG-AoA.

For the Developmental Model we establish that it demonstrates standard properties required for computational models of reading: accurate reading of words; adequate generalisation to nonwords; and effects of word frequency and spelling-sound consistency with a frequency by consistency interaction (Andrews, 1992; Brown & Watson, 1994; Hino & Lupker, 2000; J. Monaghan & Ellis, 2002a; Seidenberg, Waters, Barnes, & Tanenhaus, 1984; Waters & Seidenberg, 1985). We also examine the role of WFG-AoA on the model's ability to read words, predicting that WFG-AoA effects will be observed as a predictor of the Developmental Model's reading responses.

However, it is possible that observing effects of WFG-AoA in the Developmental Model could be a consequence not of the point at which the model was first exposed to the word, but rather of some other properties of early-acquired words making them easier to learn, or more resistant to interference from other items. In order to test these possibilities, we analyse the performance of two additional models. The first we term the *Cumulative Frequency Model*, where words were presented to the model according to the same cumulative frequency as the words used for the Developmental Model, except that sampling of the words according to frequency was not staged during training. This model is essentially a replication of the Harm and Seidenberg (1999) model of reading. An important point to note, however, is that a conventional model of this sort is not immune from potential effects of order of presentation of words over and above the simple effects of frequency. All connectionist models of reading that aim to represent the child's learning environment, including those which train all items together from the outset, are trained in an incremental fashion in that words are presented to the network one at a time. High frequency words are thus more likely to be sampled during the earliest stages of training than low frequency words, and are therefore likely to benefit more from any advantages associated with early entry into training. Apparent effects of frequency could reflect a combination of genuine frequency effects (better representations for words trained more often) and order of presentation effects arising from the earlier entry of high

than low frequency words into training. Conventional models in which all words are ostensibly entered into training together would then contain hidden but potentially significant “age of acquisition” effects, at least in models where the weights are updated after each word’s presentation (see Sibley, Kello, & Seidenberg, 2010, for an exception to this training regime). In order to assess the precise contribution of order of presentation, for each simulation, we report analyses in which we monitored the actual order in which words were presented to the Cumulative Frequency Model (i.e., the precise point at which the word was entered training). We term this Point of Entry (PoE). If PoE can be shown to be related to the Cumulative Frequency Model’s performance, that would indicate that presentation order influences model performance more broadly than in the particular training regime of the Developmental Model.

In our analyses we are careful to distinguish the relative contributions of these effects in order to test the relative contribution of frequency, PoE, as well as WFG-AoA, in explaining the models’ performance. We predicted that the Cumulative Frequency Model would demonstrate effects of frequency, and PoE, which would support our hypothesis that incremental exposure influences learning in dynamic models. We predicted that WFG-AoA would not explain additional variance in the model’s response once frequency and PoE have been taken into account, which tests whether WFG-AoA is a surrogate variable for other factors influencing the model’s performance (e.g., uncontrolled factors that influence the ease of forming orthography to phonology mappings).

In Cortese and Khanna’s (2007) set of 2,342 words, the correlation between frequency and adult AoA was  $-.689$ . The correlation between AoA and cumulative frequency would necessarily have been even higher (because cumulative frequency reflects a combination of how frequent a word is and how long it has been known). In the present Developmental Model, the correlation between AoA and (cumulative) frequency is  $-.834$ . This means that it would be impossible to manipulate frequency when training a model on a representative sample of English words without at the same time introducing potential AoA effects into the model’s performance (because of the likelihood that high frequency words will be sampled earlier than low frequency words). To break this link

between frequency and AoA effects in the model's performance, we tested an additional non-developmental model, which we refer to as the *Random Frequency Model*. This model was trained and tested in the same way as for the Cumulative Frequency Model, except that the frequencies of words were randomly reassigned before training. This model enabled us to test whether the high correlation between frequency and WFG-AoA was causing the apparent effects of order of presentation in the Developmental and the Cumulative Frequency Models. We predicted that (re-assigned) frequency and PoE would remain as significant predictors of the Random Frequency Model's performance, but we predicted that frequency of words used for the other models and WFG-AoA would not be significant predictors,.

### *Network architecture*

The three models used the same basic architecture which was based on Harm and Seidenberg's (1999) computational model of reading. A schematic diagram of the network is shown in Figure 1. The network had an orthographic input layer that was connected via a set of hidden units to an output layer from which the word's phonological representation was to be produced. Each unit in the output layer was connected to a layer of 25 "clean-up" units that connected back to the output layer. This meant that the output layer acted as a set of phonological attractor units, designed to learn patterns between phonemes in the output layer and to maintain stable phonological representations of words (see Harm & Seidenberg, 1999, 2004, for more details).

The input layer was composed of 10 letter slots, each of which comprised 26 units, one for each letter of the alphabet. Words were presented with the first vowel letter in the word at slot 4, a second vowel if there was one in slot 5, and the rest of the word in adjacent slots to the left and right of this position. In Figure 1, the word TRUTH is presented with the U in the fourth slot, as it was the first vowel in the word. The letters T and R immediately precede the U, while the letters T and H follow the U and the second (empty) vowel slot. A word like TROUGH would utilize both of the vowel slots in positions 4 and 5. Only monosyllabic words were presented, so only one pair of

adjacent vowel slots was required.

The hidden layer contained 100 units connected to all the input units and all the output units. The output layer consisted of 8 phoneme slots, 3 slots for the word onset, 1 slot for the vowel – diphthongs were encoded in a single vowel slot – and 4 slots for the coda. Within each phoneme slot were 25 units representing phonological features derived from Chomsky and Halle (1968). These were the same features as employed by Harm and Seidenberg (2004). A particular phoneme in a particular position was therefore represented in terms of activating the appropriate subset of feature units. In Figure 1, at output /t/ and /r/ are the consonants in the onset slots for the spoken word “truth”, /u:/ is the only vowel in the nucleus slots, while /θ/ (voiceless “th”) is the only consonant in the coda slots.

### *Training set*

The training set employed in the simulations comprised 6,229 monosyllabic words that had phonological representations in CELEX (Baayen et al., 1995) and semantic representations derivable from the WordNet database (Miller, 1990). Our set included the 6,103 words used by Harm and Seidenberg (2004) with the addition of a subset of 126 words that were omitted from Harm and Seidenberg’s (2004) model because only one inflected form of each word was included in their set whereas we included all inflections (so, NAME, NAMES, and NAMED were all included in our set). This set of 6,229 words was cross-indexed with the 4,114 monosyllabic words that occurred in Section I of the *Educator’s Word Frequency Guide* (WFG: Zeno et al., 1995). The WFG is based on a large corpus of words extracted from 60,527 samples from 16,333 written texts. The samples contained 154,941 different words of widely varying lengths and frequencies, and 17,274,580 words in total. The texts in WFG were graded using readability measures as being appropriate for readers at 13 different grade levels, covering the age range 5 to 18 in the American and British schooling systems. Section I of the WFG contains 19,468 words whose frequency, weighted across topics and grade levels, was at least 1 per million. For those words, Section I gives

the frequency of each word at each grade level, and it is considered the best frequency count currently available for age-appropriate vocabulary (Balota, Cortese, Sergeant-Marshall, Spieler, & Yap, 2004; Zevin & Seidenberg, 2002; 2004).

### *Training the Models*

*Training the Developmental Model.* Words were entered into training at a point representing the grade at which they first occur in the WFG with a frequency greater than a certain threshold value. That threshold was reduced with each grade to admit progressively more words into the model, and to reflect the fact that older children read more words per unit time than do younger readers, and are therefore more likely to encounter lower frequency words in the course of their reading. The threshold used is shown in Table 1 along with the number of words in the training set at each stage in the model. So, the words on which the model was trained in epoch 1 were only those words that occur in grade 1 of the WFG with a frequency of 1,000 per million or more. Additional words were added to the training set in epoch 2 if they achieved the threshold of 100 occurrences per million in that epoch but had not done so in epoch 1, and so on.

The Developmental Model had 13 initial cycles or epochs of training that corresponded to the first 13 grades in the WFG. A 14th cycle represented additional adult reading experience and gave the model an opportunity to learn the whole set of 6,229 words, some of which (late acquired, low frequency words) may not have appeared by cycle 13. The stage at which the word was first entered into training was our objective measure of AoA for the model (WFG-AoA).

Words were presented to the Developmental Model according to their log-compressed frequency for each reading age. Log compression was used in order to increase the probability that the model was exposed to each word during the training stage (for more detailed justification, see Harm & Seidenberg, 1999, and Plaut et al., 1996). Log-compressed frequencies less than .05 were increased to .05, so the frequencies of words varied from 1 for the most frequent word in each training stage to .05 for the least frequently occurring word in that training stage. All 6,229 words were entered

into the training set for stage 14, sampled according to their frequency, to ensure that the model was exposed to all the words during training.

To simulate the child's exposure to quantity of words with age, we progressively increased the number of word presentations (i.e. word tokens). For training stages 1 and 2, 50,000 word presentations were included, for stages 3 to 5 there were 100,000 presentations, while stages 6 to 13 involved 200,000 word presentations. The model was then trained on the full set of 6,229 words for a further 3 million word presentations. This meant that in some respects our test of the influence of early- versus late-acquired words in the Developmental Model was made under relatively weak conditions: by the time the model was tested at the end of training, it had been trained on late-acquired words for more than half the total training time. If there was an effect of WFG-AoA under these conditions, that suggests that the findings are robust and not due to particular parameters of the relative difference in training time on early and late acquired words. We have noted above that AoA effects in humans appear not to reduce over the adult lifespan (Barry et al., 2006; Morrison et al., 2002).

The model was initialised by setting weights on all connections to a random value in the range (-1,1). Initial activation of units was set to .5. The model was trained to produce the phonological form of the word given its input, across 7 time steps. At time 1, the orthographic representation of the word was clamped to the input units. At time 2, activation had spread from the input units to the hidden layer. At time 3, activation reached the output layer. For a further 4 time steps, activation cycled in the model and the error at the output layer was computed. At the end of the 7th time step, cross-entropy error was propagated back through the network and connection weights were adjusted according to the backpropagation learning algorithm (Rumelhart, Hinton, & Williams, 1986). Connections between all layers were trained with a learning rate of .01.

Ten complete runs of the model were performed with different randomisations of the weights and different random samplings from the words at each training stage.

*Training the Cumulative Frequency Model.* For the Cumulative Frequency Model, the cumulative frequency of words was determined from the Developmental Model, and words were sampled according to this frequency from the beginning of training. Hence, by the end of training, words had been presented with the same overall frequency as for the Developmental Model, but without incorporating the precise staging according to WFG-AoA. Again, ten simulation runs were performed, with randomised starting weights and random sampling of words according to their cumulative frequency.

*Training the Random Frequency Model.* As noted above, measures of cumulative frequency and AoA are highly correlated, and there remains the possibility that effects of order of presentation relating to the models' responses may be an accidental property of confounds with psycholinguistic properties – some unaccounted for property of high-frequency or early-acquired words that is not due to differences in presentation order. For the Random Frequency Model, the frequencies of words in the Cumulative Frequency Model were randomly reassigned to the set of words in the training set, disconfounding WFG-AoA as well as frequency from any possible contributions of other variables. Words were then sampled according to their reassigned frequencies, and the model was trained with the same parameters as for the other models. This was done separately for each of the 10 runs of the model. We will present the results for the Random Frequency Model separately for each run whereas the results for the Developmental and Cumulative Frequency Models will be averaged over the 10 runs (because for those models the key parameters for each run were the same).

#### *Procedure for Testing the Models*

After training, the performance of the three was assessed in two ways. First, we measured reading accuracy by determining for each phoneme slot in the model's output the nearest target phoneme from the set of all phonemes in the language. If the model's production was closest to the

target phoneme at all phoneme slots, it was judged to have pronounced a word (or nonword - see below) correctly. Second, we determined how close the model's output production was for each word by measuring the Euclidean distance between the actual output of the model and the target phonology for each word. Euclidean distance is determined by squaring the difference between the actual output the model produces and the target output for each unit, then summing the squared differences, and taking the square root of this sum. High values indicate that, overall, the model's production was distinct from the target output while low values mean the model was producing an output close to the target. Such measures have been related to response times and accuracies in behavioural studies of naming (e.g., Harm & Seidenberg, 1999, 2004; Plaut et al., 1996), as they indicate the relative ease or difficulty for the model of generating a pronunciation for each word. We favored this measure of determining the closeness of the model's production because it resulted in discrimination between responses for words, whereas measurements of cross-entropy error, for instance, were close to zero for all patterns. For the results of factorial studies of particular word sets in the analyses reported below, each run of the model was analysed as a separate subject. The error scores for each word used in the regression analyses are the mean of the 10 Euclidean distance scores for that word in the 10 simulation runs of the Developmental and the Cumulative Frequency Models, but the regressions are performed on each simulation run for the Random Frequency Model.

During training of each of the three models, we analyzed the order of presentation of words in training to determine the point of exposure (PoE) of each word in the model's training. The PoE of a word for a particular simulation was established in terms of when each word had been presented 5 times in total.<sup>2</sup> The Developmental Model underwent 14 epochs of training with the first epoch involving 50,000 presentations, the second involving a further 50,000 presentations, and so on (see Table 1). For that model, a word was given a PoE value between 1 and 14 corresponding to the epoch in which the word received its fifth training exposure. Thus, if a word was presented 5 times

---

<sup>2</sup> The analyses reported below that use PoE values based on determining when a word had been presented 5 times in training were repeated using alternative PoE values based on the point when a word had been presented 1, 2 or 10

during the first epoch, it was given a PoE value of 1. If it reached its 5th presentation during the second, it was given a PoE of 2, etc. If the word had still not been presented 5 times before the end of the final epoch of training, then PoE was set at 14.

The same principle was used to determine the PoE values for words in each run of the Cumulative Frequency and the Random Frequency Models. A word received a PoE value of 1 if it was sampled 5 times within the first 50,000 presentations, a value of 2 if it was sampled for the 5th time between presentations 50,001 and 100,000, and so on. If the word had again not been presented 5 times before the end of training, PoE was set at 14. Note that although all words are available to be sampled from the start of training by the Cumulative Frequency and Random Frequency Models, the probability of word being sampled at a given moment is determined by its frequency. The higher a word's frequency, the greater the probability that it will be sampled at any given point in training, and the greater the probability that it will have been sampled 5 times before the end of earlier epochs. Higher frequency words therefore dominate the early stages of training even in models where all words are available for selection from the outset.

### Analysis of the Models' Performance

The analysis of the models is presented in three stages. First, we report the effect of order of presentation of words (PoE) on the models' learning of the mappings between orthography and phonology to determine whether, as predicted, this accounts for variance in a dynamic, incrementally-trained model of reading. We test the extent to which WFG-AoA predicts variance in the models' performance before and after PoE has been partialled out. This enables us to determine whether the manipulation of WFG-AoA is effective for the Developmental Model, and also whether WFG-AoA may be a surrogate variable for other unknown effects that were not controlled in the modeling. We then establish that the key model, the Developmental Model, predicts standard hallmark effects of word naming, as simulated in previous computational models of reading.

---

times. The results were unchanged. We will therefore report the results based only on PoE<sub>5</sub>.

Finally, we present the Developmental Model's ability to reflect the detailed data on age of acquisition in behavioral studies of word processing.

### *Effects of Order of Presentation on the Models' Performance*

The first major goal of the modelling was to assess the extent to which the order in which a dynamic learning model is exposed to patterns has an effect on its learning. To test this, we performed regression analyses of each of the models for all words in the training set, first partialling out effects of psycholinguistic variables on the models' performance, then testing directly the effect of PoE and WFG-AoA against the residuals of the regression analysis. Frequency, word length, neighborhood size, and consistency are factors that are known to affect word naming latencies, as well as reading models' performance, and we wanted to examine the effect of order of presentation once these additional variables had been accounted for in the models' performance.

The Frequency measure in the model was the cumulative frequency of the model's exposure to each word by the end of training, and is thus distinct from the (final) frequency measure entered into the Cortese and Khanna (2007) analysis. Number of orthographic neighbors (N) was determined by counting the number of words of the same length in the model's training set that differed by only a single letter (Coltheart, Davelaar, Jonasson, & Besner, 1977), though we note that there are other measures of orthographic similarity that could be used (see e.g., Yarkoni, Balota, & Yap, 2008). Length was the number of input letter slots that were active for each word. For spelling-sound Consistency, we computed a continuous variable based on the proportion of words with the same orthographic rime that were pronounced in the same way as the target word. This is the measure typically employed in the analysis of distributed models of reading like Harm and Seidenberg (1999) in determining spelling-sound consistency. Thus, the rime -UMP is deemed to have a consistent pronunciation because all words ending in -UMP are pronounced alike (BUMP, DUMP, LUMP, etc), and so all of those words have a consistency of 1. The rime -INT has an inconsistent pronunciation. Most words ending in -INT are pronounced as in HINT, MINT and

STINT: those words have a high value of consistency (.92, indicating that 92% of the words with the rime -INT are pronounced in this way). PINT has a different and, in this case, unique pronunciation and has a low consistency value (.08, because the remaining 92% of the words ending -INT in the model's training set have a different pronunciation of the rime). AoA in the model was defined as the first stage in training in which the model was exposed to the word (see Table 1), and is in the range [1, 14].

Table 2 shows the mean values of the five predictor variables and the dependent variables (the Euclidean distance scores of the three models) along with their standard deviations and minimum and maximum values. The similar mean and standard deviations for the Euclidean distance scores for the models indicate that the quantity of error in the three models' productions is comparable, though the range is greater in the Random Frequency Model.

Table 3 shows the correlations among the predictor variables and their correlations with the dependent variables from the models. With 6,229 items, even very small correlations are statistically significant. The psycholinguistic variables show a similar pattern of intercorrelations to previous studies involving human participants (e.g., Balota et al., 2004; Cortese & Khanna, 2007; Morrison & Ellis, 2000). Length and N are highly negatively correlated, which reflects the fact that shorter words generally have more neighbors than longer words. Length and Frequency are also highly negative correlated (common words tend to be shorter than longer words [Zipf's Law]). WFG-AoA is positively correlated with Length and negatively correlated with N and with Frequency (early-acquired words tend to be short, with many neighbours, and of high frequency). The correlation between WFG-AoA and Frequency confirms the importance of determining whether, once cumulative frequency has been entered into the regression analysis, WFG-AoA can still account for unique variance in the Developmental Model's performance.

There was a significant correlation between each of the psycholinguistic variables and the error in both the Developmental and the Cumulative Frequency Models, and these were significant for the majority of the Random Frequency simulation runs, with the exception of frequency (as

frequency was randomised for this model). All these correlations were in the anticipated direction, and are in line with behavioral data (e.g., Balota et al., 2005). Critically, there was a significant correlation between WFG-AoA and the error for the Developmental and the Cumulative Frequency Models. This may be due to the order in which words are presented to these models (the high correlation between frequency and WFG-AoA means that early-acquired words are presented early to both models), or it could indicate that WFG-AoA is a surrogate variable for other, uncontrolled factors. Importantly however, the correlations in the Random Frequency Model show that there was no systematic relationship between a reading model's error and WFG-AoA: early-acquired words are not inherently easier to learn, because when the link between WFG-AoA and presentation order to the model is broken, as in this model, there was just one of ten simulations that showed a significant relationship between reading difficulty and WFG-AoA. Hence, apparent WFG-AoA effects in both the Developmental and the Cumulative Frequency Models are most likely to be due to presentation order effects, with the effects in the Cumulative Frequency Model occurring because higher frequency words tend also to be early acquired according to WFG-AoA and enter early into the training of the model because of the frequency-based sampling effect discussed above.

The residuals analyses, shown in Table 4, provide a direct indication of the extent to which order of presentation of words to the model influenced its reading performance. We first performed a regression analysis including a set of variables, and then determined whether the residuals from the analysis significantly related to  $PoE_n$  for each model and to WFG-AoA. The first regression partialled out the effects of psycholinguistic variables – cumulative frequency, consistency, length, and N – and the residuals were then related to WFG-AoA. We report more detailed analyses of these individual effects below when we relate the performance of the Developmental Model to detailed behavioral data. Then, we also partialled out the actual order of presentation of words to the model (PoE) and related the residuals to WFG-AoA. We predicted that WFG-AoA would be significantly related to both the Developmental and to the Cumulative Frequency Model, but that if the staged training of the Developmental Model was effective then this would result in a higher

correlation between the residuals and WFG-AoA for the Developmental Model. We also predicted that, if the staged training of the Developmental Model was effective then there would be no relationship between the residuals of the regression with psycholinguistic variables and PoE, but that there would remain a small relationship for the Cumulative Frequency Model. We also included analyses for the Random Frequency Model. We predicted that WFG-AoA would not relate to the residuals for this model, unless it is a surrogate variable for ease of learning unrelated to order of presentation of words to the model.

The results supported our hypotheses. For all the models PoE was significantly related to the residuals, after cumulative frequency had been partialled out of the Models' performance. This shows that the order of presentation had an effect on all of the models. For the Developmental Model, WFG-AoA was significantly related to the residuals of the model's performance after the psycholinguistic variables had been taken into consideration. Furthermore, this relationship reduced to zero when the actual order of presentation (PoE) had been taken into consideration. Hence, WFG-AoA was effectively operationalised in the modelling environment as reflected by PoE, and did not have an effect in the model other than as an effect of order of presentation.

For the Cumulative Frequency Model, WFG-AoA also related significantly to the residuals when the psycholinguistic variables had been taken into account. However, this relationship reduced to a small and non-significant relationship when PoE was also entered into the regression analysis. The difference in the  $\beta$  values for the Developmental and the Cumulative Frequency Models was not significant,  $\beta = .013$ ,  $p = .58$ , indicating that the Cumulative Frequency Model closely approximated the effects of staged input to the model, but did not perfectly reflect the effects of order of learning in human reading development, as measured by the WFG database.

The analysis of the residuals for the Random Frequency Model confirmed that WFG-AoA was not a surrogate variable for ease of learning orthography to phonology mappings, and that it was related in the models' performance to order of presentation. Indeed, there was a tendency for the residuals to be negatively related to WFG-AoA for the Random Frequency Model, clearly

demonstrating that the words that occur early in natural reading exposure are not easier to acquire when order of presentation is controlled.

These analyses have therefore established that order of presentation has a profound impact on dynamic models learning to read when presented with words incrementally. The Developmental Model most closely approximates the actual environment of a child learning to read, and for this model WFG-AoA was, as intended, a very close approximation to the actual order of presentation of words to the model. The next analyses test the extent to which the three models relate to psycholinguistic measures of reading, and then we extend previous models of reading to address issues of AoA effects in the Developmental Model.

#### *Hallmark Effects of Word Reading in the Three Models*

*Word and nonword reading.* All three models were tested at the end of 5 million words of training for the number of words correctly converted from orthographic input to phonological output in terms of accuracy at every phoneme slot. To assess generalisation to new, untrained items, the models were also tested on their accuracy of reading 240 pronounceable nonwords. The nonwords comprised 86 pseudohomophones from Glushko (1979) together with 160 pseudowords from McCann and Besner (1987), though 4 were omitted from the latter source because they contained letter combinations that never occurred in the training set (BINJE, FAIJE, JINJE, and WAIJE). Two further nonwords were duplicated between the Glushko (1979) and the McCann and Besner (1987) lists.

The Developmental Model's performance during training is shown in Figure 2A for both words and nonwords. The x-axis indicates the number of word tokens presented to any given point during training, while the y-axis indicates the proportion of all words or nonwords pronounced correctly at each point. Each of the 10 runs of the Developmental Model correctly produced 100% of the words in the training set correctly by the end of training (Figure 2A). At the end of each stage (epoch) of training, the Developmental Model read accurately all the words that had been presented in that

stage. Accuracy of word reading for age-appropriate words at each stage was > 99.6% for all stages. Errors made to words affected almost exclusively words that had not been entered into training by that epoch. The Developmental Model also generalised well to nonwords, pronouncing a mean of 84.6% (SD = 2.5) of the nonwords correctly across the 10 runs by the end of training (Figure 2A). At the same training stage, Harm and Seidenberg's (1999) original model correctly read between 75% and 80% of a similar set of untrained nonwords.

The Cumulative Frequency and Random Frequency Models also learned to pronounce words (both 100% accuracy) and nonwords (85% and 87%, respectively) after 5 million patterns (see Figures 2B and 2C). We note that the performance of those models on the nonwords had already asymptoted by the end of epoch 1 (50,000 training trials) and did not substantially improve further as training continued. This suggests that the models' knowledge of sublexical letter-sound correspondences was complete (or as good as it would be) by the end of epoch 1. In contrast, performance on the training set of words continued to improve from the end of epoch 1 through to around epoch 6. Improvement specific to trained words may reflect consolidation of larger grain size mappings between orthography and phonology. Plaut et al. (1996) investigated the development of non-componential "attractors" within recurrent networks that mediate the pronunciation of exception words. The late improvement for words in Cumulative Frequency and Random Frequency Models may indicate the establishment of non-componential representations for exception words.

*The frequency by consistency interaction.* Harm and Seidenberg's (1999) original simulation demonstrated the frequency by consistency interaction that has been reported in many studies of human word naming (e.g., J. Monaghan & Ellis, 2002a; Paap & Noel, 1991; Taraban & McClelland, 1987). To test the frequency by consistency interaction in the models, we used the set of words from Experiment 1 of Taraban and McClelland (1987). These words varied in terms of whether the word had neighbors with similar pronunciations (consistent) or distinct pronunciations

(inconsistent), and were also divided into sets of high and low frequency words. There were 24 high frequency inconsistent words, 24 high frequency consistent control words, and 24 low-frequency inconsistent words with 24 controls. One high-frequency consistent word (MAIN) and one low-frequency consistent word (SANK) were not in the Harm and Seidenberg (2004) training set on which the current simulations were based, so these words were omitted from the analyses. For each model an ANOVA was conducted on the Euclidean distance of the model's production for each word at the end of training, with each of the 10 runs of the model as subjects in the analysis, and with Frequency and Consistency as factors.

The Developmental Model demonstrated significant effects of Frequency and Consistency, and a significant interaction between the two: Frequency,  $F(1, 9) = 50.49$ ,  $MSE = .0003$ ,  $p < .001$ ; Consistency,  $F(1, 9) = 75.15$ ,  $MSE = .00003$ ,  $p < .001$ ; Frequency x Consistency,  $F(1, 9) = 31.68$ ,  $MSE = .0001$ ,  $p < .001$ , simulating the behavioral results effectively (see Figure 3).

The Cumulative Frequency Model also demonstrated significant main effects of Frequency,  $F(1, 9) = 53.47$ ,  $MSE < .0001$ ,  $p < .001$ , and Consistency,  $F(1, 9) = 19.30$ ,  $MSE < .0001$ ,  $p < .001$ , along with a significant interaction between Frequency and Consistency,  $F(1, 9) = 31.03$ ,  $MSE < .0001$ ,  $p < .001$ , indicating that the simulation that was intended to replicate Harm and Seidenberg's (1999) model did so on hallmark effects of word naming.

For the Random Frequency Model, there was a large effect of Consistency, as anticipated,  $F(1, 9) = 35.57$ ,  $MSE < .0001$ ,  $p < .001$ , but the effect of Frequency was only marginally significant (and was in the opposite direction to the effect for the other simulations), as frequency varied across simulations and did not tally with the frequency distinction in the experiment,  $F(1, 9) = 4.92$ ,  $MSE < .001$ ,  $p = .054$ . There was no significant Frequency by Consistency interaction,  $F < 1$ .

#### *Further Analysis of the Effects of Age of Acquisition Effects in the Developmental Model*

*The AoA by consistency interaction.* J. Monaghan and Ellis (2002a; 2002b) reported an AoA by consistency interaction using sets of words matched on two measures of word frequency. Seven of

the 80 words used by J. Monaghan and Ellis (2002a; b) are not listed in the WFG (WHACK, GIG, BRAWL, WHOOP, SWAT, GHOUL and BROOCH). There were no significant differences in frequency, length, or neighborhood size between the remaining early- and late-acquired words calculated from the training set used for the Developmental Model, all  $t$ 's  $< 1$ ,  $p \geq .5$ . The early and late word sets differed significantly, however, on point of entry into the WFG; in other words, on reading age,  $t(72) = -2.00$ ,  $p < .05$ . The Euclidean distance for the remaining words in the Developmental Model were analysed. An ANOVA conducted on the Euclidean distance for these words at the end of training showed significant main effects of AoA,  $F(1, 9) = 5.94$ ,  $MSE = .001$ ,  $p < .05$ , and of Consistency,  $F(1, 9) = 5.24$ ,  $MSE = .001$ ,  $p < .05$ , and a significant interaction between AoA and Consistency,  $F(1, 9) = 14.37$ ,  $MSE = .002$ ,  $p < .005$ . The interaction is shown in Figure 4, alongside the behavioral naming response time data from Experiment 2 of J. Monaghan and Ellis (2002a). The difference in the Euclidean distance values for early and late acquired words in the Developmental Model is greater for inconsistent words ( $.038 - .018 = .020$ ) than for consistent words ( $.015 - .020 = -.005$ ), a similar effect to that shown in the RT data of J. Monaghan and Ellis (2002a; 2002b).

The interaction between AoA and Consistency observed by J. Monaghan and Ellis (2002a; 2002b) thus depends on entering the late words into training after a human or artificial reading network has already been exposed to early words. The AoA x Consistency interaction also supports the claim that late acquisition does not impact equally on all words. If a word contains consistent, predictable letter-sound correspondences, there is little or no cost to learning it late rather than early for the orthography to phonology pathway. This is because late-acquired, consistent words can exploit the knowledge of letter-sound mappings built up by the network in response to earlier words. A word containing inconsistent, unpredictable letter-sound correspondences can also be learned well if it is encountered early and can therefore influence the orthography-to-phonology mappings while they are still pliant and malleable (see below). But if an inconsistent word is only encountered late in training (in the network or in human development) it struggles to achieve an

effective representation because doing that requires reconfiguration of a network which, by that point, is well established and resists re-structuring.

*Regression analyses of model behavior.* Mega-studies of reading provide possibilities for comparing the relative contributions of different psycholinguistic variables to explaining variance in naming times, and consequently can test whether variables found to distinguish responses in small sets of words can generalise to a larger vocabulary (though see Sibley, Kello, and Seidenberg, 2009, for a caveat on comparing regression analyses of large sets of naming time responses to studies on controlled word sets). In Cortese and Khanna's (2007) hierarchical regression analysis on naming response times for 2,342 monosyllabic words (see also Balota et al., 2004), the first step entered manner features of the onset of the word. Length, frequency, neighbourhood size and consistency variables were entered in the second step. AoA and imageability were entered in the third and final step to determine whether they contributed to variance after the effects of all the other variables had been accounted for.

In order to determine whether the effect of AoA generalised to the whole training set for the Developmental Model's performance, and also to confirm that the interactions between frequency and consistency, and between AoA and consistency, could also be seen across the whole word set, we replicated the regression analysis of Cortese and Khanna (2007) with modifications appropriate to the different conditions of the model's training. As the dependent variable, we entered the Euclidean distance of the model's production for each word, averaged across the 10 simulation runs of the model. As the independent variables, we entered a set of psycholinguistic variables. We did not use phonological features as these did not contribute significantly to the model's responses as they do for speech responses. With human naming times, onset of voicing varies according to the type of phoneme onset, which was not relevant to Euclidean distance between actual and target output in the model. For the model hierarchical regression analysis, at the first step we entered Length, orthographic Neighborhood size, cumulative Frequency, and Consistency. At the second

step, we entered WFG-AoA. At the third step, we entered interacting variables: Frequency x Consistency, WFG-AoA x Consistency, and WFG-AoA x Frequency x Consistency.

The hierarchical regression analysis results for the Developmental Model are shown in Table 5. As previously reported in the residuals analysis, in the first step, Frequency, Consistency, Length, and N all made significant contributions to accounting for variance in the model's error. High-frequency words, short words with many neighbors, and words with consistent spelling to sound mappings resulted in lower error in the model's production, and reflected the standard pattern of psycholinguistic influences found for response times in mega-studies of single word naming (Balota et al., 2004). At Step 2, WFG-AoA was found to account for independent variance after the other psycholinguistic variables had been entered: words to which the model was exposed earlier in training resulted in lower error in the model's production.

For Step 3, the interaction between Frequency and Consistency (testing the Taraban and McClelland, 1987, effect on a larger set of words) was not significant, though this was found to be significant when the three-way interaction term (Consistency x Frequency x WFG-AoA) was excluded. In the model, the interaction between Frequency and Consistency, then, seems to be crossed with WFG-AoA. The Developmental Model therefore predicts that the interaction between Frequency and Consistency will be strongest for late-acquired words (low-frequency late-acquired words with inconsistent spelling-sound correspondences have highest error). For the early-acquired words, inconsistent low-frequency words had less of an effect on error in the model. In Step 3, the interaction between WFG-AoA and Consistency was significant, indicating that the J. Monaghan and Ellis (2002a; 2002b) findings on subsets of carefully controlled words extended to the vocabulary at large in terms of the Developmental Model's performance.

*Frequency trajectory.* The term “frequency trajectory” refers to a word's relative change in frequency across the grades. Late acquired words inevitably have a positive frequency trajectory because they are represented at later grades but not at the youngest age levels. Why, then, did Zevin

and Seidenberg's (2002) simulation not show an effect of frequency trajectory, even though the high-to-low frequency words had early AoA, and the low-to-high words had late AoA, and our Developmental Model revealed such an effect of AoA? One possibility is because the low-to-high frequency words had frequency of 1 early in training, whereas the high-to-low frequency words had extremely low frequency in later stages. This will have reduced the possibility of finding an AoA effect in their model. The former effect would have meant that some of the high-to-low words will have had early AoA depending on the sampling of the word set. The latter effect will have meant that the potential advantage of early acquired words would have diminished because this will have been offset by forgetting in the model. In Zevin and Seidenberg's (2002; 2004) simulations, early acquired words were given negative (high to low) frequency trajectories, but in reality this only occurs in a small subset of words like CAT and KITE that occur more often in children's than adult reading material.

To test whether this forgetting would have had an effect on the Developmental Model's performance, as a potential explanation of one of the differences between Zevin and Seidenberg's (2002) model and the present Developmental Model, we tested whether frequency trajectory contributed additional variance to the regression analysis of the Developmental Model's performance. If it does, then this indicates that forgetting does have an impact on word naming performance in the model. If it does not, then forgetting of early AoA words that reduce in frequency later in training could not be the reason for the discrepancy between our model and Zevin and Seidenberg's (2002) simulations. For present purposes, frequency trajectory was measured as the difference between the log-compressed Grade 13 word frequency and the log-compressed Grade 1 word frequency in occurrences per million. In keeping with Zevin and Seidenberg (2002; 2004), late acquired words had positive values while those early acquired words that declined in relative frequency over time had negative values.

The regression analyses shown in Step 2 of Table 4 were repeated with the addition of frequency Trajectory as a predictor variable. For the Developmental Model, the Trajectory variable predicted a

small amount of additional variance in the model, increasing the  $R^2$  from .327 to .331. The standardized beta value for the frequency Trajectory variable was significant,  $\beta = -.067$ ,  $p < .001$ . This indicated that after cumulative frequency and AoA had been taken into account, words with high-to-low frequency trajectory were read with less ease at the end of training than those with low-to-high frequency. This, in turn, indicates that the Developmental Model is prone to forgetting words which are entered early into training but which receive little training in the late epochs. Put another way, early words are prone to forgetting (catastrophic interference) if they do not continue to be trained alongside the later words, an effect modeled in simulation 12 of Ellis and Lambon Ralph (2000).

Including frequency Trajectory increased the coefficient of the AoA predictor from .217 to .228, showing that AoA was still having a large influence on the Developmental Model's performance. The results indicate that Zevin and Seidenberg's (2002) modelling results are likely due to a combination of cumulative frequency, AoA, and frequency trajectory (or forgetting) effects, with the early AoA words with low frequency late in training (a configuration that rarely happens in natural language) opposing one another as influences on the model's performance. Yet, critically, the effect of AoA is still maintained in the Developmental Model – the effect is not an artefact of either cumulative frequency or frequency trajectory. Bonin et al. (2004) did not find a significant effect of frequency trajectory on word naming latencies, but we have shown that, for a very large set of words, frequency trajectories can be observed in the Developmental Model as having a small influence on naming performance.

#### *Plasticity in the Developmental Model*

Ellis and Lambon Ralph (2000) showed in their backpropagation model that learned to map between sets of pseudo-patterns, that the AoA effect in the model was most likely due to the greater plasticity of the learning system earlier in training. As greater changes to the connections in the model have to occur to incorporate words with inconsistent orthography-phonology mappings, any

reduction in plasticity will have a greater impact on the processing of late-acquired inconsistent words. Reduced plasticity has a smaller effect on learning consistent new words which can inherit the orthography-phonology mappings that the model has already learned.

We analysed the plasticity of the Developmental Model in terms of how rapidly weight changes occurred in the models at different points in training. For every connection in each simulation run of the model, we determined the weight of the connection at time  $t$  and at time  $t+1$ , then calculated the Euclidean distance of the weight change. These weight changes were then averaged across all connections in the model (Orthography to Hidden, Hidden to Phonology, Phonology to Attractor, Attractor to Phonology, bias unit to Hidden, bias unit to Phonology, and bias unit to Attractor). The results for the 10 simulation runs are shown in Figure 5. We conducted an ANOVA on the mean Euclidean distance weight change for each simulation run, with training block as within-subjects factor. There was a significant effect of training block,  $F(13, 117) = 49,093.46$ ,  $MSe = 7.33$ ,  $p < .001$ , showing that there was greater plasticity early in training.

As predicted, the Developmental Model demonstrated a reduction in plasticity as training progressed, with the consequence that later exposure to inconsistent words would mean greater difficulty for the model in learning the orthography-phonology mappings for these words. Zevin and Seidenberg's (2002) model was trained in a similar way to our own and we would predict that a similar pattern of plasticity would occur in their model. This suggests that the absence of AoA effects in their simulations were likely due, not to differences in plasticity, but to differences in the presentation of the training set to their model.

## **Discussion**

### *Effects of AoA and frequency in models of reading*

We have presented three computational models of word naming derived from the Harm and Seidenberg (1999) simulation. For each, we have demonstrated that the order in which words were presented to the model had a profound influence on the model's learning – those words that

occurred earlier in training were read more accurately at the end of training than those words that occurred later in training, even when cumulative frequency and other psycholinguistic factors of the words were taken into account. Both the Developmental Model and the Cumulative Frequency Model demonstrated a statistically significant relationship between a plausible measure of when children were exposed to words during reading development – the WFG-AoA – and the models' accuracy at reading the words. This was despite the different training regimes of the Developmental Model and the Cumulative Frequency Model, indicating that the precise parameters of training were not critical to observing age of acquisition effects in a dynamic model trained incrementally on reading materials. For the Cumulative Frequency Model, the high correlation between age-of-acquisition and frequency meant that higher frequency words tended to be selected earlier in training. The regression analyses enabled us to distinguish the role of order of presentation from genuine frequency effects: both were found to be significant factors in the model's performance.

The critical finding of the present study is that when the error scores of words at the end of training were analysed, the Developmental Model showed an influence of the point at which words were entered into training. This AoA effect, based on an objective measure of point of entry into the learning reader's exposure, was highly significant even with total, cumulative frequency of training as a co-predictor. Thus, point of entry into training played a part in determining the final quality of representations over and above the total number of times a word was sampled in training.

It could be argued that the words children encounter early in their reading experience differ systematically from words encountered later in ways that makes the early words easier to convert from orthography to phonology. Perhaps the authors of texts suitable for grade 1 readers choose words that can be read aloud using relatively simple letter-sound correspondences, or perhaps there is an inherent quality to the orthography-phonology mappings of those words learned early by children that makes them easier to acquire. The Random Frequency Model provided the crucial control to distinguish whether early acquired words tended to be responded to more easily than later acquired words independently of presentation order in the model. For this model, there was no

relationship between WFG-AoA measures and reading accuracy. There are doubtless many reasons why children learn to read some words before others, but the modelling shows this does not mean that AoA is necessarily a surrogate for the effects of whatever real-world factors influence order of acquisition (cf. Bonin et al., 2004; Zevin & Seidenberg, 2002; 2004).

In this respect, the situation with regard to AoA is precisely the same as that relating to word frequency. There are doubtless many reasons why some words are encountered and used more often than others. That does not mean that word frequency is just a surrogate for those many factors. A long line of cognitive and computational models have shown how frequency could influence reading regardless of the factors that cause some words to have higher frequencies than others (e.g., Morton, 1969; McClelland and Rumelhart, 1981; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). It has taken rather longer, but cognitive and computational models, including the ones presented here, are now beginning to reveal how order of acquisition could influence representation and processing irrespective of the factors that cause some words to be earlier acquired than others (e.g., Ellis & Lambon Ralph, 2000; Lambon Ralph & Ehsan, 2006; Li et al., 2004; Richardson & Thomas, 2008; Steyvers & Tenenbaum, 2005).

#### *Comparison with the Zevin and Seidenberg (2002) model*

In contrast to the results from our models, Zevin and Seidenberg's (2002) connectionist model of reading did not demonstrate an AoA effect. As we have observed above, their model differs in several ways from our own more naturalistic simulation of reading development. However, the absence of AoA effects may point to interesting interacting dynamics for models learning to read when trained incrementally on a vocabulary. When we measured frequency trajectory – the difference in frequencies of words at the first versus the last training stage – we found that this made a small, independent contribution to predicting error scores in the Developmental Model once AoA and cumulative frequency had been accounted for. In Zevin and Seidenberg's (2002) simulations, they found no effect of AoA for words presented early to the model when these words

reduced to very low frequencies for later training stages. Our analysis of frequency trajectory suggests that if a word occurs very infrequently in the late stages of training, this will have an impact on a model's ability to form the mapping for that word, though effects of AoA remain an additional influence. Therefore, the small potential effect of frequency trajectory, in both modelling and human studies of word naming (Bonin et al., 2004; Ghyselinck et al., 2004a), may be an additional factor in explaining the process of word naming beyond AoA and frequency effects. Zevin and Seidenberg's (2002) results confound the effects of frequency trajectory with AoA, whereas more naturalistic training regimes, such as those explored in our models, enable these to be assessed independently.

#### *Interactions between AoA, Frequency, and Consistency*

The Developmental Model was trained in the spirit of Ellis and Lambon Ralph (2000) and Lambon Ralph and Ehsan (2006), but this time with plausible representations of words' written and spoken forms. Early learned words were entered into the training of the Developmental Model in the first cycle. Later acquired words only entered training in subsequent cycles. The determinant of point of entry into training for the Developmental Model was an objective measure of AoA derived from the WFG (Zeno et al., 1995). The Developmental Model respected the fact that in natural reading development, children are exposed to some words at the very beginning but take a long time to build their reading vocabulary up to adult levels. Though this does increase the validity of the model's environment, the Cumulative Frequency Model demonstrated that staged training was not critical for the emergence of AoA effects in models of reading, but the Developmental Model did achieve a closer correspondence to the actual order of presentation of words in naturalistic reading.

Additionally, the Developmental Model demonstrated sensitivity to the consistency of words, as well as reading shorter words with more neighbours with greater accuracy, and a frequency effect, with lower error for high frequency words. The Developmental Model also demonstrated an interaction between frequency and consistency, with a greater effect of consistency for low

frequency words. This interaction has been observed in previous computational models of reading (e.g., Harm & Seidenberg, 1999; Plaut et al., 1996; Seidenberg & McClelland, 1989) and mirrors the interaction observed in human word naming (e.g., Andrews, 1992; Brown & Watson, 1994; Hino & Lupker, 2000; J. Monaghan & Ellis, 2002a; Seidenberg et al., 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985). Hence, phasing the entry of words into training did not undermine this classic interaction which J. Monaghan and Ellis (2002a, Experiment 1) showed to hold true for high and low frequency word sets matched on AoA.

The two additional control models provided additional tests that the AoA effects observed in the Developmental Model were genuinely due to the order of presentation of words, not to other uncontrolled properties of early- versus late-acquired words. The Cumulative Frequency Model comes closest to other connectionist models of reading that have attempted to simulate effects of word frequency and other factors in reading (e.g., Harm & Seidenberg, 1999, 2004; Plaut et al., 1996; Seidenberg & McClelland, 1989). Like those models, the Cumulative Frequency Model showed effects of (cumulative) frequency, spelling-sound consistency, length and N. When the contribution of presentation order (PoE) of words was ignored, the interaction between frequency and consistency was also significant (as in other models of reading). But by measuring the point of entry of words into training we were able to confirm our suspicion that effects of order of entry into training lie hidden within models which manipulate frequency in training when trained incrementally. As noted above, high frequency words have a greater chance of being sampled in the earliest stages of training than low frequency words. The high frequency words are therefore able to exert a disproportionate influence when the network is at its most pliable. Point of entry exerted a significant effect in the Cumulative Frequency Model even after the influence of cumulative frequency itself had been accounted for. By obtaining the same results in the Random Frequency Model (in which, as its name suggests, frequencies were assigned to words at random), we were able to show that this pattern of results was not an artefact of other differences between high and low frequency words that affected their learnability in the model. The present simulations taken

together show that order of acquisition *per se* can influence the quality of representations developed within a network, irrespective of the factors that determine natural order of acquisition. The Developmental Model demonstrates that AoA effects can be simulated using staged training, but the other models confirm that order of presentation of words has an impact on the model's performance, due to effects of early plasticity also observable within these models.

Our analysis of the behavior of the Developmental Model shows that the order in which words were entered into training interacted with their frequency and spelling-sound consistency. AoA effects in the Developmental Model were stronger for exception words from inconsistent families than for words from consistent families, echoing the findings of J. Monaghan and Ellis (2002a; 2002b) and Ellis and J. Monaghan (2002) for human word naming. In the Developmental Model, the words to which the model was exposed early in training had the opportunity to make large initial changes to the connection strengths (weights) and therefore to mould the weight space into a form well suited for representing those words. (Note that the first changes to weights are the most substantial and that as the activation of units in the output move away from a neutral value to become closer to 0 or 1, changes resulting from additional training become progressively smaller, see Figure 5 and Ellis and Lambon Ralph, 2000, for discussion.) The impact of an individual word will be greater when it is one of the 103 items in the training set for the first cycle of the Developmental Model than when it is one of 6,229 items in the entire training set in the last training cycle. This particularly benefited inconsistent words, especially if they continued to be sampled at high frequencies throughout the subsequent training of the Developmental Model.

#### *Locus of AoA effects in the human reading system*

One of the reasons we chose the Harm and Seidenberg (1999) network as a basis for our simulations was the fact that it contains only orthographic and phonological representations, which means that any AoA effects observed must reside in orthography to phonology mappings. The more recent model of Harm and Seidenberg (2004) added a pool of semantic units with arbitrary

mappings from orthography and phonology to semantics that reflect the arbitrary nature of the relationship between the spelling and pronunciation of words and their meanings. AoA effects are strong when semantic representations are used to access phonology, as in the object naming task (e.g., Bates et al., 2001; Ellis & Morrison, 1998; Ghyselinck et al., 2004a; Meschyan & Hernandez, 2002). Substantial AoA effects have also been reported in tasks involving semantic processing for written words; for example translation judgements (Izura & Ellis, 2004), production of word associations in Dutch (Van Loon-Vervoorn, 1989; Van Loon-Vervoorn et al., 1988), proper noun/object noun judgements (Brysbaert et al., 2000b), and living/non-living decisions (Ghyselinck, Custers, & Brysbaert, 2004b).

Our modeling shows that semantic representations are not necessarily the only contributor to AoA performance in word naming (cf. Brysbaert et al., 2000b; Cortese & Khanna, 2007; Steyvers & Tenenbaum, 2005): the incremental nature of exposure to words in the language is also likely to have an influence in the formation of mappings between orthography and phonology in the reading system. However, our modeling results are consistent with the finding that increasing arbitrariness in the mapping results in increasing effects of AoA, and so we predict that when orthography to semantics mappings are also implicated in a model of reading, AoA effects are likely to be more emphatic than in the quasi-systematic orthography to phonology mappings we have explored here (see also Lambon Ralph & Ehsan, 2006).

#### *Limitations of the modeling*

We used the WFG of Zeno et al. (1995) to determine the point at which words were entered into training in the Developmental Model (when their frequencies exceeded a reducing threshold for each age) and the frequency with which they were sampled during each subsequent cycle of training. That has the advantage of being entirely objective and based on a frequency count that is promoted by Balota et al. (2004) and Zevin and Seidenberg (2002; 2004) as the best available. But although the WFG is convenient for present purposes, and may be the best source currently

available for English, it is not perfect. Zeno et al. (1995) did not sample texts actually being read by readers of different ages. Rather they took samples from 16,333 written texts which they then classified as suitable for readers at different grade levels on the basis of a “readability” measure involving factors like the length of sentences used. Such a procedure could result in text that is written for adults but uses relatively simple language being classified as a grade 1 or 2 text. This may explain in part why many of the single-syllable words listed in Section I of the WFG had frequencies of 2 or more in grade 1. To avoid this over-representation of early stages of word learning, we used the reducing threshold, so that only reliably occurring words in early reading grades were entered into the model’s training. Future work needs to find an entirely objective basis for building up to the adult vocabulary from a much smaller initial training set.

We have reported for our main analyses the  $R^2$  values which indicate the proportions of variance in error scores achieved collectively by predictors like frequency, consistency, N and AoA. The best that was achieved was around 33%. That does not compare very favorably with the 70% of the variance in word naming latencies accounted for by Morrison and Ellis (2000) with AoA as one of the predictors, or the 50% accounted for by Cortese and Khanna (2007) with AoA, or the 40-50% accounted for by Balota et al. (2004) without AoA. Recently, accounting for the variance in responses in such computational models (e.g., Spieler & Balota, 1997), and relating that to large scale studies of human word naming response times, has become a criterion for assessing model’s performance (Perry et al., 2007). However, large-scale studies of word reading may not be perfectly valid as reflections of lexical access, in that there may be large error in the human responses and a large proportion of variance related to articulatory and acoustic parameters that are not reflected in model performance (Seidenberg & Plaut, 1998). More problematically still, the patterns of naming responses to large sets of single words may become strategic and may not always replicate factorial studies of naming (Sibley et al., 2009). Instead of positioning our own model alongside other, extremely successful, attempts to replicate the data from mega-studies of reading (e.g., Perry et al., 2007), we have instead focused here on exploring the *principles* that contribute to the structure of

the vocabulary in reading.

We have provided an existence proof of the fact that the order of entry of words into training can influence the quality of the representations formed by a computational model that simply learns to convert English orthography into phonology. However, in Harm and Seidenberg (1999), their model was pre-trained with spoken forms of words, before training on the orthography to phonology mappings. We excluded pre-training in our own simulations, in order that we might isolate the AoA effects within the orthography to phonology mappings. Nevertheless, we suggest that pre-training on phonology may even increase the observed effect of AoA on the model's performance. An age-appropriate spoken corpus is likely to contain mostly words that children learn to read early in training. Hence, early-acquired words will have greater efficiency both of phonological representations, as well as mappings from orthography to phonology.

### *Conclusions*

Zevin and Seidenberg (2002, p. 2) wrote that “the finding that AoA affects performance independent of frequency seems to present a challenge for models of word reading”. This is indeed true for practical reasons because incremental training of a model sampling words according to frequency tends to sample the higher frequency words earlier for presentation to the model. The three models we present demonstrate this effect in practice, yet we have shown that PoE, as a reflection of AoA for the model, can be tracked and does contribute to explaining the model's responses in addition to cumulative frequency.

However, it is certainly the case that AoA effects present a challenge to models that do *not* learn through training and experience. J. Monaghan and Ellis (2002a) showed that the DRC model of Coltheart et al. (2001), which successfully predicts many aspects of normal and impaired visual word recognition, was unable to reproduce the effect of AoA on word naming or the interaction of AoA with consistency. J. Monaghan and Ellis (2002a) argued that this is because the DRC model lacks any learning mechanism: the strengths of connections between units are hand-coded and do

not depend on the experience the DRC model has with words. Frequency effects only occur in the DRC model because links between lexical units representing high frequency words are hand-coded to be at different levels from those representing low frequency words. AoA effects have no opportunity to arise because the DRC model does not learn. Similar arguments apply to more recent generations of computational models that use hard-wired localist word-level representations in mapping between orthography and phonology (such as CDP+; Perry et al., 2007). A DRC-type model that learned over time might have the ability to simulate AoA effects, but the current versions do not.

But whereas AoA effects present genuine problems for models that do not learn, we have shown that AoA effects are actually predicted by models like that of Harm and Seidenberg (1999) which learn over time. Words without competitors (i.e., words from wholly consistent orthographic families) can be acquired at almost any time and trained with almost any frequency because they can exploit the orthography to phonology connections established by their uniformly friendly neighbors. But if words with many enemies (inconsistent words) or few neighbours (low N words) are to develop good representations, they need to influence the shape of the input-output mappings while the network is still plastic and major changes to connection strengths are occurring. They then need to be trained at relatively high frequencies to maintain those representations against growing competition. If words with many enemies or few neighbours only enter into training after a network has been shaped by rivals, and are only trained at low frequencies, they will struggle to establish themselves and will show a cost to being late acquired. The key to understanding AoA effects in artificial and natural neural networks is competition for the mappings between representations.

In conclusion, far from representing a challenge to models of reading, AoA effects emerge as a natural property of models that learn over time provided that the learning involves input-output mappings which contain some degree of inconsistency and unpredictability. The cost of being late acquired is greater for inconsistent or exception words than for words with consistent spelling-sound correspondences. When naturalistic learning conditions prevail, it is the absence rather than

the presence of age (or order) of acquisition effects that would represent a challenge to computational models that have learning as a central tenet.

## References

- Alario, F.-X., Ferrand, L., Laganaro, M., New, B., Frauenfelder, U. H., & Segui, J. (2005). Predictors of picture naming speed. *Behavior Research Methods, Instruments and Computers*, 36, 140-155.
- Anderson, K. L., & Cottrell, G. W. (2001). Age of acquisition and connectionist networks. *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society* (pp 27-32). Mahwah, N.J.: Lawrence Erlbaum.
- Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 234-254.
- Baayen, R.H., Pipenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania,.
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yapp, M. J. (2004). Visual word recognition for single syllable words. *Journal of Experimental Psychology: General*, 133, 283-316.
- Barry, C., Johnston, R. A., & Wood, R. F. (2006). Effects of age of acquisition, age, and repetition priming on object naming. *Visual Cognition*, 13, 911-927.
- Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency and name agreement. *Quarterly Journal of Experimental Psychology*, 50A, 560-585.
- Bates, E., Burani, C., D'Amico, S., & Barca, L. (2001). Word reading and picture naming in Italian. *Memory & Cognition*, 29, 986-999.
- Bonin, P., Chalard, M., Méot, A., & Fayol, M. (2002). The determinants of spoken and written picture naming latencies. *British Journal of Psychology*, 93, 89-114.
- Bonin, P., Barry, C., Méot, A., & Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: A never ending story? *Journal of Memory and Language*, 50, 456-476.

- Brown, G. D. A., & Watson, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition*, *15*, 208-216.
- Brysbaert, M., Lange, M., & Van Wijnendaele, I. (2000a). The effects of age of acquisition and frequency of occurrence in visual word recognition: Further evidence from Dutch. *European Journal of Cognitive Psychology*, *12*, 65-85.
- Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000b). Age of acquisition effects in semantic tasks. *Acta Psychologica*, *104*, 215-226.
- Butler, B., & Hains, S. (1979). Individual differences in word recognition latency. *Memory & Cognition*, *7*, 68-76.
- Carroll, J. B., & White, M. N. (1973). Word frequency and age-of-acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, *25*, 85-95.
- Chalard, M., Bonin, P., Méot, A., Boyer, B., & Fayol, M. (2003). Objective age-of-acquisition (AoA) norms for a set of 230 object names in French: Relationships with psycholinguistic variables, the English data from Morrison et al. (1997), and naming latencies. *European Journal of Cognitive Psychology*, *15*, 209-245.
- Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.
- Colombo, L., & Burani, C. (2002). The influence of age of acquisition, root frequency, and context availability in processing nouns and verbs. *Brain and Language*, *81*, 398-411.
- Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance, VI* (pp. 535-555). New York: Academic Press.
- Coltheart, V., Laxon, V. J., & Keating, C. (1988). Effects of word imageability and age of acquisition on children's reading. *British Journal of Psychology*, *79*, 1-12.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204-256.
- Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision

- performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology*, *60*, 1072-1082.
- Cuetos, F., Ellis, A. W., & Alvarez, B. (1999). Naming times for the Snodgrass and Vanderwart pictures in Spanish. *Behavior Research Methods, Instruments and Computers*, *31*, 650-658.
- Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1103-1123.
- Ellis, A. W., & Monaghan, J. (2002). Reply to Strain, Patterson, and Seidenberg (2002). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 215-220.
- Ellis, A. W., & Morrison, C. M. (1998). Real age of acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 515-523.
- Gerhand, S., & Barry, C. (1999). Age of acquisition and frequency effects in speeded word naming. *Cognition*, *73*, B27-B36.
- Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004a). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, *115*, 43-67.
- Ghyselinck, M., Custers, R., & Brysbaert, M. (2004b). The effect of age of acquisition in visual word processing: Further evidence for the semantic hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 550-554.
- Gilhooly, K. J., & Logie, R. H. (1981). Word age-of-acquisition, reading latencies and auditory recognition. *Current Psychological Research*, *1*, 251-262.
- Glushko, R.J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 674-691.
- Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading and dyslexia: Insights from connectionist models. *Psychological Review*, *106*, 491-528.
- Harm, M. W., & Seidenberg, M. S. (2004). Computing the meaning of words in reading:

- cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662-720.
- Hernandez, A. E., & Li, P. (2007). Age of acquisition: its neural and computational mechanisms. *Psychological Bulletin*, *133*, 638-650.
- Hino, Y., & Lupker, S. J. (2000). Effects of word frequency and spelling-to-sound regularity in naming with and without preceding lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 1166-1183.
- Izura, C., & Ellis, A. W. (2002). Age of acquisition effects in word recognition and production in first and second languages. *Psicológica*, *23*, 245-281.
- Izura, C., & Ellis, A.W. (2004). Age of acquisition effects in translation judgment tasks. *Journal of Memory and Language*, *50*, 165-181.
- Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, *13*, 789-845.
- Juhasz, B. J. (2005). Age of acquisition effects in word and picture identification. *Psychological Bulletin*, *131*, 684-712.
- Juhasz, B. J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1312-1318.
- Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Lake, B.M. & Cottrell, G.W. (2005). Age of acquisition in facial identification: A connectionist approach. *Proceedings of the 27<sup>th</sup> Cognitive Science Society Conference* (pp. 1236-1341). Mahwah, NJ: Lawrence Erlbaum.
- Lambon Ralph, M. A., & Ehsan, S. (2006). Age of acquisition effects depend on the mapping between representations and the frequency of occurrence: Empirical and computational evidence. *Visual Cognition*, *13*, 292-248.

- Lewis, M. B. (1999a). Age of acquisition in face categorisation: is there an instance-based account? *Cognition*, *71*, B23-B39.
- Lewis, M. B. (1999b). Are age-of-acquisition effects cumulative-frequency effects in disguise? A reply to Moore, Valentine and Turner (1999). *Cognition*, *71*, 311-316.
- Lewis, M. B., Gerhand, S., & Ellis, H. D. (2001). Re-evaluating age-of-acquisition effects: are they simply cumulative frequency effects? *Cognition*, *78*, 189-205.
- Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, *17*, 1345-1362.
- Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*, 313-330.
- McCann, R.S., & Besner, D. (1987). Reading pseudohomophones: Implications for models of pronunciation assembly and the locus of frequency effects in naming. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 13-24.
- McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*, 375-407.
- McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. In G.H. Bower (Ed.), *The psychology of learning and motivation*, Vol. 24, (pp. 109-165). New York: Academic Press.
- Meschyan, G., & Hernandez, A. (2002). Age of acquisition and word frequency: Determinants of object-naming speed and accuracy. *Memory and Cognition*, *30*, 262-269.
- Miller, G. A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*, 235-312.
- Monaghan, J., & Ellis, A. W. (2002a). What, exactly, interacts with spelling-sound consistency in word naming? *Journal of Experimental Psychology: Learning, Memory & Cognition*, *28*, 183-206.
- Monaghan, J., & Ellis, A. W. (2002b). Age of acquisition and the completeness of phonological

- representations. *Reading and Writing*, 15, 759-788.
- Morrison, C. M., & Ellis, A. W. (1995). The roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 21, 116-133.
- Morrison, C. M., & Ellis, A. W. (1999). Age of acquisition, lexical processing and ageing: changes across the lifespan. *Proceedings of the 21st Annual Meeting of the Cognitive Science Society* (pp. 444-447). Mahwah, N.J.: Erlbaum.
- Morrison, C. M., Hirsh, K. W., Chappell, T. D., & Ellis, A. W. (2002). Age and age of acquisition: An evaluation of the cumulative frequency hypothesis. *European Journal of Cognitive Psychology*, 14, 435-459.
- Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76, 165-178.
- Murray, D. J. (1986). Characteristics of words determining how easily they will be translated into a second language. *Applied Psycholinguistics*, 7, 353-372.
- Nagy, W. E., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly*, 24, 262-282.
- Nation, K. (2009). Form-meaning links in the development of visual word recognition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 27, 3665-3674.
- New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, 13, 45-52.
- Paap, K. R., & Noel, R. W. (1991). Dual-route models of print to sound: Still a good horse race. *Psychological Research*, 53, 13-24.
- Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychological Review*, 114, 273-315.

- Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56-115.
- Ratcliff, R. (1990). Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review*, *97*, 285-308.
- Richardson, F.M. & Thomas, M.S.C. (2008). Critical periods and catastrophic interference effects in the development of self-organizing feature maps. *Developmental Science*, *11*, 371-389.
- Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. Volume 1: Foundations* (pp.318-362). Cambridge, MA: MIT Press.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523-568.
- Seidenberg, M.S., & Plaut, D.C. (1998). Evaluating word-reading models at the item level: Matching the grain of theory and data. *Psychological Science*, *9*, 234 –237.
- Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior*, *23*, 383-404.
- Sibley, D.A., Kello, C.T., & Seidenberg, M.S. (2009). Error, error everywhere: A look at megastudies of word reading. *Proceedings of the 31<sup>st</sup> Annual Conference of the Cognitive Science Society* (pp.1036-1041). Amsterdam, Netherlands: Cognitive Science Society.
- Smith, M. A., Cottrell, G. W., & Anderson, K. L. (2001). The early word catches the weights. In T.K. Leen, T.G. Dietterich, & V. Tresp (Eds), *Advances in Neural Information Systems 13* (pp. 52-57). Cambridge, MA: MIT Press.
- Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, *8*, 411– 416.
- Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks:

- Statistical analyses and a model of semantic growth. *Cognitive Science*, 29, 41-78.
- Strain, E., Patterson, K. E., & Seidenberg, M. S. (2002). Theories of word naming interact with spelling-sound consistency. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 28, 207-214.
- Taraban, R., & McClelland, J.L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language*, 25, 608-631.
- Van Loon-Vervoorn, W. A. (1989). *Eigenschappen van basiswoorden*. Lisse: Swets and Zeitlinger.
- Van Loon-Vervoorn, W. A., Van Ham, P., & Van der Koppen, M. (1988). The importance of age of acquisition for imageability in word processing. In M. Dennis, J. Engelkamp, & J. T. E. Richardson (Eds), *Cognitive and neuropsychological approaches to mental imagery* (pp. 99-107). Dordrecht: Martinus Nijhoff.
- Waters, G.S., & Seidenberg, M.S. (1985). Spelling-sound effects in reading: Time-course and decision criteria. *Memory & Cognition*, 13, 557-572.
- Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, 17, 143-154.
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin and Review*, 15, 971-979.
- Zeno, S. M., Ivens, S. H., Hillard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NJ: Touchstone Applied Science Associates.
- Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, 47, 1-29.
- Zevin, J. D., & Seidenberg, M. S. (2004). Age of acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory and Cognition*, 32, 31-38.

### **Acknowledgements**

We are grateful to Marc Brysbaert, Daragh Sibley, and an anonymous reviewer for very helpful comments on earlier versions of this paper.

Table 1. Epoch, frequency cut-off for entry in training, number of words in training, and number of training presentations for the Developmental Model.

Epoch (stage) of training	Cut-off (frequency/ million)	Number of words in training	Number of training presentations
1	1000	103	50,000
2	100	535	50,000
3	50	775	100,000
4	25	1107	100,000
5	10	1675	100,000
6	8	1833	200,000
7	6	2183	200,000
8	5	2322	200,000
9	4	2492	200,000
10	3	2695	200,000
11	2	2996	200,000
12	1	3307	200,000
13	1	2738	200,000
14	0	6229	3,000,000

Table 2. Mean, SD, minimum and maximum values for the continuous predictor variables used in the analyses of the Developmental, Cumulative Frequency, and Random Frequency Models, and for the Euclidean distance (ED) measures in the three models. For the Random Frequency Model the range of values across the ten simulations are shown.

	WFG- AoA	Cumulative Frequency	Length (in letters)	Consistency	N	Developmental Model ED	Cumulative Frequency Model ED	Random Frequency Model ED
Mean	9.90	160.54	4.81	.92	6.11	.003	.004	[.004, .005]
SD	4.53	198.74	5.18	.19	5.12	.002	.002	[.003, .004]
Min	1	36.43	1	.04	0	.000	.000	.000
Max	14	1458.09	9	1	26	.030	.032	[.034, .143]

*Note.* WFG-AoA = Grade level (determining the point of entry into training in the Developmental Model), frequency is frequency per million in presentations to Developmental Model, N = number of orthographic neighbours.

Table 3. Correlations between variables used in the analyses of the Developmental, Cumulative Frequency, and Random Frequency Models, against Euclidean Distance scores for words in the Models.

	1	2	3	4	5
1. WFG-AoA	—				
2. Cumulative frequency	-.838	—			
3. Consistency	.118	-.154	—		
4. Length (letters)	.144	-.191	-.038	—	
5. N	-.143	.125	-.011	-.625	—
Developmental Model	.287	-.257	-.123	.415	-.487
Cumulative Frequency Model	.286	-.242	-.102	.400	-.483
Random Frequency Model	[-.018, .031] <sup>1</sup>	[-.004, .039] <sup>1</sup>	[-.100, -.013] <sup>9</sup>	[.045, .209] <sup>10</sup>	[-.063, -.251] <sup>10</sup>

Cut-off values for  $r$  to give significance levels:  $r > .042, p < .001$ ;  $r > .033, p < .01$ ;  $r > .025, p < .05$ . Range of correlations is shown for Random Frequency Model with each randomised frequency assignment, with superscript indicating how many of the ten simulations were significantly correlated.

Table 4. Regressions for model error residuals with WFG-AoA after: (1) psycholinguistic variables (cumulative frequency, consistency, length, and N) have been taken into account; and (2) psycholinguistic variables and PoE have been taken into account.

Regression variables	Developmental Model	Cumulative-Frequency Model	Random-Frequency Model
(1) Psycholinguistic variables	$R^2 = .311^{***}$	$R^2 = .295^{***}$	$R^2 = [.120, .165]^{10}$
Residuals with PoE	$\beta = .090^{***}$	$\beta = .123^{***}$	$\beta = [.059, .094]^{10}$
Residuals with WFG-AoA	$\beta = .090^{***}$	$\beta = .080^{***}$	$\beta = [-.043, .011]^3$
(2) Psycholinguistic variables + PoE	$R^2 = .327^{***}$	$R^2 = .314^{***}$	$R^2 = [.127, .173]^{10}$
Residuals with WFG-AoA	$\beta = 0$	$\beta = .013$	$\beta = [-.042, .011]^3$

\*\*\*  $p < .001$ . Superscripts for Random-Frequency Model indicate how many of the ten simulation runs were significant.

Table 5. Standardised regression coefficients for hierarchical multiple regression analysis of Euclidean distance error for the Developmental Model, with cumulative Frequency, Consistency, word Length, and N entered in Step 1, WFG-AoA in step 2, and Frequency x Consistency, WFG-AoA x Consistency, and Frequency x WFG-AoA x Consistency entered in Step 3.

Predictor variable	Standardized $\beta$
Step 1	
Cumulative frequency	-.194***
Consistency	-.162***
Length	.158***
N	-.362***
$R^2 = .311***$	
Step 2	
WFG-AoA	.217***
$R^2 = .327***$	
Step 3	
Consistency x Frequency	.015
Consistency x WFG-AoA	-.394***
Consistency x Frequency x WFG-AoA	-.109***
$R^2 = .336***$	
*** $p < .001$	

### Figure captions

Figure 1. The architecture of the network used for the Developmental Model, Cumulative Frequency and Random Frequency Models.

Figure 2. Performance of (A) Developmental Model, (B) Cumulative Frequency Model, and (C) Random Frequency Model on words and nonwords. The x-axis indicates the number of training cycles. For the Developmental Model, 50K represents the end of the cycle during which the network was trained on words from Grade 1 of the WFG guide. 100K represents the end of Grade 2 training, and so on up to 2M which represents the end of training for Grade 13 words. The network was then trained on to 5 million cycles on the entire vocabulary to ensure adequate learning of even late acquired, low frequency words.

Figure 3. Interaction between consistency and frequency for the Developmental Model with behavioral data from Taraban and McClelland (1987).

Figure 4. Interaction between consistency and early-/late-acquired words for the Developmental Model with behavioral data from Monaghan and Ellis (2002a).

Figure 5. Plasticity in the Developmental Model in terms of Euclidean distance of weight changes between each training block. The weight changes were averaged across all connections in the models, and are given in terms of change per 1000 training patterns.

Figure 1.

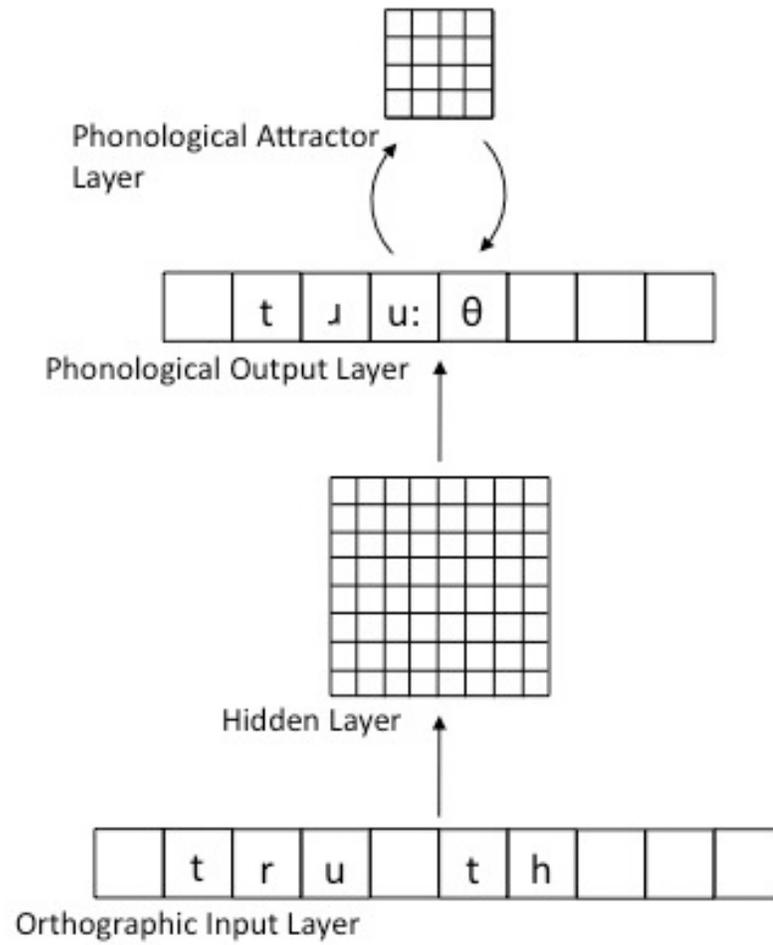


Figure 2.

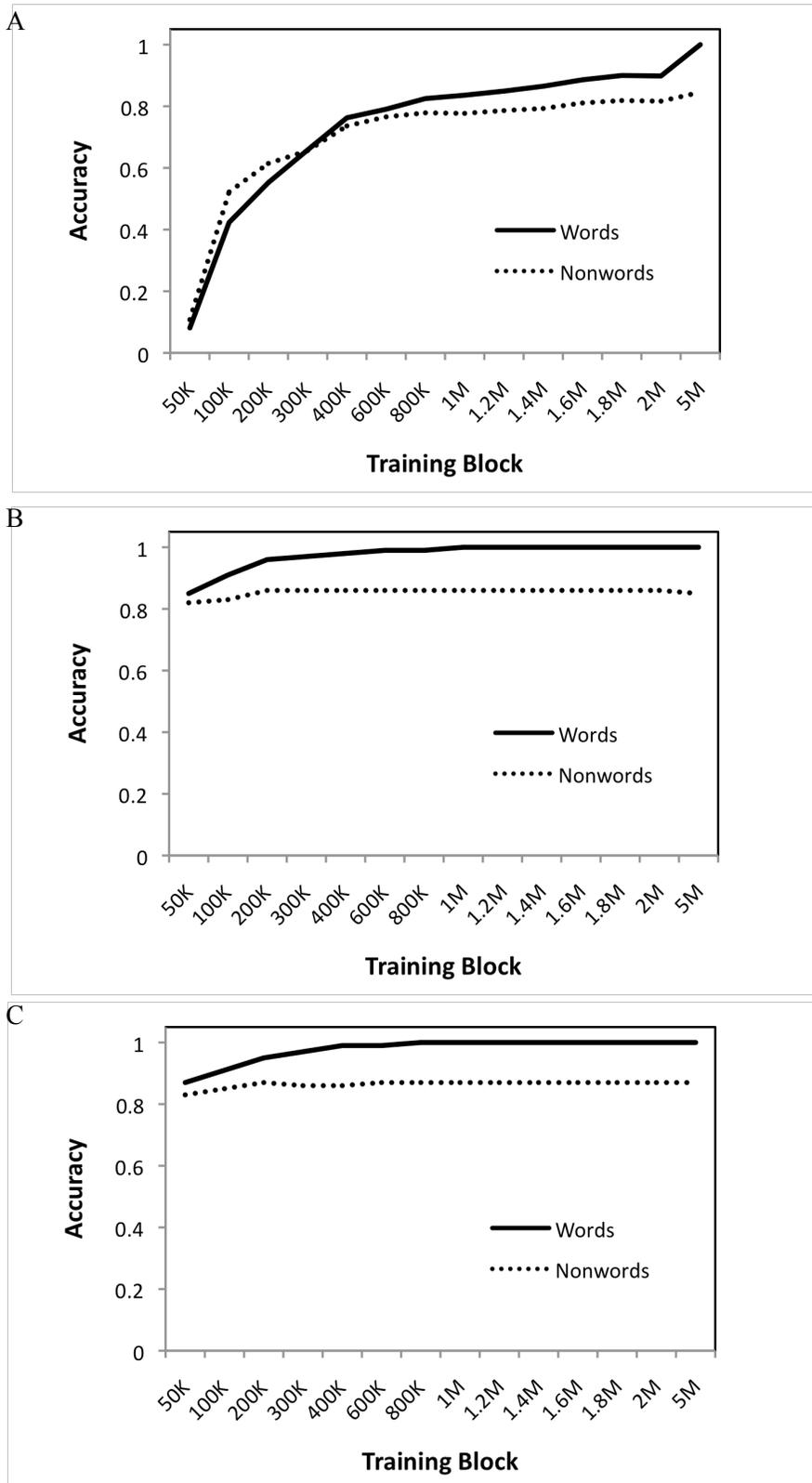


Figure 3.

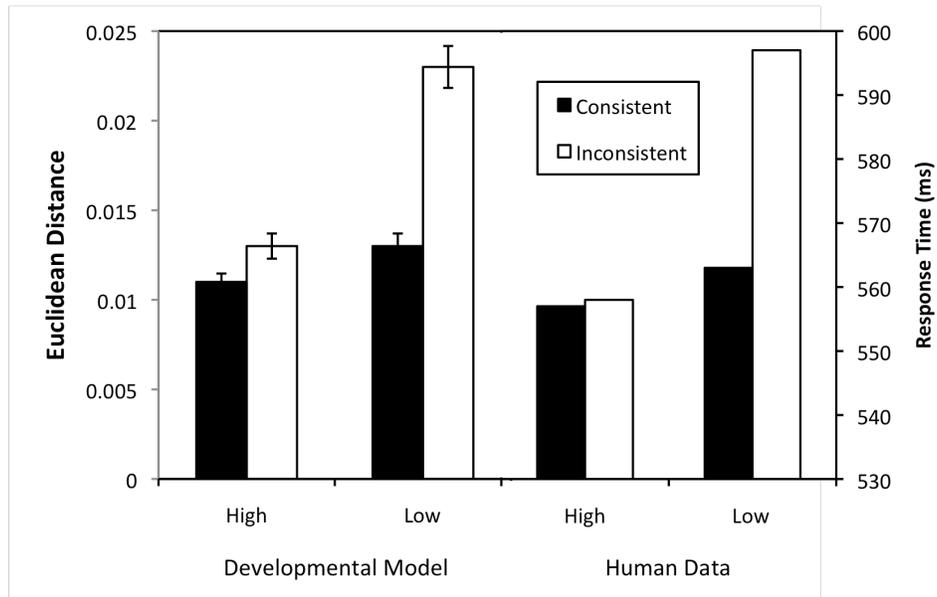


Figure 4.

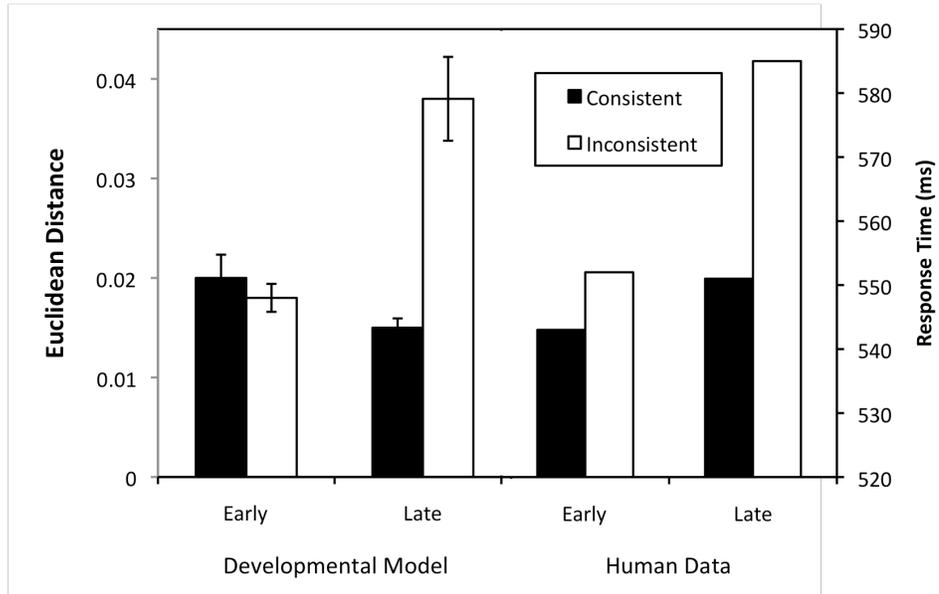


Figure 5.

