Constraints for computational models of reading: Evidence from learning lexical stress

Padraic Monaghan (p.monaghan@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster LA1 4YF, UK
Joanne Arciuli (j.arciuli@usyd.edu.au)
Faculty of Health Sciences, University of Sydney, Australia
Po Box 170, Lidcombe, 1825
Nada Seva (n.seva@psych.york.ac.uk)
Department of Psychology, University of York,
York YO10 5DD, UK

Abstract

Models of reading have typically focused on monosyllabic words, consequently lexical stress assignment has been neglected in such models. Yet, determining how and when stress is applied in naming is an important consideration in providing a comprehensive account of reading processing. A corpus analysis of English age-appropriate reading materials between ages 5-12 indicated two properties: a general tendency that bisyllabic words have initial stress; and the orthographic endings of words increase in importance as indicators of stress position. A behavioural study on children aged 5 to 12 reading nonwords showed that children are sensitive to these properties in their responses. Finally, a developmental computational model mapping orthography onto stress showed the same developmental trend, suggesting that probabilistic orthographic information is used for stress assignment, and that the child's biases may be a consequence of learning the statistical properties of written words.

Models of Reading and Stress Assignment

Models of reading have typically focused on processing the mapping between the orthography and the phonology of the word. The debate over the architecture of the reading system has converged to the question of whether the mapping between written and spoken forms of words is mediated by a single route or a dual route. The single route perspective assumes that the statistical regularities of the properties of letters and sets of letters mapping onto phonemes and sets of phonemes are learned by a mechanism where all these sources of information are available in parallel (Seidenberg & McClelland, 1989). The alternative dual-route view is that the mapping is served by a system that maps the regularities between graphemes and phonemes, and another, separate system that consults a lexicon for the pronunciation of the whole word (Coltheart et al., 2001; Perry, Ziegler, & Zorzi, 2007).

The role of stress assignment provides a test case for constraining the architecture of such models of reading, but it has generally been ignored, presumably due to the limitation of models of reading to monosyllabic words, where stress assignment is trivial. Yet, considering how stress assignment is applied to the word is critical in understanding reading processes. This is because stress must be assigned before the word can be correctly pronounced. In the DRC dual-route model of reading, for regularly pronounced words, the lexical route is accessed after the grapheme-phoneme mapping system is completed. Thus, stress assignment for regular words cannot be stored in the lexicon for this model.

Rastle and Coltheart (2000) attempted a solution to this difficulty by deriving a rule-based system for stress assignment that performed a sublexical search through the word for morphemes, and then consulted a database for whether the morpheme carried stress or not. This model was able to classify a large proportion of the lexicon of English, and could generalize to nonwords when they contained identifiable morphemes. Yet, the model performed poorly when no morphemes were present in the nonwords (Kelly, 2004; Seva, Monaghan, & Arciuli, submitted).

In contrast, the single-route model of reading predicts that stress position can be determined by the learning of regularities in the mapping from orthography to stress position. Zevin and Joanisse (2000) undertook some pilot modeling that examined the regularities potentially available in the orthography, morphology, and phonology of words for determining stress assignment, and found that such cues were highly predictive of stress position (see also Daelemans, Gillis, & Durieux, 1994, for a word-analogy approach to stress assignment). It is one aim of the current study to determine the extent to which a single-route computational model with no pre-determined structure can effectively learn stress assignment from the regularities in the orthography alone for reading in English.

Models of Reading Development

A critical aspect of the single-route, statistical regularity approach to modeling of reading is that the mapping between written and spoken forms of words is learned. So, additional constraints on the architecture of the reading system can be provided by developmental studies. A second principal aim of this paper is to determine the patterns of stress assignment during reading, and determine whether a computational model can produce a similar performance trajectory when trained on a realistic reading corpus.

Previous studies of the development of stress assignment in children have indicated that morpheme frequency has an influence on stress assignment. Jarmulowicz (2002) found that 9 year old children were more likely to stress multi-morphemic words correctly than 7 year olds, but the advantage increased for more frequent morphemes. Protopapas, Gerakaki, and Alexandri (2006) report that Greek children make more errors in stress assignment for nonwords than for words, suggesting that stress may be lexically stored in Greek. The same study also found a default stress assignment on the penultimate syllable in early reading development. Similarly, Gutierrez-Palma and Reyes (2007) found that, for Spanish children aged 7 to 8 years, sensitivity to stress predicted reading ability, and that children with lower reading ability showed a bias toward penultimate stress, the most common stress position in Spanish.

However, there is no even distribution of information across words regarding stress position. The distributions of phonemes and letters at the very beginning and ending of words have been found to be particularly reliable indicators of word boundaries (Hockema, 2006) and grammatical category (Arciuli & Monaghan, 2006), as well as stress position (Arciuli & Cupples, 2006, 2007). In the current study, we extend these previous analyses by combining corpus analysis, experimental, and computational modeling to investigate the role of stress in the reading system. In particular, we:

1. Investigate the extent to which the start and the end of words' orthography can predict stress assignment in the lexicon of English;
2. Determine whether these cues are available in the words that children are exposed to at early stages of reading;
3. Determine the extent to which beginnings and endings of words are utilized by children at different stages of reading;
4. Test whether a computational model instantiating a single route that learns the statistical regularities mapping orthography onto stress position, and without in-built structure, is sensitive to words' beginnings and endings in the same way as children learning to read.

Study 1: Corpus Analyses of Stress

Method

Corpus preparation

The corpus was derived from Section I of the Educator's Word Frequency Guide (WFG, Zeno, Ivens, Hillard, & Duvvuri, 1995). The WFG is constructed from a large corpus of words extracted from 60,527 samples taken from 16,333 written texts. In the corpus, there are approximately 155,000 different words, and the total corpus size is over 17 million words. Each text was graded using readability measures and its age-appropriate level was determined for readers at 13 different grades, covering the age range 5 to 18 in the American and British schooling systems. The 19,468 words with frequency at least 1 per million were included in Section I.

The pronunciation of each word was taken by comparing the WFG wordlist with the CELEX database (Baayen, Pipenbrock, & Gulikers, 1993). When the pronunciation was ambiguous with respect to the orthography from WFG, then the most frequent pronunciation cited in CELEX was used. The analyses were applied to all the bisyllabic words extracted from the WFG. There were 6,531 such words altogether, 5393 with first- and 1138 with second-syllable stress. The number of words with stress on each position for each set of age-appropriate words is shown in Table 1.

Table 1. Corpus size at each age group for the analyses.

AGE	STRESS POSITION (SYLLABLE)	
	1ST	2ND
5/6	3296	447
7/8	4171	683
9/10	4791	903
11/12	4876	957

Corpus analysis

Similar to the analyses of Arciuli and Cupples (2007), we took for each word the letters that corresponded to the pronunciation up to and including the first vowel and vowel cluster as a beginning cue (i.e., 'ze' in 'zebra'). The ending cue were the letters that corresponded to the rime of the second syllable (i.e., "a' in 'zebra'). Each word, then, had just one beginning and one ending cue encoded as a binary variable. For each age-appropriate corpus, we conducted a stepwise discriminant analysis with either the word beginnings or endings as independent variables, and had as dependent variable whether the word had first- or second-syllable stress. We selected all the words with a frequency > 2/million. Each word was weighted according to its frequency as contributing to the analyses. As a baseline, we randomly reassigned the stress positions to words and repeated the discriminant analyses. This enabled us to determine the extent to which the variability of the cues alone, when disconnected from their relationship to stress position, could distinguish two groups.

Results and Discussion

For each age group, both beginnings and endings were highly significantly effective in determining stress position. Figure 1 shows the results for beginnings and endings for each age group.

For each age-appropriate corpus of written words, word endings were more accurate at predicting stress position than were the word beginnings, $F(1, 7) = 140.67$, $p < .001$, $\eta^2 = .95$. However, the difference between beginnings and endings in terms of how accurate they classified words increased with each age-level, with endings remaining highly accurate predictors, and beginnings reducing in predictive value with age. A regression of difference between beginnings and endings and age, was highly significant, $r^2 = .97$, $p < .001$

The corpus analyses confirmed previous studies of adult lexica that indicated that both the beginnings and endings of words were extremely useful in categorizing words according to their stress position. So "ze" from zebra provides information that the word carries first syllable

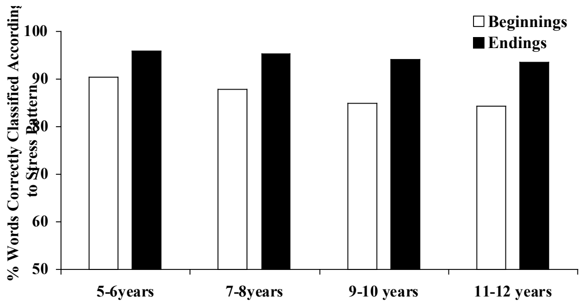

Figure 1. Discriminant analyses of first and second syllable stressed words from age-appropriate text for either beginnings or endings.

stress, however, of increasing importance, the word's ending, "a", indicates that stress is carried on the first syllable. The current study provides a novel contribution by demonstrating that this was true for all age-appropriate groups of words. Intriguingly, the accuracy of the discriminant analysis was greatest for the earliest age-appropriate grouping of words. These words carried more orthographic information about stress position than the words that children were exposed to at later stages of reading development. However, the greater overall accuracy of the early ages was due in part to the smaller set of age-appropriate words for these groups.

As in previous studies of adult lexica of English, the endings carried more information about stress position than beginnings (Arciuli & Monaghan, 2006). This is consistent with affixes carrying information about stress position (e.g., Fudge, 1984) and the greater incidence of suffixes in English, though most of the beginning and ending cues used in the current analyses were not affixes.

However, though we have established that there is substantial orthographic information to stress position, and that this can be found in the very start and end of words, at various ages, the corpus analyses cannot show whether such information is actually utilized by children learning to read. In the next study we constructed a set of nonwords based on cues that were found to be reliable indicators for stress assignment in the corpus analyses, and tested children's sensitivity to these cues at different ages.

Study 2: Development of Stress Assignment

Method

Participants

The participants were primary-aged children from 6 schools in New South Wales, Australia. There were 7 children of 5 years, 31 of 6 years, 29 of 7 years, 25 of 8 years, 27 of 9 years, 28 of 10 years, 25 of 11 years, and 14 of 12 years of age. The children had begun learning to read at school at age 5. Parents of the participants confirmed that their children were monolingual speakers of English without hearing or learning problems. Classroom teachers confirmed that each child in the study was reading at an age-appropriate level.

Table 2. Nonwords used in Study 2, by predicted stress position for beginning and ending cues.

Consistent		Inconsistent	
Beg: 1st syll End 1st syll	Beg: 2nd syll End: 2nd syll	Beg: 1st syll End: 2nd syll	Beg: 2nd syll End: 1st syll
mandol	bedoon	copoon	beppet
mapet	bemade	mapade	ambol
coddol	apoon	comade	andet
combet	abade	madoon	bendol
mambey	bedince	mabince	beddan
condan	adact	codact	ambey

Materials

Using the results of the corpus analyses we identified beginnings that were strongly associated with first syllable stress and endings that were strongly associated with second syllable stress for all age-appropriate groups. So, for instance, the beginnings ma- and co- and the endings –ol and –et were highly associated with first syllable stress. Whereas the beginnings be- and a- and the endings –oon and –ade were associated with second syllable stress. We constructed two sets of nonwords, one where the beginning and ending cues were consistent indicators of stress patterns, and one set where the beginnings and endings were inconsistent. There were 12 consistent and 12 inconsistent nonwords, though the youngest age groups (5 and 6) were only tested on 8 of each, due to testing fatigue. The nonwords are shown in Table 2.

Procedure

Children were tested individually in a quiet room. Nonwords were presented on individual cards, and children were told they were to try and read nonsense words and that there was no right or wrong answer. Responses were coded in terms of whether they read the nonsense word with first syllable stress, second syllable stress, even stress or did not respond. Children produced bisyllabic responses for most of the nonwords, and the vast majority of responses fell into the first two categories so we report only these in the analyses.

Results and discussion

Figure 2 shows the results for children reading nonwords with consistent cues, where the cues predicted either first or second syllable stress (error bars show SE). We conducted an ANOVA on the proportion of nonwords pronounced as predicted by the cues, with age and 1st or 2nd syllable stress cues as factors. There was a main effect of 1st/2nd stress cues, F(1, 182) = 216.02, $p < .001$, a main effect of age, F(1, 182) = 8.04, $p < .001$, and a significant interaction between stress cues and age, F(1, 182) = 10.70, $p < .001$.

Figure 3 illustrates the results for the nonwords when cues were inconsistent, according to whether stress was applied based on the predictions of the beginning or the ending of the nonword. An ANOVA with age and type of cues (beginning 1st a nd ending 2nd, and beginning 2nd and ending 1st) was performed on the proportion of nonwords

Figure 2. Classification of nonwords as first or second syllable stress for consistent cues.

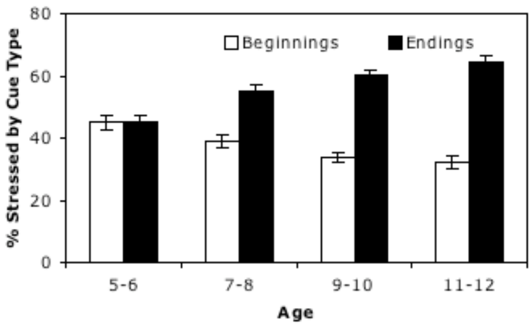

Figure 3. Classification of nonwords according to beginnings or endings for inconsistent cues.

pronounced according to the ending cue. There was a main effect of type of cue, $F(1, 182) = 235.22$, $p < .001$, a main effect of age, $F(1, 182) = 13.19$, $p < .001$, and a significant interaction, $F(1, 182) = 11.55$, $p < .001$.

The results demonstrate two principal effects of learning stress assignment in reading in English. First, young children indicate a greater tendency than older children to pronounce nonwords with first syllable stress. This suggests an over-generalisation, consistent with the distribution of first- and second-syllable stressed words in the language, evident from the corpus analyses – 85% of the bisyllabic words in the WFG database have first syllable stress. This over-generalisation bias applied for all nonwords types. For nonwords with consistent beginning and ending markers for first and second syllable stress, children aged 5/6 matched their responses closely to the actual distributions in their age-appropriate lexicon: approximately 80% of responses to the consistent nonwords were pronounced with first syllable stress. Similarly, for the inconsistent words, children aged 5/6 pronounced almost 67% of the words with first syllable stress. Yet, it would appear that as the child has more practice with reading (corresponding to increasing age), more words are read with second syllable stress – for the consistent nonwords, 54% are pronounced with first syllable stress by the 11/12 year old group, and for inconsistent words this was 57%.

The second main property of the reading development responses is the change in the dominance of the cues as age increases. When reading begins, at age 5/6, stress assignment was generally placed on the first syllable, and neither beginning nor ending orthographic cue had a large influence on stress assignment. As age increases, however, words' endings become more predictive of stress position than word beginnings, until by age 9 upwards, almost 70% of stress assignments are made in accordance with the stress position of regularities at word endings. This dominance of ending cues reflects the relative value of endings from the corpus analyses as being more reliable predictors of stress position. Children, as they extend their lexicon for reading, learn to rely more heavily on the region of the word that is most predictive of stress position.

The third study we performed was a computational model of learning to assign stress from written words in English. We tested whether a computational model was sensitive to the same statistical properties of the language that children appear to be, and whether a similar developmental profile in the model could also be observed. The computational model is a statistical mechanism, but in contrast to the discriminant analyses, it enables us to test whether the relationship between orthographic cues and stress position can be discovered by the model, and it also enables us to plot a developmental trajectory – the discriminant analysis provides the solution for the beginnings/endings to stress position mapping, but does not provide insight into a computational system's *approach* to this solution.

Study 3: Modeling Stress Assignment

The computational model was trained on all bisyllabic words in English according to their frequency of occurrence in age-appropriate literature for ages 5-12, and was required to predict from the orthographic input the stress position of the word. Critically, we wanted to determine whether the model could: (1) learn only from orthographic cues the stress position of the word in English; (2) demonstrate early in learning the same bias as young children in terms of assignment of first syllable stress to words and nonwords; and (3) whether the model could reveal the developing dominance of word-endings as indicators of stress position.

Method

Architecture

We constructed a supervised feedforward network that learned to map the orthography of words onto their stress position. There were two orthographic input layers, composed of 14 letter slots each. The first contained a left-aligned representation of the word and the second contained a right-aligned version of the word. Consequently, the potential influence of beginnings and endings on the model's performance was balanced. In each letter slot, one of 26 units was active to represent the letter, or for an empty slot no units were active. The input layers were fully connected to a layer of 100 hidden units, which in turn was fully connected to one output stress unit (see Figure 4). The output unit was trained to be active for second syllable stress and inactive for first syllable stress.

Training and testing

The model was trained incrementally with age-appropriate words taken from the WFG list. For the age 5 words, we

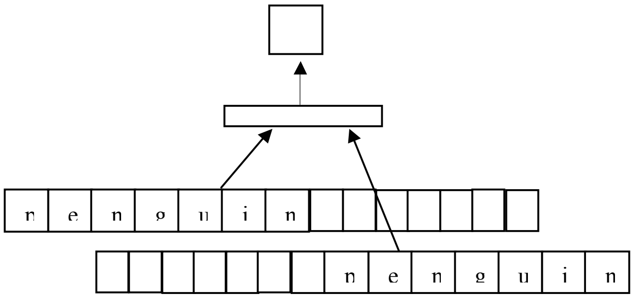

Figure 4. The computational model of stress assignment.

trained the model for 10,000 examples, for age 6 words we trained for 20,000 examples, and this increased up to 150,000 examples of training for the age 12 words. The model was trained with the backpropagation learning algorithm with learning rate .005. For first-syllable stress, the stress unit's target activity was 0, for second syllable stress it was 1. The model was tested at each learning stage, and the simulation was repeated 10 times.

The model was judged to have assigned first syllable stress if activation of the output unit was less than .5, and second syllable stress for activity greater than .5. The model was tested on all the words in the age-appropriate set. The model was also tested on the age appropriate lists of consistent and inconsistent nonwords used for the behavioural study (shown in Table 2).

Results and Discussion

The model learned to assign stress to words in the training set accurately, reading 88% of words at age 5/6 correctly, rising to 92% correct at age 11/12. Additional training resulted in 99% accuracy, though we do not report the model's performance at this more advanced reading stage.

For the consistent nonwords, Figure 5 shows the models stress assignment. An ANOVA with $1^{st}/2^{nd}$ syllable stress cues and age as factors was performed on the proportion of stress assignments consistent with the cues. As with the behavioural data, there was a main effect of $1^{st}/2^{nd}$ stress cues, $F(1, 9) = 19.76$, $p < .001$, a main effect of age, $F(3, 27) = 768.23$, $p < .001$, and an interaction, $F(3, 27) = 19.76$, $p < .001$.

For the inconsistent nonwords, Figure 6 illustrates the results, according to whether stress was applied based on the predictions of the beginning or the ending cue. An ANOVA with type of cue (beginning 1^{st} and ending 2^{nd}, and beginning 2^{nd} and ending 1^{st}) and age on the proportion of nonwords given 1^{st} syllable stress by the model. As with the behavioural study, there was a main effect of type of cue, $F(1, 9) = 1046.10$, $p < .001$, a main effect of age, $F(3, 27) = 49.25$, $p < .001$, and an interaction between type of cue and to rely on nonword endings at a slightly earlier stage than age, $F(1, 9) = 136.84$, $p < .001$, though the model appeared did the children in Study 2.

Overall, the simulation results indicated that the computational model was able to learn the stress position of words and nonwords in English from orthographic cues

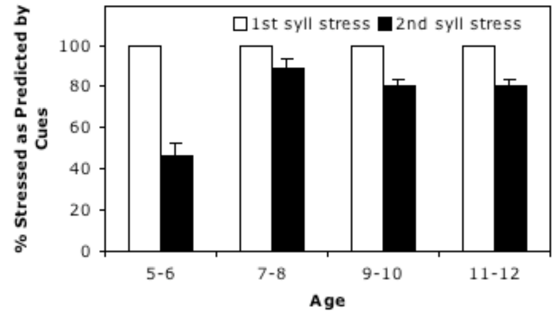

Figure 5. Model classification of nonwords as first or second syllable stress for consistent cues.

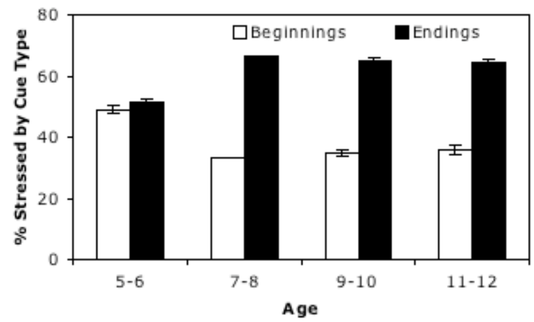

Figure 6. Model classification of nonwords according to beginnings or endings for inconsistent cues.

alone. In addition, early in training on realistic corpora of age-appropriate written language, the model was sensitive to the same biases as young children, both in terms of assignment of first syllable stress to words and nonwords and for the developing dominance of word-endings as indicators of stress position.

General Discussion

We have explored the statistical properties of written words in terms of the information they provide about the stress position of bisyllabic words. The corpus analyses indicate that, for age-appropriate lexica, there is substantial information in the letters at the beginning and ending of the word to determine with a high degree of accuracy stress assignment. As age increases, the relative advantage of endings compared to beginnings as determinants of stress becomes greater. The behavioural experiments indicate that children are sensitive to these statistical properties of English. As children have more practice in reading, they lose an initial bias to pronounce unfamiliar words with first syllable stress, and become more attuned to the orthographic properties of word endings in guiding their decisions about pronunciation.

The corpus and behavioural results provide insights into the potential processing involved in the reading system for applying stress. A computational model trained to map orthography onto stress position for bisyllabic words indicated similar performance to children learning to read. Initially, the first-syllable stress bias was found, but with an increasing dependency on word endings in guiding pronunciation. The model indicated that learning the statistical properties of the lexicon resulted in effects found in reading development.

Importantly, the computational model we have presented does not have a lexical component for stress assignment. The lexical level is potentially available, as all letters within the word are inputted to the model simultaneously, yet the accurate encoding of stress by the model is not due to access to a specified lexicon. Furthermore, the finding that (non-morphological) parts of the word can provide valuable information about stress, and the fact that generalization to novel words can take place argues against stress being stored in the lexicon (though it is still possible that a word-analogy approach may also be able to reflect the behavioural data). Our model suggests instead that sublexical properties of words drive their pronunciation.

The model, however, provides insight into only one aspect of pronunciation of polysyllabic words – that of stress assignment – but as such it indicates that assumptions made for models of monosyllabic words do not generalize adequately to the entire lexicon. A notable absence from the current model is the mapping from orthography to phonology, which is necessary for providing a full account of polysyllabic word reading (e.g., Ans, Carbonnel, & Valdois, 1998). Yet, incorporating phonology into the model as well may improve performance further. Rastle and Coltheart's (2000) rule-based model of stress, for instance, bases its decision in part on the pronunciation of the vowel which contributes substantially to accurate performance.

Another absence is the availability of other cues that may assist in stress assignment. The correspondence between grammatical category and stress position is well-attested (Arciuli, & Slowiaczek, 2007) as well as phonological and phonotactic indicators of stress position (Kelly, 2004). A sublexical, probabilistic approach to modeling word reading, such as we have taken in this paper, can be extended using the same principles to incorporate these additional sources of information.

Author Note

All authors of this paper contributed equally. The research was supported by a bilateral funding scheme: ESRC grant RES-000-22-1975 and ARC grant LX0775703.

References

Ans, B., Carbonnel, S., & Valdois, S. (1998). A connectionist multi-trace memory model of polysyllabic word reading. *Psychological Review, 105*, 678-723.

Arciuli, J., & Cupples, L. (2007). Would you rather 'embert a cudsert' or 'cudsert an embert'? How spelling patterns at the beginning of English disyllables can cue grammatical category. In Schalley, A. & Khlentzos, D. (Eds.), *Mental States: Language and Cognitive Structure.*

Arciuli, J., & Cupples, L. (2006). The processing of lexical stress during visual word recognition: Typicality effects and orthographic correlates. *Quarterly Journal of Experimental Psychology, 59*(5), 920-948.

Arciuli, J., & Monaghan, P. (2006). Hidden cues to grammatical category in the spelling of English disyllables. The 15th Australian Language and Speech Conference. *Australian Journal of Psychology, 58*, 1-13.

Arciuli, J. & Slowiaczek, L. (2007). The where and when of linguistic word-level prosody. Neuropsychologia, 45, 2638-2642.

Baayen, R.H., Pipenbrock, R. & Gulikers, L. (1993). *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., Ziegler, J. (2001). DRC: Dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204-256.

Daelemans, W., Gillis, S., & Durieux, G. (1994). The Acquisition of Stress: A Data-oriented Approach. *Computational Linguistics, 20*, 421-451.

Fudge, E. (1984). *English word stress*. London: Allen & Unwin.

Gutierrez-Palma, N. & Reyes, A.P. (2006). Stress sensitivity and reading performance in Spanish: A study with children. *Journal of Research in Reading, 30,* 157-168.

Hockema, S.A. (2006). Finding words in speech: An investigation of American English. *Language Learning and Development, 2*, 119-146.

Jarmulowicz, L. (2002). English derivational suffix frequency and children's stress judgments. *Brain & Language, 81*, 192-204.

Kelly, M. (2004). Word onset patterns and lexical stress in English. *Journal of Memory and Language, 50*, 231-244.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modelling in the development of computational theories: the CDP+ model of reading aloud. *Psychological Review, 114*, 273-315.

Protopapas, A., Gerakaki, S., & Alexandri, S. (2006). Lexical and default stress assignment in reading Greek. *Journal of Research in Reading, 29*, 418-432.

Rastle, K. & Coltheart, M. (2000). Lexical and nonlexical print-to-sound translation of disyllabic words and nonwords. *Journal of Memory and Language, 42*, 342-364.

Seidenberg, M.S. & McClelland, J.L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review, 96*, 523-568.

Seva, N., Monaghan, P., & Arciuli, J. (submitted). Stressing what is important: Orthographic cues and lexical stress assignment.

Zeno, S.M., Ivens, S.H., Millard, R.T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

Zevin, J. & Joanisse, M. (2000). Stress assignment in nonword reading. *Journal of Cognitive Neuroscience, 41B*, S5.