Journal of Memory and Language 93 (2017) 169–192

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Moving beyond the monosyllable in models of skilled reading: Mega-study of disyllabic nonword reading

Petroula Mousikou *, Jasmin Sadat, Rebecca Lucas, Kathleen Rastle

Department of Psychology, Royal Holloway, University of London, United Kingdom

ARTICLE INFO

Article history:
Received 14 December 2015
revision received 6 September 2016
Available online 20 October 2016

Keywords:
Mega-study
Reading aloud
Computational reading models
Generalization
Stress assignment
Pronunciation

ABSTRACT

Most English words are polysyllabic, yet research on reading aloud typically focuses on monosyllables. Forty-one skilled adult readers read aloud 915 disyllabic nonwords that shared important characteristics with English words. Stress, pronunciation, and naming latencies were analyzed and compared to data from three computational accounts of disyllabic reading, including a rule-based algorithm (Rastle & Coltheart, 2000) and connectionist approaches (the CDP++ model of Perry, Ziegler, & Zorzi, 2010, and the print-to-stress network of Ševa, Monaghan, & Arciuli, 2009). Item-based regression analyses revealed orthographic and phonological influences on modal human stress assignment, pronunciation variability, and naming latencies, while human and model data comparisons revealed important strengths and weaknesses of the opposing accounts. Our dataset provides the first normative nonword corpus for British English and the largest database of its kind for any language; hence, it will be critical for assessing generalization performance in future developments of computational models of reading.

Introduction

Research on the mental processes involved in visual word recognition and reading aloud has flourished in the past 30 years. This advancement is demonstrated most clearly by the fact that precise accounts of how people recognize printed words and read them aloud are now available in the form of computational models that seek to mimic human reading behavior (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Perry, Ziegler, & Zorzi, 2007; Plaut, McClelland, Seidenberg, & Patterson, 1996). These models have been remarkably successful in explaining how skilled readers translate print into sound, and have also offered fresh insights into our understanding of typical and atypical reading development (e.g., Jackson & Coltheart, 2001), acquired impairments of reading (e.g., Coltheart, 2006), and the genetic and neural basis of reading (Bates et al., 2010; Taylor, Rastle, & Davis, 2013).

However, most empirical and computational work in this domain has focused on monosyllables, which constitute only a very small minority of the words in English and are practically absent from many of the world's other languages. Moving beyond the monosyllable presents more than a problem of scale – at least for theories concerned with reading aloud. Indeed, English polysyllables present not only greater opportunities for inconsistency in the spelling-to-sound mapping than monosyllables, but also raise special additional challenges such as syllabification, stress assignment, and vowel reduction that are not relevant to monosyllables.

This article seeks to advance our understanding of how skilled readers process letter strings with more than one syllable. We report a mega-study in which 41 adults read aloud 915 disyllabic nonwords (yielding over 37,000 read-

* Corresponding author at: Max Planck Institute for Human Development, REaD (Reading Education and Development), Lentzeallee 94, 14195 Berlin, Germany.
E-mail address: mousikou@mpib-berlin.mpg.de (P. Mousikou).

http://dx.doi.org/10.1016/j.jml.2016.09.003
0749-596X/

ing aloud responses). Our analyses seek to identify the cues to stress assignment in English, and to uncover the factors that influence pronunciation variability and reading aloud latencies across adult readers. We also use these data to assess for the first time the adequacy of opposing computational accounts of disyllabic reading aloud, namely the CDP++ model (Perry, Ziegler, & Zorzi, 2010), the rule-based disyllabic algorithm of Rastle and Coltheart (2000), and the connectionist print-to-stress network of Ševa, Monaghan, and Arciuli (2009). In doing so, our data further offer an insight into the question of whether the print-to-stress and print-to-sound mapping is best described in terms of rules or learned statistical relationships. Finally, we make available this entire database of human and model data to facilitate rapid theoretical advancement in the area of reading.

Monosyllabic models of reading aloud

There is now strong agreement among reading theorists that two kinds of procedures are used in the translation of orthography to phonology (Taylor et al., 2013; Woollams, Lambon Ralph, Plaut, & Patterson, 2007). One of these procedures is particularly important for supporting the reading aloud of irregular/inconsistent words that do not respect typical spelling-to-sound relationships (e.g., pint, yacht). The other procedure specializes in the computation of phonology using sublexical information and is particularly important for supporting the reading aloud of unfamiliar words or nonwords (e.g., vib, slint), which are not stored in lexical memory. There is strong evidence that both pathways are routinely used in parallel in skilled reading (Paap & Noel, 1991; Rastle & Coltheart, 1999).

Though all models of reading aloud subscribe to this general dual-pathway architecture, they differ in important ways. The DRC model (Coltheart et al., 2001) proposes that the lexical pathway consists of local orthographic and phonological representations of known words, while the sublexical pathway consists of rules that relate graphemes to phonemes. Conversely, the 'triangle model' (Harm & Seidenberg, 2004; Plaut et al., 1996) comprises learned mappings between orthographic, phonological, and semantic units; lexical knowledge in this model is represented in a distributed manner and the learned mappings between representations are probabilistic rather than rule-based. The more recent CDP+ model (Perry et al., 2007) combines features of these two types of models. Its lexical pathway is identical to that of the DRC model while its sublexical pathway consists of a two-layer network that learns the mappings between graphemes and phonemes, as well as a grapheme identification and parsing procedure.

Different approaches have been adopted to adjudicate between these models. One such approach involves investigating their generalization performance (i.e., the models' ability to read aloud nonwords). Besner, Twilley, McCann, and Seergobin (1990) were the first to use this approach in relation to models of reading, demonstrating that an early version of the triangle model (Seidenberg & McClelland, 1989) showed far worse generalization performance than human readers. Subsequently, investigating nonword reading performance has become a key aspect of model evaluation. Most recently, Pritchard, Coltheart, Palethorpe, and Castles (2012) used this approach to evaluate the reading performance of the sublexical pathways of the DRC (Coltheart et al., 2001) and CDP+ (Perry et al., 2007) models. They selected 412 monosyllabic nonwords for which the models disagreed on pronunciation, and compared these model responses to the pronunciations given by 45 adult readers. Results showed that the modal human pronunciation matched the DRC pronunciation in 74% of cases, whereas it only matched the CDP+ pronunciation in 12% of cases. However, one could argue that models should not be expected to behave as the average human participant, but rather as an individual participant within the normal distribution. This is because many nonwords yield more than one pronunciation when read aloud by different people (Andrews & Scarratt, 1998; Glushko, 1979; Masterson, 1985; Seidenberg, Plaut, Petersen, McClelland, & McRae, 1994). Hence, in considering what counts as an acceptable response in the models, Pritchard et al. (2012) took into account not only the most frequent response given by participants, but also whether the model pronunciations matched *any* human response. Results showed that the DRC model produced a response that was unlike any human response for 2% of the nonwords; for the CDP + model, this figure was far higher, at 49% of the nonwords.

Disyllabic models of reading aloud

Rastle and Coltheart (2000) claimed that the dual-pathway framework is ideally suited to consider the problem of disyllables, in particular, with respect to the assignment of stress. They argued that there are many examples in which stress cannot be predicted by rule (e.g., **ca**mel versus ca**nal**), and thus, that it must be stored in lexical memory along the lexical pathway. However, in a small-scale nonword reading aloud study, they found that individuals assign quite consistently first- or second-syllable stress to nonwords – for example, 100% of their participants gave first-syllable stress to the nonword 'laifun' while 93% gave second-syllable stress to the nonword 'itesque' – suggesting that stress computation must also be arising along the sublexical pathway. This dual-pathway approach to stress assignment is consistent with previous research on reading aloud in Italian (Colombo, 1992).

Rastle and Coltheart (2000) considered how this dual-pathway theory of disyllabic word reading could be implemented in the DRC model. They argued that while it would be straightforward to add entries with stress information for disyllabic words to the lexical pathway, implementing a set of spelling-to-sound and spelling-to-stress rules suitable for disyllables along the sublexical pathway poses a greater challenge. As an initial step, they proposed a partial implementation of a rule-based sublexical pathway that translates printed disyllables to sound and applies a stress marker. This partial implementation calls on the grapheme-to-phoneme translation rules used by the DRC model (Rastle & Coltheart, 1999; and later, Coltheart et al., 2001), and in addition, identifies orthographic strings corresponding to prefixes and suffixes to determine stress

placement. Evaluation of the algorithm revealed good performance in stressing disyllabic words and nonwords. It accurately predicted stress for nearly 90% of disyllabic words in the CELEX database (Baayen, Piepenbrock, & van Rijn, 1993). Further, in an experiment in which participants read aloud 210 disyllabic nonwords, the algorithm accurately predicted modal human stress for 84% of the items. The algorithm is presented in Fig. 1.

Despite its relative success in predicting stress assignment for words and nonwords, Rastle and Coltheart's (2000) algorithm has some substantive limitations. For example, the algorithm expresses a set of hypotheses about the rules relating spelling to sound and spelling to stress for disyllabic letter strings. Yet, in order to be fully tested, it would need to be implemented as part of a processing model (e.g., the DRC model) that produces reaction times in addition to pronunciations and stress. However, Rastle and Coltheart (2000) identified significant difficulties in implementing this algorithm as part of the DRC model. These difficulties arise because the proposed algorithm requires information from all parts of the printed stimulus, from beginning to end, to compute a pronunciation and stress marker. For example, the algorithm looks for a suffix at the end of the word prior to the application of grapheme-to-phoneme translation rules. However, the sublexical pathway of the DRC model computes a pronunciation serially, from left to right (Rastle & Coltheart, 1999). Thus, Rastle and Coltheart (2000) argued that further work would be required to understand how the hypotheses advanced in their algorithm could be reconciled with the serial left-to-right operation of the sublexical pathway of the DRC model.

In more recent years, researchers have considered the problem of polysyllabic word reading within models that consist of a learned mapping of the spelling-to-sound and spelling-to-stress relationship. The best developed of these is the CDP++ model (Perry et al., 2010), which is a dual-pathway model of reading aloud comprising lexical and sublexical processes for mapping print-to-sound. It is very similar to the CDP+ model (Perry et al., 2007), except for a number of minor modifications, including an increase in the number of letter and phoneme slots to accommodate longer words, a change to the input coding template to accommodate disyllables, the introduction of the schwa phoneme to deal with vowel reduction, the introduction of stress nodes to represent the position of stress, and the use of a far larger training corpus and lexicon. Unlike the algorithm presented by Rastle and Coltheart (2000), the CDP++ model is a full processing model that produces a pronunciation, stress marker, and reaction time. The model is presented in Fig. 2.

The evaluation of the CDP++ model showed very good performance against the available datasets on word reading (Perry et al., 2010). The model accurately read aloud over 32,000 words in its lexicon (with an error rate of less than 1%), and it simulated a number of key benchmark effects in the monosyllabic domain. The model also showed very strong performance in capturing variance in reading aloud latency in large-scale studies of word reading (e.g., Balota et al., 2007), explaining over 49% of variance in reading aloud latencies in a selection of monomorphemic monosyllables and disyllables (e.g., Yap & Balota, 2009), after variance due to phonetic onset was taken into account. However, the model's performance on reading nonwords aloud appears to be more mixed. Perry et al. (2010) reported that the model correctly read aloud approximately 95% of the monosyllabic nonwords in the corpus of Seidenberg et al. (1994), but this was using a lenient scoring criterion in which responses were considered as correct if they contained *any* grapheme-phoneme or body-rime response that exists in English words. Evaluating the model's generalization performance using the Pritchard et al. (2012) monosyllabic corpus (which offers the range of possible pronunciations across a sample of participants) gave a very different picture. Although the performance of CDP++ was better than that of CDP+ for this set of nonwords (see above), the CDP++ model gave the modal human pronunciation in only 38% of cases. Further, the model gave a pronunciation unlike any human response in 27% of cases. Robidoux and Pritchard (2014) further analyzed the Pritchard et al. (2012) dataset using hierarchical clustering techniques, which involved grouping participants on the basis of the overall similarity of their pronunciations. In particular, they compared individual subjects' reading profiles and they then examined whether the DRC and CDP++ models fitted any of the participants' profiles. They observed that while the DRC model fitted other participants' reading profiles at least as well as participants' fitted one another, the CDP++ model did not match the reading profile of any of the 45 participants that took part in the study. It is worth noting that the CDP++ model has not been tested extensively against disyllabic nonword reading aloud data, though Perry et al. (2010) tested the model's performance on stress assignment against the small nonword reading aloud dataset of Rastle and Coltheart (2000). These simulations showed very good capture of the human data, yet the CDP++ model was slightly more biased toward first-syllable stress than was the case for human readers.

Ševa et al. (2009) also investigated whether stress placement in disyllables could be determined through orthographic regularities using a distributed-connectionist framework. They developed a model that learns to map an orthographic input onto a stress pattern. However, this model does not provide a pronunciation or reaction time. The architecture of the model consists of 364 orthographic input units (26 letters $*$ 14 slots), 100 hidden units, and 1 output unit. The activation of the output unit is graded, but for the purposes of the simulations reported by Ševa et al. (2009), activations below 0.5 were treated as first-syllable stress while activations above 0.5 were treated as second-syllable stress. The network's performance was tested on the set of words and nonwords used by Rastle and Coltheart (2000). The model performed slightly better than the rule-based algorithm on assigning stress to disyllabic *words*. However, it showed substantially inferior performance to the rule-based algorithm when assigning stress to *nonwords*. This inferior performance was due to the model assigning first-syllable stress to nonwords that were given second-syllable stress by a majority of participants. The model of Ševa et al. (2009) is shown in Fig. 3.

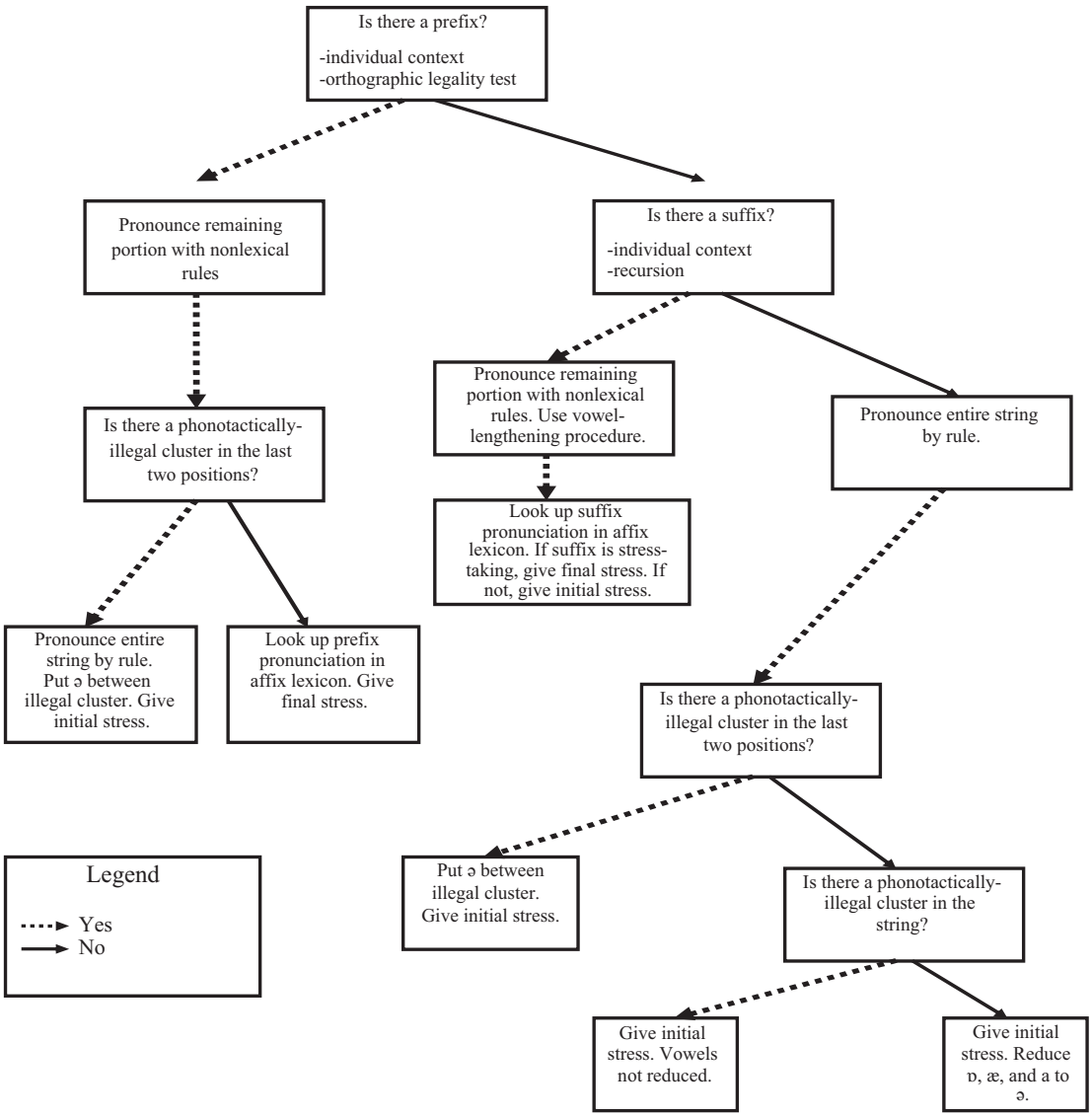

Fig. 1. The rule-based algorithm of Rastle and Coltheart (2000).

Other connectionist approaches to disyllabic reading include the Junction model (Kello, 2006; Sibley, Kello, & Seidenberg, 2010), which was benchmarked against naming latency data from the English Lexicon Project (Balota et al., 2007). However, the first of these models showed very poor generalization performance on the monosyllabic nonwords used by Seidenberg et al. (1994) and the monosyllabic and disyllabic nonwords used by Sibley, Kello, Plaut, and Elman (2008). The generalization performance of the second model on the Seidenberg et al. (1994) nonwords was significantly better; however, its ability to read nonwords decreased substantially as their length increased. As the authors acknowledge, "it is likely that performance would not decrease as much as in the present simulation if skilled readers were to name bisyllabic pseudowords from our corpus. Thus an important task for future modeling work will be to investigate methods of improving pseudoword naming." (Sibley et al., 2010, p. 664). The multiple-trace memory model (Ans, Carbonnel, & Valdois, 1998) is another connectionist approach to simulating polysyllabic reading in French. However, this model does not have a component in its architecture that allows it to deal with stress assignment. This is not a problem in a fixed-stress language like French, in which stress is always placed on the last syllable; however, this model cannot offer an account for polysyllabic reading in a language with variable stress such as English. For all of the above-mentioned reasons, we did not consider any of these models in the present work.

Finally, in order to account for polysyllabic reading in languages with variable stress patterns, a probabilistic approach to stress assignment within the Bayesian framework has recently been considered in Russian (Jouravlev & Lupker, 2015). According to a Bayesian model of stress

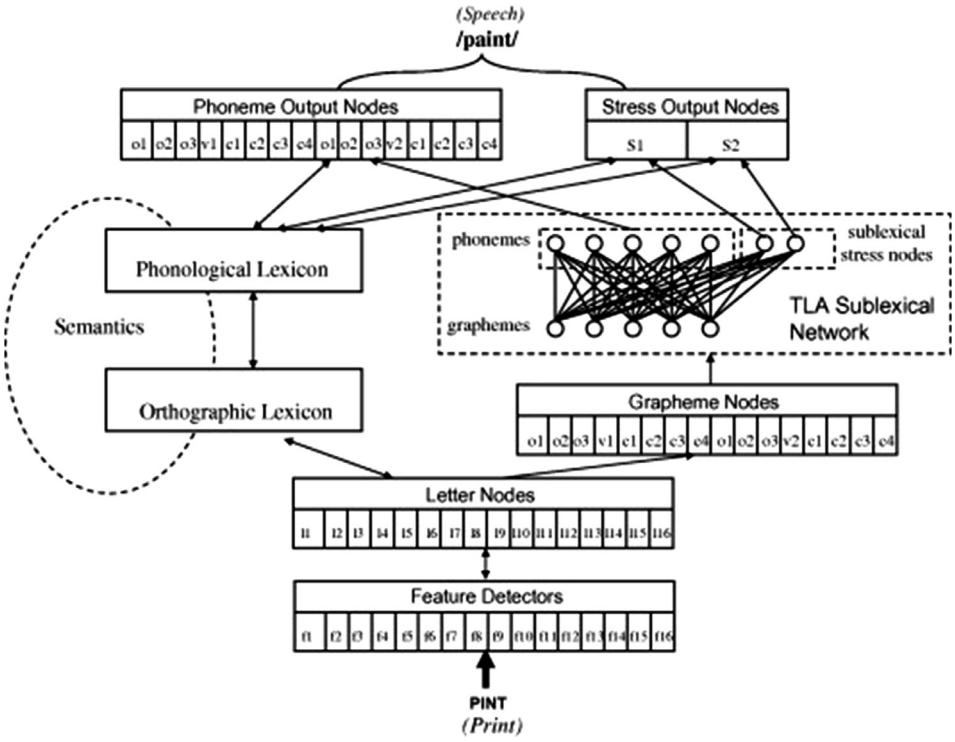

Fig. 2. The CDP++ model (Perry et al., 2010).

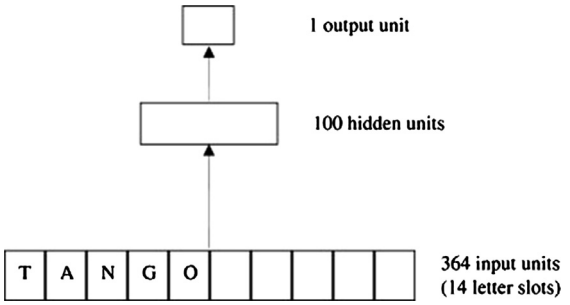

Fig. 3. The Ševa et al. (2009) network of stress assignment.

assignment in reading, readers compute probabilities of stress patterns by assessing prior beliefs about the likelihoods of stress patterns in a certain language, and combining this information with sublexical cues to stress that are language-specific. In this approach then, readers are thought to be sensitive to the frequency with which various stress patterns occur in their language (for example, in the English language, approximately 75% of disyllabic words are stressed on the first syllable and 25% are stressed on the second syllable), *and* language-specific sublexical cues to stress that are present in the orthographic input. The idea that skilled readers may integrate sublexical orthographic cues to stress with prior beliefs about stress patterns in their language to decide about stress placement is worth exploring. However, this approach has not been tested so far in the English language, and the language-specific sublexical cues to stress in English are as yet unknown. For this reason, we did not consider this approach in the present study.

In summary, while most empirical and theoretical work on reading aloud has been undertaken using monosyllables, there has been some progress in modeling the reading aloud of disyllables. However, of the three models that propose to simulate disyllabic reading aloud in English, only the CDP++ model is a full processing model that provides a reaction time, pronunciation, and stress marker. The rule-based algorithm of Rastle and Coltheart (2000) provides a pronunciation and stress marker, and the distributed-connectionist model of Ševa et al. (2009) provides only a stress marker. Similarly, just as the number of available models of reading aloud that can handle disyllables is small, the empirical database needed to test these models is very limited. For example, the few large-scale studies on English disyllabic reading aloud include only monomorphemic words (Chateau & Jared, 2003; Yap & Balota, 2009), which constitute less than 30% of disyllabic words in the English language (CELEX; Baayen, Piepenbrock, & Gulikers, 1995). Further, neither of these large-scale reading aloud studies investigated how people assign stress to disyllables. Most importantly, the only available dataset of disyllabic nonword reading that can be used to test generalization performance in models of reading is that of Rastle and Coltheart (2000). However, this dataset is small in size, the nonwords used have critical limitations (as outlined below), and their characteristics are not fully representative of the characteristics of English disyllabic words.

Cues to stress assignment in English

The fact that people are relatively consistent when assigning stress to disyllabic nonwords (Rastle & Coltheart, 2000) suggests that stress can be computed along the sublexical pathway. One possibility that has been considered is that the sublexical pathway simply assigns a default stress pattern corresponding to the most frequent stress pattern in a given language (Colombo, 1992; Colombo & Tabossi, 1992). For example, because most Italian words are stressed on the penultimate syllable (e.g., tac**chi**na), Italian nonwords may also attract stress on that syllable; similarly, because approximately 75% of English disyllabic words are stressed on the first syllable, English disyllabic nonwords may also attract first-syllable stress. On this type of distributional rule, words that follow the rule are considered 'regular' while words that do not follow the rule are considered 'irregular'. Colombo (1992) and Colombo and Tabossi (1992) found that Italian words stressed on the penultimate syllable were read aloud more quickly than words stressed on the antepenultimate syllable. However, this effect was limited to low-frequency words, while Burani and Arduino (2004) failed to observe such an effect. Other studies in Italian also failed to observe a reading aloud latency advantage for 'regularly-stressed' words compared to 'irregularly-stressed' words (see Colombo & Zevin, 2009; Sulpizio, Job, & Burani, 2012). Hence, the available empirical evidence from Italian does not support the notion that knowledge about the distribution of stress patterns in a language influences stress assignment (see also review by Sulpizio, Burani, & Colombo, 2015). Accordingly, Rastle and Coltheart (2000) did not find support for the distributional rule in English word reading aloud; English disyllabic words stressed on the first syllable were read aloud as fast as those stressed on the second syllable. Further, though most English words take first-syllable stress, there were instances in which people routinely assigned second-syllable stress to nonwords (e.g., emvoke → /ɛmˈvəʊk/), arguing against the notion that the sublexical pathway implements a default stress pattern (Rastle & Coltheart, 2000).

Various other kinds of proposals have been put forward as to the sublexical cues that readers use to assign stress. These include phonological cues such as vowel length (i.e., the syllable with the long vowel takes stress; Baker & Smith, 1976) and phonological weight (i.e., syllables with more phonemes draw stress; Guion, Clark, Harada, & Wayland, 2003). Research has also suggested that there might be strong orthographic cues to stress assignment. Kelly, Morris, and Verrekia (1998) and Kelly (2004) suggested that syllables with many letters tend to attract stress, in particular in those cases where these letters are redundant for pronunciation. For example, the word 'succumb' does not require the final 'b', and the word 'gazelle' does not require the final 'le', because both words would be pronounced in the same way had they not contained these 'extra' letters. Similarly, Arciuli and Cupples (2006) analyzed 340 orthographic word endings and found clear correlates of stress assignment to which readers appear to be sensitive. In addition to orthographic cues, Rastle and Coltheart (2000) have argued that morphological cues provide an important cue to stress, namely, that prefixes and suffixes typically repel stress, except for a handful of 'stress-taking suffixes' (e.g., –een, –ique, –oo) as defined by Fudge (1984). Finally, Burani and Arduino (2004) argued that an important cue to stress assignment is the stress neighborhood in which a stimulus resides. In Italian, the stress neighborhood is defined as those words that share the nucleus of their penultimate syllable and their final syllable, such as the –ola in 'pistóla' (see Sulpizio & Colombo, 2013). Burani and Arduino (2004) showed that irrespective of whether an Italian word had the typical penultimate stress pattern, reading aloud latency was influenced by the consistency of a stimulus within its stress neighborhood. Words with many stress 'friends' were read aloud more quickly than words with many stress 'enemies'. However, these cues have often been confounded in investigations of stress assignment in the literature. Hence, it is unclear whether readers take into account some or all of these cues when assigning stress to disyllables. Determining the cues to stress assignment in English and assessing their relative importance is critical for understanding how people read the vast majority of words in the English lexicon, and for evaluating the next generation of computational models of reading. Our study is the first to examine the combined influence of several cues on stress assignment and their relative importance while taking into account potential confounds of previous studies.

The present study

The present article reports a large-scale study in which 41 participants read aloud 915 disyllabic nonwords. For each response, we report stress placement, phonemic transcription, and reaction time. One major aim of the study was to determine the factors that influence stress, pronunciation, and naming latency, and thus to learn more about the operation of the sublexical pathway in reading disyllables aloud. We anticipated that like similar studies of monosyllabic nonword reading (Andrews & Scarratt, 1998; Pritchard et al., 2012), we would uncover a range of permissible responses for each nonword. We quantified this pronunciation variability using the *H*-statistic (e.g., Shannon, 1949; Treiman, Mullennix, Bijelac-Babic, & Richmond-Welty, 1995), a measure of entropy that takes into account the proportion of participants producing each alternative pronunciation.

The other major aim of the study was to test the adequacy of the three computational models of disyllabic reading aloud discussed previously – the CDP++ model (Perry et al., 2010), the rule-based algorithm of Rastle and Coltheart (2000), and the distributed-connectionist model of stress assignment (Ševa et al., 2009) – against the obtained human data. Our study is the first to provide stress, pronunciation, and naming latency data on a large number of disyllabic nonwords whose characteristics are representative of the characteristics of English disyllabic words in the lexicon. Further, our dataset forms the largest corpus of nonword pronunciation data generated to date. Hence, it will be critical for assessing generalization perfor-

mance in future developments of computational models of reading.

Method

Participants

Forty-one monolingual native speakers of Southern British English (11 males) participated in the study for £20. Participants were undergraduate students at Royal Holloway, University of London. They were 21 years old on average (range [19,27], $SD = 1.6$), had normal or corrected-to-normal vision, and reported no reading impairments.

Materials

The experimental stimuli consisted of 915 disyllabic nonwords. These nonwords were derived by submitting a subset of disyllabic words from the English Lexicon Project (Balota et al., 2007) to the Wuggy algorithm (Keuleers & Brysbaert, 2010), a program that generates nonwords by maintaining the subsyllabic structure and transition frequencies of the base words. The nonwords that we selected for the present study were 5–8 letters long, with an average neighborhood size of 0.5, taking into account substitution, addition, and deletion neighbors (CLEARPOND database; Marian, Bartolotti, Chabal, & Shook, 2012). Nonword stimuli had an average orthographic weight of 0.9 (orthographic weight was defined as a ratio of number of letters in the first syllable to number of letters in the second syllable).

To overcome limitations of previous studies (Pritchard et al., 2012), we ensured that our stimuli had characteristics that mirrored those of disyllabic English words. According to the CELEX database (Baayen et al., 1995), 71% of disyllabic words in the English language are morphologically complex. Hence, to approximate the morphological structure of English disyllables, 60% of our nonwords contained affixes (546 items, of which 396 were prefixed and 150 were suffixed). We further ensured that only genuine affixes were used by (a) selecting from the ELP (Balota et al., 2007) disyllabic base words that were denoted as morphologically-complex; (b) verifying the status of these stimuli as comprising 'stem + affix' or 'affix + stem' using the CELEX database (Baayen et al., 1995); and (c) selecting nonwords generated by Wuggy that preserved the affix of the base word. The prefixed and suffixed items in our set are indicated in the supplementary materials.

We subsequently ran the 915 nonwords through the CDP++ model (Perry et al., 2010), the rule-based algorithm for pronouncing disyllables (Rastle & Coltheart, 2000), and the distributed-connectionist network (Ševa et al., 2009). The CDP++ model and the rule-based algorithm (hereafter referred to as RC00) agreed on pronunciation and stress assignment for 21% of the nonwords ($N = 192$), and disagreed on pronunciation and stress assignment for 20% of the nonwords ($N = 183$). These models agreed on pronunciation only for 13% of the nonwords ($N = 123$), and on stress assignment only for 46% of the nonwords ($N = 417$). The Ševa et al. (2009, hereafter referred to as SMA09) model produced the same stress as the RC00 algorithm for 83% of the nonwords ($N = 756$); it produced the same stress as the CDP++ model for 77% of the nonwords ($N = 702$). The 915 items and their corresponding pronunciations and stress assignments by the three models are provided in the supplementary materials.

Design and procedure

Eight experimental lists were created. The order of presentation of the nonwords in each list was pseudo-randomised based on the following criteria: (a) nonwords on two successive trials were not phonologically related (i.e., they did not share their initial phoneme or rime), and (b) affixed and non-affixed nonwords were equally distributed across each experimental list. Participants were randomly assigned to one of the eight experimental lists.

Participants were tested individually, seated in front of a CRT monitor in a quiet room. Stimulus presentation and data recording were controlled via DMDX software (Forster & Forster, 2003) and verbal responses were recorded by a head-worn microphone. Each trial started with a fixation cross displayed at the center of the computer screen for 500 ms. Nonwords were presented in the same position, in lowercase (Courier New, 14-point font), in white on a black background, and remained on the screen for 3000 ms or until participants responded, whichever happened first. The inter-stimulus interval randomly varied between 400 and 600 ms.

The experiment consisted of five practice trials followed by five blocks of 183 nonwords each. A break was administered between the blocks. Participants were instructed to read aloud the nonwords as naturally as possible, as if they were real words, at their own pace and without hesitation. An experimenter monitored participants' performance throughout the session. In addition to the main experiment, we collected demographic information and standard measures of participants' vocabulary, spelling, and reading ability to ensure that participants did not have language impairments, which could potentially affect their reading performance. The whole session, including breaks, lasted approximately 90 min.

Data preparation

Reaction times, stress, and pronunciation

The main experiment generated 37,515 digitized sound files (915 items $*$ 41 participants). If any one of the three main measures – stress, pronunciation, or reaction times (RTs) – could not be obtained, then the trial as a whole was excluded from any further analysis. This process yielded the exclusion of 693 trials (1.8% of the data).

The acoustic onsets of verbal responses were hand marked using CheckVocal (Protopapas, 2007), following the criteria specified by Rastle, Croot, Harrington, and Coltheart (2005). In addition, the full set of sound files were phonetically transcribed by one of the authors (R. L.), who had been trained to use the coding scheme of the RC00 algorithm and the CDP++ model for expressing

English phonemes. To facilitate transcription of this very large dataset, the transcriber was provided with the pronunciations produced by the CDP++ model and the RC00 algorithm for each item (note that for 315 of the items the models produced the same pronunciations). The transcriber's task was to record whether the speaker's pronunciation matched the pronunciation of the CDP++ model or that of the RC00 algorithm (or both, in cases where models agreed on pronunciation), or was a different pronunciation. In the latter case, the transcriber documented the alternative pronunciation. The transcriber was naïve to the purposes of the study.

Stress judgements were made by one of the authors (K.R.) who had previous training and experience in this task (Rastle & Coltheart, 2000). Responses where primary stress could not be detected (e.g., where nonwords were given two stressed syllables) were coded as missing. Such responses were also non-fluent (i.e., produced with clear syllabification). Even though the rater was blind to the predictions of the CDP++ and RC00 models during transcription, we sought to verify her judgments using an independent measure of stress for a proportion of trials. Stressed syllables have been associated with longer vowel duration, higher pitch, and greater intensity than unstressed syllables (Lehiste, 1970), though the relative importance of each of these cues to stress perception is still unknown. More recently, Kochanski, Grabe, Coleman, and Rosner (2005) conducted a study on a large corpus of natural speech, which demonstrated that prominent syllables were louder and longer compared to non-prominent syllables. Of the two acoustic cues, loudness (intensity) was the stronger predictor. For this reason, a trained research assistant, who was naïve to the purposes of the experiment, labeled the acoustic boundaries of the vowels in each syllable for a random subset of nonword responses (5.2%, 1932 trials) using the criteria established in the ANDOSL database (Croot, Fletcher, & Harrington, 1992). The vowel intensity values were subsequently extracted using Praat (Boersma, 2001). The syllable with the highest vowel intensity value within each nonword was assigned primary stress. When comparing the stress judgements of the rater to those yielded by the acoustic measure, the match between the two was 79%. In order to reassure ourselves of the quality of the human ratings, we inspected all instances of disagreement between the human ratings and the acoustic measure. In all of these cases, we judged that the human ratings provided a better approximation to the actual stress position.

We also assessed K.R.'s stress judgements against the stress judgements of 10 undergraduate students who had no previous training or experience of similar nature. In particular, these students listened to the reading aloud responses of 10 out of our 41 participants, with each student listening to the responses of a different participant. Despite the fact that the students were untrained, the kappa inter-rater reliability between K.R. and the students (listeners) was very high in most cases, with values between 0.81 and 1 indicating 'almost perfect agreement' (4 cases), values between 0.61 and 0.80 indicating 'substantial agreement' (5 cases), and values between 0.21 and 0.40 indicating 'fair agreement' (1 case). All values were highly significant ($p < .001$). The kappa inter-rater reliability values between the stress judgements made by K.R. and those made by each listener are reported in Table 1.

Mean reaction time and modal human stress and pronunciation for each of the 915 nonwords, as well as the full dataset of every response are provided in the supplementary materials.

Results

We analyzed stress placement, pronunciation, and reaction time for the human data and (where possible) for the three computational models under investigation: CDP++ (Perry et al., 2010), RC00, and SMA09. These three models vary substantially in their completeness, with only CDP++ yielding pronunciation, stress assignment, and reaction time (measured in cycles). The RC00 algorithm outputs pronunciation and stress assignment but no reaction time, and the SMA09 network only outputs stress assignment. Therefore, the analyses that we were able to conduct on each model varied substantially. The outputs of the CDP++ model, the RC00 algorithm, and the SMA09 network, for each stimulus, are provided in the supplementary materials.

Stress assignment

We conducted two sets of analyses on the stress data. In the first set, we used chi-square analyses to assess the success of each model in capturing the modal stress assigned to each item across participants. In the second set, we conducted logistic regression analyses at the item level to probe the factors that influence stress assignment in human readers. These same analyses were then conducted on the model responses to determine whether the models were sensitive to the same factors as human readers. Thus, the analyses allowed us to determine whether the three models under investigation successfully predicted modal human stress.

Model performance in capturing human stress assignment

The modal stress produced by human participants was calculated for each item, and these values were compared with the stress placements provided by each model. Participants produced 77% of the items with first-syllable stress and 23% of the items with second-syllable stress, thus mirroring the distribution of stress in the English language. Eight items were produced with first- and second-syllable stress an equal number of times. Table 2 shows model performance in assigning stress against the modal human stress. Percentages are derived by dividing counts by 907 items, in which participants favoured first- or second-syllable stress.

CDP++. The CDP++ model stressed 81% of the items in line with the modal stress given by participants, a result that was highly significant, $\chi^2(1) = 237.55$, $p < .001$. Importantly, although the model has a strong bias toward initial stress, it also performed reasonably well for those items

Table 1
Kappa (k) inter-rater reliability values between stress judgements of K.R. and stress judgements of 10 listeners.

Listener	1	2	3	4	5	6	7	8	9	10
k	.81	.37	.95	.79	.87	.92	.66	.75	.76	.78

Table 2
Counts of items (with percentages in parentheses) that predominantly received first- and second-syllable stress by human readers and were correctly assigned first- and second-syllable stress by CDP++, RC00, and SMA09.

	Human modal stress	CDP++	RC00	SMA09
1st syllable	694 (77%)	588 (65%)	508 (56%)	555 (61%)
2nd syllable	213 (23%)	148 (16%)	150 (17%)	161 (18%)
Total model match		736 (81%)	658 (73%)	716 (79%)

that were given second-syllable stress by most participants.

In addition to reporting the ultimate stress given to a disyllabic stimulus, CDP++ reports the activation levels of the stress nodes for the first and second syllables at the point at which the stimulus is read aloud. Using these activations, we computed a measure of *CDP++ stress certainty*, consisting of the absolute value of the difference between the activations of the two stress nodes. The higher the value of this measure the greater the model's certainty of the given stress. Similarly, we computed human stress certainty, expressed as the absolute value of the difference between the percentage of people that assigned 1st-syllable stress and 2nd-syllable stress to each item, with a bigger difference reflecting higher stress certainty for a certain item. We then compared the CDP++ stress certainty measure against human stress certainty and found a strong positive correlation between them, $r = .56$, $p < .001$. In other words, items that yielded high certainty about which syllable to stress across the 41 participants also yielded high stress certainty in the model.

RC00. The RC00 algorithm stressed 73% of the items according to the modal stress given by human participants, a result that was highly significant, $\chi^2(1) = 132.98$, $p < .001$. A close inspection of the stress errors made by the model revealed that these were mainly due to producing second-syllable stress in response to prefixed nonwords.

SMA09. The SMA09 network stressed 79% of the items according to the modal stress given by participants, a result that was highly significant, $\chi^2(1) = 227.26$, $p < .001$. A close inspection of the stress errors made by SMA09 revealed that these were mainly due to incorrectly assigning second-syllable stress to prefixed nonwords.

The SMA09 network outputs both stress placement and the activation of the stress node scaled between 0 and 1, with values above 0.5 denoting second-syllable stress and values below 0.5 denoting first-syllable stress. As for the CDP++ model, we computed stress certainty in SMA09 and compared it with stress certainty across participants. Model stress certainty for items with second-syllable stress was the activation value of the stress node; for items with first-syllable stress, it was 1 minus the activation value of the stress node. For example, the item 'abast' yielded a stress value of 0.998, which denotes high certainty (in this case, for second-syllable stress). The item 'burbam' yielded a stress value of 0.002. We calculated the network's stress certainty for this item as $1 - 0.002 = 0.998$, which denotes high certainty (in this case, for first-syllable stress). Analyses of the model and human stress certainty measures revealed a strong positive correlation between them, $r = .44$, $p < .001$; hence, items that yielded high certainty about which syllable to stress across participants also yielded high stress certainty in the SMA09 network.

Factors influencing stress assignment

In order to determine the factors that influence human stress assignment, we conducted logistic regression analyses at the item level, with modal human stress as the outcome variable (first- versus second-syllable stress). The predictors consisted of variables that are thought to influence stress assignment. These include the orthographic weight (Kelly, 2004; Kelly et al., 1998) and vowel length (Baker & Smith, 1976) of a syllable as well as the stress pattern of word neighbors (Burani & Arduino, 2004; Sulpizio & Colombo, 2013).[1]

Morphology has also been suggested as a cue to stress assignment in English (Ktori, Tree, Mousikou, Coltheart, & Rastle, 2016; Rastle & Coltheart, 2000). However, this cue is unavoidably confounded with other potentially important cues to stress, including orthographic and phonological weight, and vowel length. In particular, items containing prefixes tend to have more letters and phonemes in the second syllable than the first, while the vowel in the second syllable is likely longer than the vowel of the prefix. Similarly, items containing suffixes tend to have more letters and phonemes in the first syllable than the second, while the vowel in the first syllable is likely longer

[1] As we mentioned in the introduction, the phonological weight of a syllable is also known to influence stress assignment (Guion et al., 2003). However, we cannot reliably count the number of phonemes in our set of nonwords, because each nonword may receive a number of alternative pronunciations, and so the number of phonemes in each syllable varies as a function of its pronunciation. Also, vowel length is typically confounded with number of vowel letters in English (i.e., long vowels are usually spelled with two letters). This was also the case in our study, so it could be that any observed effects of vowel length on stress assignment denote effects of number of vowel letters.

than the vowel of the suffix. Such confounds were apparent in the stimuli used by Rastle and Coltheart (2000). For example, in their study, all prefixed nonwords (49 in total) had more letters in the second syllable compared to the first, while the majority of suffixed nonwords (48/88) had more letters in the first syllable compared to the second. Because the nonwords in our study were modeled on English disyllabic words, we hypothesized that prefixation and suffixation were likely confounded with other sublexical cues to stress. For this reason, we first examined the relationship between affixation and orthographic weight in our items. We observed that all except one of our prefixed nonwords had more letters in the second syllable than the first while 69% of our suffixed nonwords had more letters in the first syllable than the second.[2] Similarly, we examined the relationship between affixation and vowel length in our items; while almost half of the prefixed nonwords (48%) had a long vowel in the second syllable, only 25% of the suffixed nonwords had a long vowel in the second syllable.

To avoid potential confounds in our analyses while still being able to assess how morphological properties of printed letter strings may influence stress assignment, we calculated a graded metric of spelling-to-stress consistency that expressed the consistency with which orthographic onsets and rimes in the first and second syllable map to a particular stress pattern in the lexicon. We considered that this metric would capture morphological effects on stress because (a) the CELEX database of disyllables that we used to calculate this metric comprises a very high proportion of morphologically-complex words (Baayen et al., 1995), and (b) the syllabic units in our affixed nonwords always corresponded to a prefix or a suffix. Further, the hypothesis that prefixes and suffixes repel stress (Rastle & Coltheart, 2000) is derived from the observation that the vast majority of prefixes and suffixes in English words do not take stress. Thus, we predicted that the sublexical orthographic units in the first syllable of prefixed items (i.e., the onset and rime units that make up prefixes) would likely map consistently onto second-syllable stress, while the sublexical orthographic units in the second syllable of suffixed items (i.e., the onset and rime units that make up suffixes) would likely map consistently onto first-syllable stress.

The calculation of this metric required us to syllabify our items. There is no consensus on the syllabification of English words (Treiman & Danis, 1988). However, in our case, we wanted to derive sublexical units that would allow us to compare our nonwords to similar words in the CELEX database (Baayen et al., 1995). For this reason, we closely followed the method for orthographic syllabification used in CELEX. These syllabifications generally follow the Maximum Onset Principle (Kahn, 1976; Selkirk, 1982), whereby the phonological onset of the second syllable is maximized (e.g., kelvin → kel-vin instead of kelv-in). However, orthographic constraints are also respected (e.g., afford → af-ford instead of a-fford). Following these syllabifications allowed us to derive onset and rime units from our nonwords that matched the onset and rime units of similar words in CELEX, thus increasing the validity of our spelling-to-stress consistency metric.

In the following logistic regression analyses, 'first-syllable stress' was considered as the baseline category, so that regression coefficients reflect the probability of assigning second-syllable stress.[3] Accordingly, our predictors included the measure of spelling-to-stress consistency of the first and the second syllable described above (which reflects whether the onset and rime units of each syllable point toward first- or second-syllable stress), the item's orthographic weight (expressed as the ratio of number of letters in the first syllable to the number of letters in the second syllable), the vowel length of the second syllable (which denotes whether the second syllable contains a long vowel), and the stress pattern of the item's orthographic neighbors (which denotes whether an item has any word neighbors that are stressed on the second syllable). Further, we considered stress certainty as an additional predictor of stress assignment.

In order to determine then whether stress assignment in the models was sensitive to the same factors as in the human data, the same logistic regression analyses were carried out for each model. These analyses treated model stress as the outcome variable (again, with first-syllable stress being the baseline category) and used the same predictor variables as in the human data.[4] All analyses of stress placement used the glm function in R (R Core Team, 2015, version 3.2.3) and the packages car (Fox & Weisberg, 2011) and mlogit (Croissant, 2013). Further, given that we had no logical or theoretical basis for considering any variable to be prior to any other, we entered all of the variables into the analysis simultaneously. Our predictions about how each of these variables was likely to influence stress assignment are outlined below:

i. *Spelling-to-stress consistency*. We calculated two variables, one for the onset and rime units of the first syllable and another for the onset and rime units of the second syllable for all 915 nonwords. These variables expressed the consistency with which these units in each syllable map to a particular stress pattern according to the CELEX database (Baayen et al., 1995). The values for onsets and rimes were then averaged to form a composite measure of unit consistency within each syllable. Our hypothesis was that the more consistently the onset and rime units in a syllable map onto a particular stress pattern, the more likely it would be for that syllable to have this

[2] Indeed, after extracting all of the disyllabic words in CELEX that had either a prefix or a suffix contained in one or more of our nonwords, we observed that 95% of those prefixed words had more letters and phonemes in the second syllable than the first, while 98% of those suffixed words had more letters and phonemes in the first syllable than the second.

[3] Eight of the nonwords were assigned first- and second-syllable stress an equal number of times. These items were excluded from the analyses of both the human and the model stress data.

[4] Stress certainty in the CDP++ model consisted of the absolute value of the difference between the activations of the model's stress nodes. Stress certainty in the SMA09 network was also calculated on the basis of the stress node's activation value (see details of its calculation on p. 19). In both models, higher values indicated greater stress certainty.

stress pattern. Values toward 1 pointed toward first-syllable stress while values toward 0 pointed toward second-syllable stress. In line with our predictions, the average spelling-to-stress consistency of the first syllable of prefixed items in our set was 0.40 (thus, showing bias toward second-syllable stress) while the corresponding value of the second syllable of suffixed items was 0.80 (thus, showing bias toward first-syllable stress). The average spelling-to-stress consistency for non-affixed items in our set was 0.80 for both syllables (indicating strong bias toward first-syllable stress). Further details on the calculation of this metric can be found in Appendix A.

ii. *Orthographic weight*. We calculated a metric expressing orthographic weight by dividing the number of letters in the first syllable by the number of letters in the second syllable. Values below 1 indicated more letters in the second syllable, while values above 1 indicated more letters in the first syllable. We hypothesized that the more letters a syllable contained the more likely it would be for that syllable to attract stress (see Kelly, 2004; Kelly et al., 1998).

iii. *Long vowel (second syllable)*. We created a binary variable, which expressed for each nonword whether there is likely to be a long vowel in the second syllable. This was based on a rule-based pronunciation of the vowel orthography (e.g., vowel digraphs and split digraphs would normally yield a long vowel). In the analysis, the category 'no long vowel in the second syllable' was treated as the baseline. Our hypothesis was that long vowels in the second syllable would tend to attract stress (see Baker & Smith, 1976).

iv. *Orthographic neighbor with second-syllable stress*. We created a binary variable, which expressed for each nonword whether it has an orthographic word neighbor that is stressed on the second syllable. The calculation of this metric included addition, deletion, and substitution neighbors. In this analysis, the category 'no neighbor with second-syllable stress' was treated as the baseline. We hypothesized that nonwords that had one or more neighbors pronounced with second-syllable stress would be more likely to attract second-syllable stress than nonwords with no such neighbors (see Burani & Arduino, 2004; Sulpizio & Colombo, 2013).

v. *Stress certainty*. We created a continuous variable by calculating the absolute difference between the percentage of people that assigned 1st-syllable stress and 2nd-syllable stress to each nonword. A bigger difference reflected higher stress certainty for a certain item. We hypothesized that nonwords that yielded less stress certainty may be more likely to take 2nd-syllable stress. This is because 2nd-syllable stress is not the default stress pattern in English and so readers are likely less exposed to sublexical cues that are typically associated with 2nd-syllable stress in the English lexicon (and hence more uncertain when assigning 2nd-syllable stress).

Some general characteristics of the nonwords that are relevant for the regression analyses on the stress data are shown in Table 3. The results from the analyses of the stress data are shown in Table 4.

We further assessed the relative importance of individual predictors in the model. Relative importance refers to the quantification of an individual regressor's contribution to a multiple regression model. The varImp metric of the caret package in R (Kuhn, 2015) is a statistical test that assesses the relative importance of individual predictors in the model by estimating the absolute value of the *t*-statistic for each model parameter. This function automatically scales the importance scores to be between 0 and 100. Table 5 presents the results from this analysis.

Humans. The analyses of the human stress data were all in the predicted direction. In particular, second-syllable stress becomes less likely when (i) the onsets and rimes of both syllables are consistently associated with first-syllable stress; (ii) the orthographic weight is biased toward the first syllable; and (iii) there is more stress certainty. Similarly, second-syllable stress becomes more likely when (i) there is a long vowel in the second syllable; and (ii) the nonword has one or more orthographic neighbors with second-syllable stress.

The odds ratios further showed that for a unit increase in the spelling-to-stress consistency of the onset and rime units in the first syllable, which indicates more association with first-syllable stress, there was a 97% decrease in the odds of 2nd-syllable stress assignment. The equivalent decrease for the onset and rime units in the second syllable was 88%. Similarly, for a unit increase in the orthographic weight, which indicates more letters in the first-syllable, there was a 99% decrease in the odds of 2nd-syllable stress assignment. However, for a unit increase in stress certainty there was only a 1% decrease in the odds of 2nd-syllable stress assignment. The odds of 2nd-syllable stress assignment for items with a long vowel in the second syllable were 149% higher than the odds for items without a long vowel in the second syllable, while the odds of 2nd-syllable stress assignment for items with 2nd-syllable stressed neighbors were 193% higher than the odds for items without such neighbors.

Table 3
Characteristics of nonwords (ranges for means are shown in parentheses).

	Total
Prefixed	396
Suffixed	150
Orthographic weight	0.9 (0.2–5)
Long vowel 2nd syllable	314
Neighbors stressed on 2nd-syllable	64
Neighborhood	0.5 (0–12)
	Mean
Onset and rime spelling-to-stress consistency (1st syllable)	0.7 (0.1–1)
Onset and rime spelling-to-stress consistency (2nd syllable)	0.7 (0.2–1)
Alternative pronunciations	5.9 (1–22)

Table 4
Logistic regression analyses on stress data for humans, CDP++, RC00, and SMA09.

	Humans		CDP++		RC00		SMA09	
	B(*SE*)	*OR*	*B*(*SE*)	*OR*	*B*(*SE*)	*OR*	*B*(*SE*)	*OR*
Spelling-to-Stress Consistency (1st syllable)	−3.52*** (0.58)	0.03	−3.39*** (0.54)	0.03	−4.11*** (0.45)	0.02	−3.07*** (0.51)	0.05
Spelling-to-Stress Consistency (2nd syllable)	−2.15*** (0.59)	0.12	−2.29*** (0.56)	0.10	−2.37*** (0.49)	0.09	−1.61** (0.54)	0.20
Orthographic Weight	−4.46*** (0.66)	0.01	−2.57*** (0.53)	0.08	−0.24 (0.24)	0.79	−3.43*** (0.50)	0.03
Vowel Length	0.91*** (0.23)	2.49	1.74*** (0.22)	5.72	−0.02 (0.19)	0.98	0.52* (0.21)	1.68
Neighbors with 2nd syllable stress	1.07** (0.35)	2.93	0.12 (0.34)	1.13	0.76* (0.35)	2.13	2.20*** (0.47)	8.99
Stress Certainty	−0.01* (0.00)	0.99	−1.02** (0.36)	0.36			−0.83 (0.64)	0.44
R^2 (Hosmer–Lemeshow)	.45		.44		.26		.42	
Chi-square statistic	$\chi^2(6) = 447.25$		$\chi^2(6) = 469.52$		$\chi^2(5) = 307.35$		$\chi^2(6) = 479.84$	

Note. *B*(*SE*): Unstandardized coefficients with Standard Errors in parentheses; OR: Odds Ratios.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 5
Analyses of variable importance on stress data for humans, CDP++, RC00, and SMA09.

	Humans	CDP++	RC00	SMA09
Spelling-to-Stress Consistency (1st syllable)	83.68	78.18	100.00	86.03
Spelling-to-Stress Consistency (2nd syllable)	29.03	49.16	51.91	30.73
Orthographic Weight	100.00	59.02	9.81	100.00
Vowel Length	37.62	100.00	0.00	21.70
Neighbors with 2nd syllable stress	17.31	0.00	22.84	60.49
Stress Certainty	0.00	32.45		0.00

The analysis of variable importance further showed that the most important predictor in the regression model on the human stress data was orthographic weight, followed by the spelling-to-stress consistency of the onset and rime units in the first syllable. Vowel length of the 2nd syllable was the next most important predictor, followed by the spelling-to-stress consistency of the onset and rime units in the second syllable. The stress neighborhood of the second syllable was a less important predictor in the model while stress certainty had no importance. Collinearity was not an issue as variance inflation factors (VIF) for all predictors in the model were less than 1.14 while the lowest observed value of the tolerance statistic (1/VIF) was 0.88.

CDP++. The analyses of the CDP++ stress data revealed the same effects as the human data except that having an orthographic neighbor with 2nd-syllable stress did not influence stress assignment in this model. The odds ratios further showed that for a unit increase in the spelling-to-stress consistency of the onset and rime units in the first syllable, which indicates more association with first-syllable stress, there was a 97% decrease in the odds of 2nd-syllable stress assignment. The equivalent decrease for the onset and rime units in the second syllable was 90%. Further, for a unit increase in the orthographic weight, which indicates more letters in the first-syllable, there was a 92% decrease in the odds of 2nd-syllable stress assignment, while for a unit increase in stress certainty, there was a 64% decrease in the odds of 2nd-syllable stress assignment. Also, the odds of 2nd-syllable stress assignment for items with a long vowel in the second syllable were 472% higher than the odds for items without a long vowel in the second syllable. The analysis of variable importance showed that the most important predictor in the regression model was the vowel length of the 2nd syllable, followed by the spelling-to-stress consistency of the onset and rime units in the first syllable. Orthographic weight was the next most important predictor followed by the spelling-to-stress consistency of the onset and rime units in the second syllable. Stress certainty was a less important predictor while the stress neighborhood of the second syllable had no importance. Collinearity was not an issue as variance inflation factors (VIF) for all predictors in the regression model were less than 1.32 while the lowest observed value of the tolerance statistic (1/VIF) was 0.76.

RC00. The analyses of the RC00 stress data revealed that second-syllable stress becomes less likely when the onsets and rimes of both syllables are consistently associated with first-syllable stress, and that second-syllable stress becomes more likely when the nonword has one or more orthographic neighbors with second-syllable stress. The odds ratios further showed that for a unit increase in the spelling-to-stress consistency of the onset and rime units in the first syllable, which indicates more association with first-syllable stress, there was a 98% decrease in the odds of 2nd-syllable stress assignment. The equivalent decrease for the onset and rime units in the second syllable was 91%. Further, the odds of 2nd-syllable stress assignment for items with 2nd-syllable stressed neighbors were 113% higher than the odds for items without such neighbors. The analysis of variable importance showed that the most important predictor in the regression model on the RC00 stress data was the spelling-to-stress consistency of the onset and rime units in the first syllable, followed by the

spelling-to-stress consistency of the onset and rime units in the second syllable. We believe that this effect was picked up the RC00 algorithm, because the spelling-to-stress consistency metric captures morphological effects on stress and prefixation and suffixation are hard-coded in the model, so that both prefixes and suffixes repel stress. The stress neighborhood of the second syllable was the next most important predictor while orthographic weight and the vowel length of the second syllable had minimal or no importance. Collinearity was not an issue as variance inflation factors (VIF) for all predictors in the model were less than 1.53 while the lowest observed value of the tolerance statistic (1/VIF) was 0.65.

SMA09. The analyses of the SMA09 stress data revealed the same effects as the human data except that stress certainty in the model did not influence stress assignment. The odds ratios further showed that for a unit increase in the spelling-to-stress consistency of the onset and rime units in the first syllable, which indicates more association with first-syllable stress, there was a 95% decrease in the odds of 2nd-syllable stress assignment. The equivalent decrease for the onset and rime units in the second syllable was 80%. Further, for a unit increase in the orthographic weight, which indicates more letters in the first-syllable, there was a 97% decrease in the odds of 2nd-syllable stress assignment. The odds of 2nd-syllable stress assignment for items with a long vowel in the second syllable were 68% higher than the odds for items without a long vowel in the second syllable, while the odds of 2nd-syllable stress assignment for items with 2nd-syllable stressed neighbors were 799% higher than the odds for items without such neighbors. The analysis of variable importance showed that similarly to the human data, the most important predictor in the regression model was orthographic weight, followed by the spelling-to-stress consistency of the onset and rime units in the first syllable. The stress neighborhood of the second syllable was the next most important predictor followed by the spelling-to-stress consistency of the onset and rime units in the second syllable. Vowel length of the 2nd syllable was a less important predictor while stress certainty had no importance. Collinearity was not an issue in the model as variance inflation factors (VIF) for all predictors were less than 1.22 while the lowest observed value of the tolerance statistic (1/VIF) was 0.82.

Taken together, these results show that the SMA09 network was the most successful in predicting modal human stress assignment, followed by the CDP++ model. The RC00 algorithm was the least successful model in predicting modal human stress.

Human stress certainty

Considering modal stress belies variability across items in the consistency with which participants assign stress. For this reason, we carried out additional analyses on the stress certainty with which people assigned first- and second-syllable stress. For each item, we calculated the percentage of people that assigned 1st-syllable and 2nd-syllable stress. The absolute difference between the two percentages reflected stress certainty, with bigger differences reflecting greater stress certainty. For example, the item 'bafeness' was assigned 1st-syllable stress 100% of the time and so its stress certainty was 100. The item 'contonse' was assigned both 1st- and 2nd-syllable stress 50% of the time and so its stress certainty was 0. The item 'abast' was assigned 1st-syllable stress 17.5% of the time and 2nd-syllable stress 82.5% of the time, hence its stress certainty was 65.

The items were then separated into two groups. One group contained the items that received modal 1st-syllable stress and the other group contained the items that received modal 2nd-syllable stress. The items that did not receive modal stress ($N = 8$), which were the items that had stress certainty of 0, were excluded from this analysis. Four stress certainty levels were then created (i.e., very low, low, high, and very high stress certainty), with each level containing the total number of items that received 1st and 2nd modal stress. Fig. 4 shows that the vast majority of the items that yielded high stress certainty tended to be those given 1st-syllable stress, while most 2nd-syllable stressed items yielded lower stress certainty. This result indicates great consistency across participants in reading aloud items with first-syllable stress, but lower consistency across people in producing items with second-syllable stress.

Pronunciation

Two sets of analyses were conducted on the pronunciation data. In the first set of analyses, we assessed how well the CDP++ model and the RC00 algorithm captured the pronunciations given by human readers. In the second set of analyses, we quantified variability in the pronunciations given by human readers through a measure of entropy known as the *H statistic* (Shannon, 1949). We then used regression analyses at the item level to determine the factors that predict variability in pronunciation across participants. In all of the analyses in this section, we consider pronunciation only, without regard to stress assignment. Thus, the calculation of the modal human response treated pronunciations stressed on the first and second syllable as the same, as long as they contained exactly the same phonemes. Similarly, model pronunciations were considered as matching human pronunciations if the same string of phonemes was produced, irrespective of stress assignment.

Model performance in capturing human pronunciation

The analyses that we carried out first sought to determine how well the model pronunciations matched those given by human readers. Table 6 summarizes how often the models' pronunciation matched the 1st most common pronunciation given by participants, the 2nd most common, the 3rd most common and so on. However, in the light of potential difficulty in distinguishing similar-sounding vowels due to accent, which may have resulted in some variability in the transcription of human pronunciations, we additionally created a lenient score (see Table 6). According to this score, all pronunciations of nonwords containing a vowel that had been transcribed as a schwa (e.g., d@nEst for 'danest') were merged with pronunciations of the same nonwords containing a vowel in the same position that had been transcribed as a short vowel (e.g., d

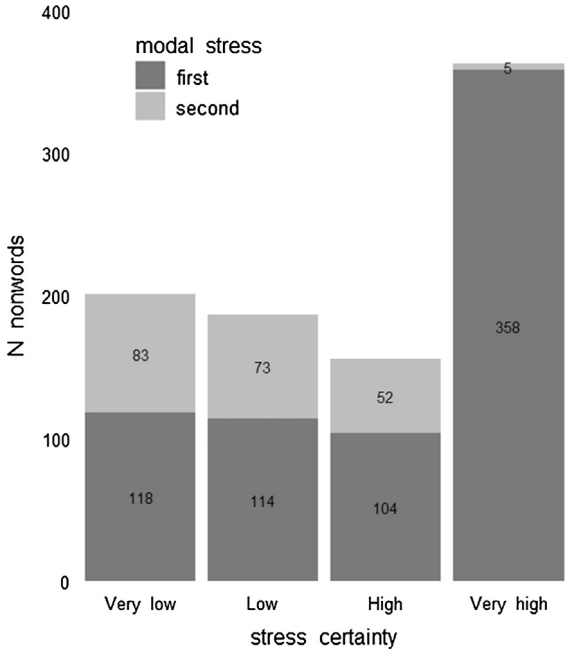

Fig. 4. Number of first- and second-syllable stressed nonwords that yielded different levels of stress certainty.

{nEst). As is shown in the table, when the lenient scoring criteria applied, there was an additional 7% of items whose modal pronunciation the CDP++ model was able to capture and an additional 14% of items whose modal pronunciation the RC00 algorithm could capture. The second last column of the table (titled 'Absent') gives the percentage of items for which the models gave a pronunciation that no human reader gave. These pronunciations are provided in Appendices B and C for the CDP++ model and the RC00 algorithm, respectively.

Typically, models of reading have been assessed relative to average performance (e.g., Coltheart et al., 2001; Perry et al., 2007); for example, their match against the most frequent pronunciation (e.g., Andrews & Scarratt, 1998) or mean reaction time (e.g., Spieler & Balota, 1997). However, it is also possible to treat models not as the average participant, but as an individual participant within the normal distribution (e.g., Pritchard et al., 2012; Robidoux & Pritchard, 2014; Zevin & Seidenberg, 2006). One way of evaluating models within this conceptualization is by assessing the extent to which their pronunciations accord with those of at least one human reader in the sample (Pritchard et al., 2012). Table 6 shows that even when we used a lenient scoring criterion, a substantial portion of pronunciations given by CDP++ and RC00 were not given by any human reader. Close inspection of these pronunciations revealed substantial differences in the kinds of errors produced by the two models. For the CDP++ model, most errors were due to:

(a) unusual vowel shortening in the second syllable (e.g., bomly → bQmlI; ungourt → VNg@t). In some cases, this led to impossible pronunciations, so that both syllables contained schwa vowels, one of which was stressed (e.g., adant → @d@nt);
(b) unlikely print-to-sound mappings (e.g., lictoun → lIst6n; bethove → bITVv);
(c) lexicalisations (e.g., insance → Inst@ns; leavy → lIvIN);
(d) assuming the existence of final –e (e.g., astond → @st5nd; enmil → Enm2l);
(e) phoneme deletions (e.g., droseful → dr5sEf; bafeness → b{finz);
(f) phoneme additions (e.g., afflave → @fl1vz; sorglom → s9gl@md).

In contrast, the cases in which the RC00 algorithm produced a response that no human reader gave could be more clearly categorized in terms of the application of certain rules. These included:

(a) reducing an initial vowel to schwa and repelling stress as a result of the identification of a prefix (e.g., combal → k@mb{l; forsive → f@s2v);
(b) treating two graphemes as one (e.g., pemble → pem@l; blingle → blIN@l);
(c) reducing the second vowel to a schwa (e.g., adnarb → {dn@b; evact → Ev@kt);
(d) pronouncing silent 'e' (e.g., bafeness → b1fin@s; droseful → dr5sifUl);
(e) producing unusual pronunciations (e.g., byfane → bIf1n; dakey → d1k2);
(f) pronouncing c followed by i as /k/ (e.g., bancing → b{NkIN; ucide → Vk2d).

The large number of model pronunciations given by no human reader would appear to suggest that these models do not provide good descriptions of sublexical processes in disyllabic reading. However, human readers also produce unique pronunciations. If a model is conceptualized as an individual participant, then perhaps these unique pronunciations are to be expected. In order to evaluate this

Table 6
CDP++ and RC00 pronunciation accuracy against human pronunciation data.

Pronunciations	1st	2nd	3rd	4th	5th	6th	7th	Match	Absent	Total
Strict scoring criteria										
CDP++	44%	17%	9%	4%	2%	0%	0%	76%	24%	100%
RC00	55%	22%	7%	3%	1%	0%	0%	88%	12%	100%
Lenient scoring criteria										
CDP++	51%	15%	8%	4%	1%	0%	0%	79%	21%	100%
RC00	69%	16%	5%	2%	0%	0%	0%	92%	8%	100%

possibility, we carried out an analysis that allows us to visualize in a fine-grained manner the extent to which the models lie within the distribution of human responses. We first calculated the similarity of each participant in the study to every other participant; that is, we compared the responses of participant 1 to those of every other participant, the responses of participant 2 to those of every other participant, and so on. This process yielded 40 similarity proportions per participant or a total of 820 unique similarity proportions across all participants. We then compared the pronunciations of CDP++ and RC00 to each individual subject, yielding 41 similarity proportions for each model. Fig. 5 shows the distribution of similarity values for human participants, the CDP++ model and the RC00 algorithm. As is shown in the figure, both models show overall lower similarity to human participants than human participants show to each other. Further, while the RC00 similarity values overlap the lower third of the values from human participants, the similarity values from CDP++ overlap only very few of the plotted human similarity values. We can infer from this analysis that if the models are treated as if they were individual participants, then they are certainly not very typical individuals, particularly in the case of the CDP++ model.

Factors influencing variability in pronunciation

Nonwords attracted an average of 5.9 alternative pronunciations, ranging from 1 to 22 pronunciations. Nonwords that attracted only a single pronunciation and nonwords that attracted 12 or more pronunciations are shown in Appendix D. Pronunciation variability across participants was quantified using the *H statistic*. *H* is calculated using the formula $\Sigma[-\text{pi} \times \log_2(\text{pi})]$, where pi is the proportion of participants giving a certain pronunciation (see also Andrews & Scarratt, 1998; Zevin & Seidenberg, 2006). For the item 'amett', for example, three alternative pronunciations were given by a total of 41 participants. The most frequent pronunciation was given by 30 participants (i.e., a proportion of .732); the second most frequent was given by 8 participants (i.e., a proportion of .195); and the third most frequent was given by 3 participants (i.e., a proportion of .073). Using the above formula, *H* was calculated as:

a. $-.732 \times \log_2(.732) = -.732 \times -0.450 = 0.33$ for the first pronunciation;
b. $-.195 \times \log_2(.195) = -.195 \times -2.358 = 0.46$ for the second pronunciation;
c. $-.073 \times \log_2(.073) = -.073 \times -3.776 = 0.28$ for the third pronunciation.

The three values were then added up (0.33 + 0.46 + 0.28) to give an *H* value of 1.07. An *H* value of 0 denotes that all participants produced a single pronunciation for a given item, whereas higher *H* values indicate pronunciation variability across participants.

In order to understand the factors that influenced pronunciation variability across participants, we conducted item-level regression analyses with *H* as the outcome variable. In line with the stress analyses, we calculated a metric of spelling-to-sound consistency that expressed the consistency with which orthographic units in the first and second syllable map to particular sounds. The same syllabifications as for the analyses on stress were used, which allowed us to derive onset and rime units from our nonwords that matched the onset and rime units of similar words in CELEX, thus increasing the validity of our spelling-to-sound consistency metric. The factors that influence the pronunciation of items containing more than one syllable are not established in the literature. For this reason, we opted to include in our analyses only variables whose properties we could reliably calculate or identify in disyllabic nonwords. Hence, our predictors included the spelling-to-sound consistency of the first and the second syllable, the item's letter length, and the total number of the item's orthographic neighbors. Analyses used the lm function in R and the packages car (Fox & Weisberg, 2011) and QuantPsyc (Fletcher, 2012). All of the variables were entered into the analysis simultaneously. Our predictions about how each of these variables was likely to influence pronunciation variability are outlined below:

i. *Spelling-to-sound consistency*. We calculated two variables, one for the onset and rime units of the first syllable and another for the onset and rime units of the second syllable for each of the 915 nonwords. These variables expressed the consistency with which these units in each syllable map to particular sounds according to the CELEX database (Baayen et al., 1995). The spelling-to-sound consistency measure was expressed using the *H* statistic, so that higher values indicate less consistency in the spelling-to-sound mapping. *H* values for onsets and rimes were then averaged to form a composite measure of unit consistency within each syllable.[5] Our hypothesis was that more consistent mappings between the spellings of onset and rime units and their sounds in a syllable would yield more homogeneous pronunciations, hence less pronunciation variability. Details of the calculation of this metric can be found in Appendix A.
ii. *Neighborhood size*. This variable refers to the number of orthographic neighbors that can be obtained by adding, deleting, or substituting one letter in the nonword. Our hypothesis was that items with more word neighbors would yield less pronunciation variability.

[5] During these calculations, we noted that this composite measure of consistency for items that ended in –ble, –cle, –dle, –fle, etc. did not reflect accurately the spelling-to-sound consistency of these units in the CELEX database. This was because –e as a rime in the second syllable is relatively inconsistent, whereas if such units are considered as a whole, their pronunciations in CELEX are very consistent (except for the unit –tle whose spelling-to-sound mappings are rather inconsistent because of the tendency to drop the /t/ sound, as in 'castle'). These units always corresponded to the second syllable of real words in CELEX. Hence, for all of our items that contained these units (e.g., churble, gibtle, steafle, etc.), we calculated the spelling-to-sound consistency of their second syllable as a whole, and these were the values that we entered in our analyses on pronunciation and naming latencies. It is worth pointing out that it was not possible to calculate whole-syllable spelling-to-sound consistencies for all of our stimuli, because half of the syllables contained in our nonwords did not exist as such in CELEX.

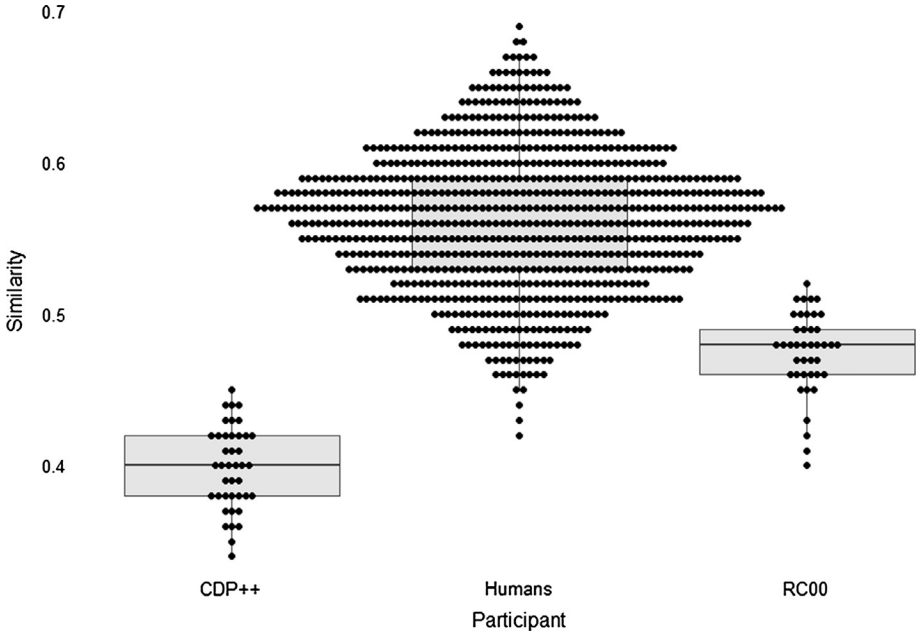

Fig. 5. Pronunciation similarity between participants and between models and every other participant.

Table 7
Regression analyses on pronunciation variability data for humans.

	B(*SE*)	*β*	Relative contributions
Spelling-to-Sound Consistency (1st syllable)	0.45 (0.07)	.22***	40.49
Spelling-to-Sound Consistency (2nd syllable)	0.48 (0.07)	.20***	32.19
Neighborhood size	−0.13 (0.03)	−0.16***	23.41
Item length	−0.04 (0.03)	−0.04	3.91
R^2	.13		

Note.
*** $p < .001$.

iii. *Item length*. This variable refers to the number of letters in each nonword. We hypothesized that longer items would yield more pronunciation variability.

The results of this regression analysis are shown in Table 7.

Results revealed that less consistent spelling-to-sound mappings in the onset and rime units of both the first and the second syllable induced greater pronunciation variability. Further, nonwords with more word neighbors yielded less pronunciation variability. Item length did not influence pronunciation variability. Collinearity was not an issue as variance inflation factors (VIF) for all predictors in the model were less than 1.21 while the lowest observed value of the tolerance statistic (1/VIF) was 0.83. We further sought to quantify the relative contributions of the regressors to the model's total explanatory value. The lmg metric of the relaimpo package in R (Grömping, 2006) estimates the importance of each regressor by decomposing R^2 into non-negative contributions that automatically sum to the total R^2. Relative importance estimates are then adjusted to sum to 100% (see Table 7). The spelling-to-sound consistency of the first syllable contributed the most to the model, followed by the spelling-to-sound consistency of the second syllable and neighborhood size.

Similarly as for stress, considering modal pronunciation belies variability across items in the consistency with which participants assign pronunciation. For this reason, we calculated the percentage of responses that corresponded to each given pronunciation across all nonwords. Nonwords were grouped in terms of the alternative pronunciations they yielded. As is shown in Fig. 6, the vast majority of the responses that participants gave across all items corresponded to the most frequent pronunciation followed by the second most frequent pronunciation. Further, we calculated the percentage of items that yielded the different numbers of alternative pronunciations in order to gauge the overall pronunciation variability. As can be seen in Fig. 7, half of the items (51%) received from 1 to 5 alternative pronunciations, indicating low pronunciation variability; another 40% of the items received 6–10 alternative pronunciations, indicating greater pronunciation variability; 8% of the items received 11–15 pronunciations, and only 1% received over 16 pronunciations.

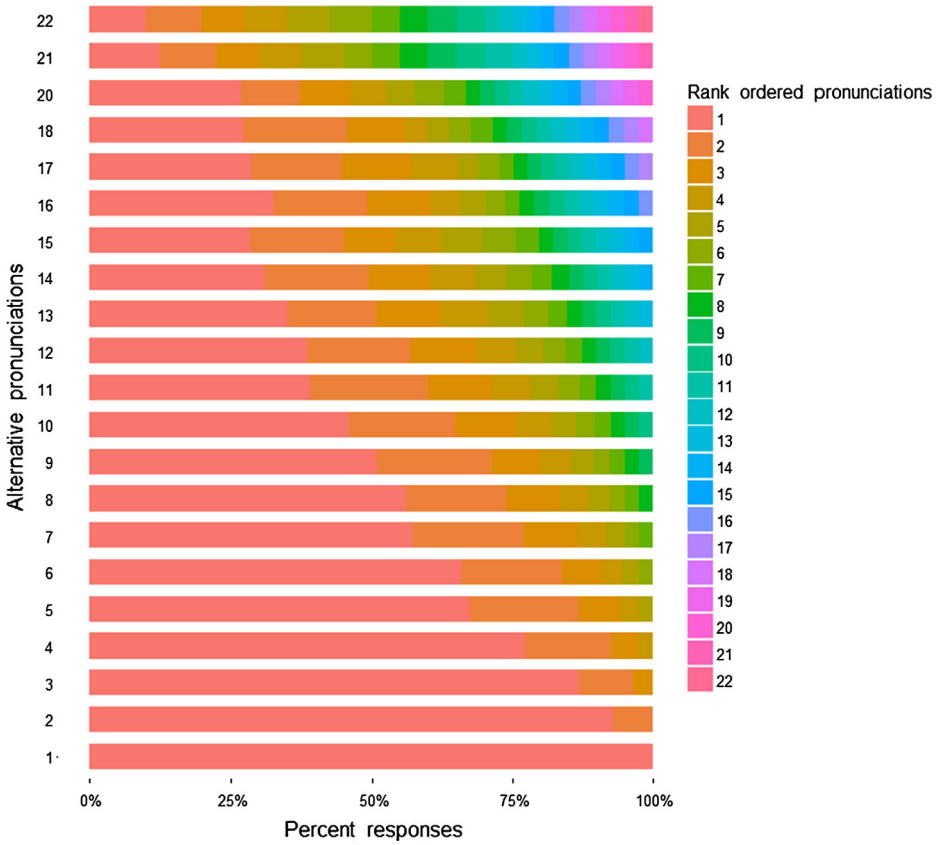

Fig. 6. Percentage of responses given to each alternative pronunciation in items that generated one or more pronunciations.

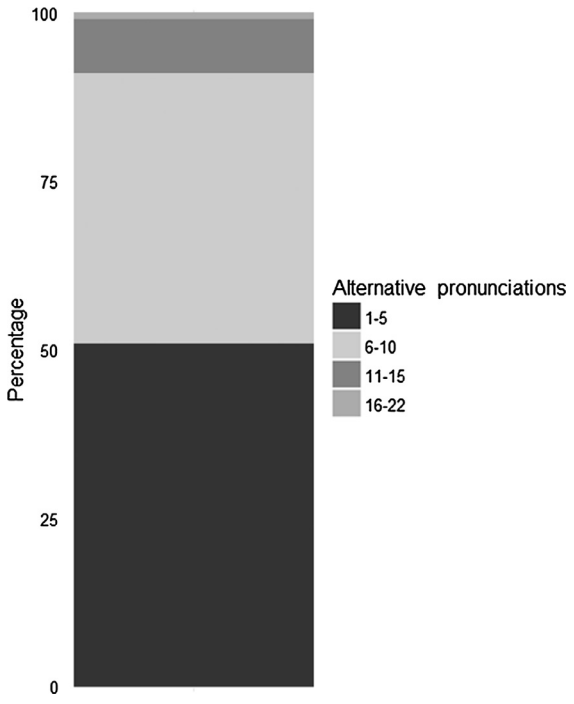

Fig. 7. Percentage of nonwords that yielded 1–5, 6–10, 11–15, and 16–22 alternative pronunciations.

Taken together, the results from these analyses show that even though the nonwords in our study yielded a number of alternative pronunciations ($M = 5.9$, range 1–22), they were overall pronounced in a relatively consistent manner.

RT

Our analysis of RT first considered how well the CDP++ model's reading aloud latencies captured human reading aloud performance. We then conducted regression analyses at the item level to determine the factors that contributed to human and model reading aloud latency.

Latencies in the CDP++ model

Correlational analyses revealed a small but significant relationship between the model's RTs and human RTs, $r = .20$, $p < .001$. RTs from the CDP++ model also correlated significantly with human pronunciation variability (H), $r = .26$, $p < .001$, with longer RTs associated with items that yielded greater pronunciation variability across participants. Finally, the model's RTs correlated significantly with human stress certainty, $r = -.15$, $p < .001$, so that longer RTs were associated with items that yielded lower stress certainty in human readers.

Factors influencing reading aloud latency

In order to determine the factors influencing reading aloud latency, we conducted a regression analysis at the

item level with human RT as the outcome variable. On the basis of previous empirical work (for a review see Perry et al., 2010), a number of factors were included as predictors. Analyses were undertaken using the lm function in R and the packages car (Fox & Weisberg, 2011), MASS (Venables & Ripley, 2002), and QuantPsyc (Fletcher, 2012). Further, the Box-Cox procedure indicated that inverse RT (1/RT) was the optimal transformation to meet the precondition of normality. For the analysis, we multiplied 1/RT by −1000 (−1000/RT) to maintain the direction of effects, so that a larger inverse RT meant a slower response. All of the variables were entered into the analysis simultaneously. The predictor variables included:

i. *Spelling-to-sound consistency*. This is the same variable that we included in the analysis of the pronunciation variability data. Our hypothesis was that more consistent mappings between the spellings of onset and rime units and their sounds in each syllable would yield faster RTs.
ii. *Neighborhood size*. This variable refers to the number of orthographic neighbors that can be obtained by adding, deleting, or substituting one letter in the nonword. Our hypothesis was that items with more word neighbors would yield faster RTs.
iii. *Item length*. This variable refers to the number of letters in each nonword. We hypothesized that longer items may yield slower RTs.
iv. *Onset coding*. This variable refers to potential articulatory parameters that could influence RTs. On the basis of previous findings (Rastle et al., 2005), we considered voicing (with two levels: voiced and unvoiced) and manner of articulation (with five levels: stops/affricates, fricatives, nasals, approximants, and vowels) as potential predictors of naming latencies. For the voicing factor we treated voiced initial consonants as the reference level, and for the manner of articulation factor we treated stops as the reference level. This is because the acoustic onset of items that start with voiced consonants and/or a stop consonant is known to occur later than for other voicing and manner classes of phoneme (see Table 1, Rastle et al., 2005, p. 1089).
v. *Stress certainty*. This is the same variable that we used in the analysis of the stress data and so higher values indicated greater stress certainty. Our hypothesis was that greater stress certainty may be associated with faster RTs.
vi. *Stress*. This is a binary variable that refers to the modal stress assigned to each item across participants (for the analysis of the human RT data) and the stress assigned to each item (for the analysis of the CDP++ RT data). The baseline category was 'first-syllable stress'. First-syllable stress is the most common stress pattern for disyllables in English, hence we hypothesized that first-syllable stressed nonwords may yield faster RTs than second-syllable stressed nonwords.

Results of the regression analysis are shown in Table 8. Standardized coefficients for dummy (factor) variables are meaningless and therefore are not included.

Results showed that nonwords with more word neighbors yielded faster RTs while longer nonwords yielded slower RTs. The place and manner of articulation of the initial consonant also influenced naming latencies, so that items with a voiceless initial consonant yielded faster RTs and items that started with an approximant, fricative, nasal, or a vowel produced faster RTs than items with an initial stop consonant. The stress certainty that an item yielded and its modal stress did not influence naming latencies. Collinearity was not an issue as variance inflation factors (VIF) for all predictors in the model were less than 1.88 while the lowest observed value of the tolerance statistic (1/VIF) was 0.53. We further sought to quantify the relative contributions of the regressors to the model's total explanatory value using the lmg metric of the relaimpo package in R (Grömping, 2006). Relative importance estimates were then adjusted to sum to 100% (see Table 8). Onset coding followed by item length and neigh-

Table 8
Regression analyses on naming latency data for humans and CDP++.

	Humans			CDP++		
	B(*SE*)	*β*	Relative contributions	*B*(*SE*)	*β*	Relative contributions
Spelling-to-Sound Consistency (1st syllable)	0.01 (0.01)	0.04	0.67	8.08** (2.61)	0.11	8.80
Spelling-to-Sound Consistency (2nd syllable)	0.01 (0.01)	0.02	0.41	4.34 (2.73)	0.05	2.57
Neighborhood size	−0.03*** (0.00)	−0.23	26.15	−2.74** (0.95)	−0.10	19.86
Item length	0.04*** (0.00)	0.25	26.97	7.66*** (1.23)	0.22	39.72
Onset coding (voicing)	−0.03** (0.01)		4.10			
Onset coding (manner)			36.66			
Approximant	−0.06*** (0.01)					
Fricative	−0.10*** (0.01)					
Nasal	−0.07*** (0.02)					
Vowel	−0.05*** (0.01)					
Stress Certainty	−0.00 (0.00)	−0.06	4.57	−10.63** (3.39)	−0.11	19.59
Stress	−0.01 (0.01)		0.46	4.49 (2.46)		9.46
R^2	.24			.10		

Note.
** $p < .01$.
*** $p < .001$.

borhood size were the predictors that contributed the most to the model's explanatory value.

In order to determine whether the CDP++ model was sensitive to the same factors, we conducted a similar regression analysis using model RTs as the outcome variable. The same predictors as for the analyses of the human RT data were used except for the onset coding variable that is not relevant for model RTs. Further, stress certainty in the model was expressed as the absolute value of the difference between the activations of the two stress nodes, with higher values indicating greater stress certainty. In agreement with the human data, nonwords with more word neighbors yielded faster RTs and longer nonwords were associated with slower RTs. However, in contrast to the human data, the model yielded a spelling-to-sound consistency effect of the first syllable, so that nonwords with more inconsistent units in the first syllable were associated with slower RTs. Further, greater stress certainty in the model yielded faster RTs. Collinearity was not an issue in the regression model as variance inflation factors (VIF) for all predictors were less than 1.27 while the lowest observed value of the tolerance statistic (1/VIF) was 0.79. As for the human data, we sought to quantify the relative contributions of the regressors to the model's total explanatory value. Item length, followed by neighborhood size and stress certainty were the predictors that contributed the most to the model's explanatory value. The results from these analyses are shown in Table 8.

General discussion

Theoretical and empirical work on the cognitive processes that underpin reading aloud has flourished over the past 30 years. Despite this, our understanding of these processes is largely confined to monosyllables, which form only a very small part of the English vocabulary and are virtually absent from many of the world's languages. Recent developments in computational models of reading aloud (Perry et al., 2010) have yielded some theoretical progress in this domain of research, yet the empirical work is very limited. The present study provides the first database of stress assignments, pronunciations, and reading aloud latencies for a variety of monomorphemic and polymorphemic disyllabic English nonwords.

Our analyses of over 37,000 reading aloud responses to 915 disyllabic nonwords sought to identify cues to stress assignment in English, and to uncover the factors that influence pronunciation variability and reading latency across skilled adult readers. We also used these data to evaluate the adequacy of three approaches to disyllabic reading aloud that have been computationally instantiated: the CDP++ model (Perry et al., 2010), the rule-based disyllabic algorithm (Rastle & Coltheart, 2000), and the connectionist print-to-stress network (Ševa et al., 2009). In addition to revealing new insights about how skilled English readers process disyllabic letter strings, and evaluating the adequacy of extant computational approaches to reading aloud, the present database is by far the largest of its kind for assessing generalization in computational models of reading. By making the full dataset available with this article, we hope to foster rapid development of the next generation of these models, particularly as they move beyond the monosyllabic domain.

Stress assignment

Our analyses of the most frequent (modal) stress indicated that participants assigned first-syllable stress to 77% of the items, and second-syllable stress to 23% of the items, thus mirroring the distribution of stress in the language (i.e., approximately 75% of English disyllabic words are stressed on the first syllable). Finer-grained analyses showed that nearly 40% of items yielded very high certainty in stress assignment across participants, and that almost all of these items had modal first syllable stress. Just over 20% of items had very low certainty in stress assignment across participants, and a much higher percentage of these had modal second syllable stress. Regression analyses revealed several item-level factors that influenced participants' assignment of stress: these included the spelling-to-stress consistency of the onset and rime units in the first and the second syllables, the relative orthographic weight of the two syllables, the second syllable's vowel length, the stress pattern of the item's orthographic neighbors, as well as the certainty with which stress was assigned across participants. Several of these variables are strongly related to the morphological structure of the English writing system. Because of these strong intercorrelations (in our stimuli and in the language), we cannot determine whether the simple presence of an affix within a letter string has a direct influence on stress assignment as postulated by Rastle and Coltheart (2000). However, we can conclude that morphology likely has a powerful indirect influence on stress assignment by virtue of its significance in the nature of the mapping between sublexical orthographic units and stress, and its significance in the relative salience of syllables within words across the English lexicon.

Recently, Ktori et al. (2016) reported that people with acquired surface dyslexia frequently assign incorrect stress to prefixed words with a strong-weak stress pattern (e.g., reflex). Ktori et al. (2016) argued that these errors provide strong evidence for the functional role of prefixes in stress assignment during reading. However, in light of the findings reported in this paper, it is uncertain whether the patients in this study were indeed sensitive to prefixation when assigning second-syllable stress to most prefixed disyllabic words, or whether it was some related factor that led them to show this stress pattern (e.g., orthographic weight, spelling-to-stress consistency). Because it is impossible to find sets of English monomorphemic and polymorphemic words matched on these factors, it would be difficult to disentangle experimentally these two possibilities. One possibility may be to conduct factorial experiments with nonwords purposely designed to pull these interrelated cues apart; this is an approach that we are currently investigating (Ktori, Mousikou, & Rastle, 2016).

Our analysis of the model data revealed that all models performed well in producing the modal stress given by participants, predicting the human response in 81% (CDP ++), 79% (SMA09), and 73% (RC00) of cases. Further analysis

revealed that nonwords that yielded high stress certainty across participants also yielded high stress certainty in the CDP++ model and the SMA09 network, with the former model performing slightly better than the latter. However, when we conducted the same item-level regression analyses as for the human participants, we observed that not all of the models performed well for the right reasons. While the CDP++ model and the SMA09 network were successful in capturing human modal stress, with the SMA09 network approaching more the human data according to our analyses of the relative importance of the individual predictors in the regression models, the RC00 algorithm failed to capture the effects of orthographic weight and vowel length that human readers showed. However, it is worth noting that the underlying reason why both the SMA09 network and the CDP++ model showed a vowel length effect on stress assignment is likely because vowel length was confounded with number of vowel letters in our nonwords, and so both models may capture this effect on the basis of orthographic information (as may also be the case for human readers).

Pronunciation

We quantified variability in pronunciation using the *H statistic*, a measure of entropy that takes into account the proportion of participants producing each alternative pronunciation. Lower values of *H* denoted fewer pronunciations for a given item, whereas higher values of *H* denoted more pronunciation variability across participants. Results showed that the factors that determine variability in pronunciation are the spelling-to-sound consistency of sublexical orthographic units in the first and the second syllable and the orthographic neighborhood in which the items reside in the lexicon. Additional analyses that we carried out to gauge the overall pronunciation variability across participants showed that even though the nonwords in our study yielded a number of alternative pronunciations (ranging from 1 to 22), they were overall pronounced in a relatively consistent manner.

The analyses of the model pronunciation data revealed that the CDP++ model produced the modal human pronunciation in 44% of cases while it produced a response given by no human participant for 24% of the items. The model responses in these cases reflected a combination of different types of errors, which would require intensive study of the model to understand fully.[6] Conversely, the RC00 algorithm captured the human modal pronunciation in 55% of cases, and gave a response produced by no human participant in just 12% of cases. These were all cases that arose as a result of individual hard-wired rules in the algorithm; hence, it is not difficult to envisage how these could be altered to improve the fit to the human data. Human readers sometimes produce unique responses, and so we went even further to assess whether either of the models could be considered as a typical individual participant. These analyses revealed that the models showed much lower similarity to human participants than human participants showed to each other. In fact, while the RC00 algorithm overlapped the lower third of the similarity distribution for human readers, the CDP++ model fell almost totally outside of the distribution for human readers. These data allow us to infer that if we treat models as if they were individual participants, they do not behave like very typical participants.

RT

Though we did not emphasize speed in our instructions to participants, the average speed with which they responded was relatively fast (818 ms). In the analyses of the reaction time data, we observed that onset coding, item length, and the items' orthographic neighborhood were all significant predictors of human naming latencies. However, the CDP++ model only approached the human data in terms of the factors that gave rise to its latencies. In agreement with the human data, item length and orthographic neighborhood were significant predictors of naming latencies in the model, yet the model also showed sensitivity to stress certainty and the spelling-to-sound consistency of the first syllable, neither of which influenced human RTs.

Model limitations and future directions

One of the aims of the present study was to assess the reading performance of statistical (CDP++; SMA09) and rule-based (RC00) computational models of reading against a large dataset of human reading aloud responses to disyllabic nonwords. The SMA09 network was the most successful model in capturing human performance on stress assignment, followed by the CDP++ model, thus providing support for a statistical-learning approach to relating spelling to stress in disyllabic reading aloud. However, in terms of pronunciation, the CDP++ model performed less well. A close inspection of the different types of pronunciation errors that the CDP++ model made (see Appendix B) revealed that these were primarily due to (a) the way the model assigns graphemes to the different slots in the graphemic buffer, and (b) the model overgeneralizing the spelling-to-sound relationships that it learns during training.

In relation to (a), the CDP++ model uses the Maximum Onset Principle, according to which consonant graphemes after the first vowel are placed in all available onset positions of the second syllable. However, most phoneme omissions that occurred in the model (e.g., grametul → gr {mIt; lafeless → l{filz) seemed to be due to grapheme assignment problems that were caused by this principle. For example, in the case of 'grametul', the consonant that follows the first vowel (i.e., /m/) is activated in the onset of the second syllable. The vowel that follows this consonant (i.e., /e/) is activated in the second syllable, and so the following consonant (i.e., /t/) is activated in the coda of the second syllable. No English word finishes in /tl/, so /l/ never gets activated in the coda of the second syllable. It is worth noting here that the CDP++ parser was designed

[6] We also ran these simulations with the CDP++ parser (Perry, Ziegler, & Zorzi, 2013) and the results were very similar to those obtained with the CDP++ model. We chose to report the simulations with the CDP++ model because it has been tested more extensively than the parser implementation.

to get around the pronunciation problems that were due to the onset maximization principle implemented in the CDP++ model. In relation to (b), statistical information about spelling-to-sound relationships is captured by the network during training. However, phoneme additions and unlikely print-to-sound correspondences were often observed in the model's pronunciations (e.g., astond → @st5nd; bethove → bITVv; strastle → str#s@l; arreme → @r1m), because the model does not have context-independent knowledge of spelling-to-sound relationships that seems to be used by skilled adult readers when they are reading aloud nonwords. One could argue that some of these errors may be due to idiosyncracies in the training set (e.g., CELEX often produces the vowel in 'golf' as /5/, hence the model's pronunciation of astond as /@st5nd/). However, other similar words in CELEX (e.g., pond) are not pronounced in the same unusual way, and so the few idiosyncracies in CELEX cannot account for the full range of errors reported in Appendix B. Therefore, future developments of the model would need to consider these two points.

Rule-based vs. statistical-learning approaches to reading aloud

Our data provide important insights into the question of whether sublexical knowledge is best characterized in terms of rules or in terms of learned statistical mappings. In order to address this question, we focused on how participants stressed and pronounced prefixed nonwords. The choice of prefixed items was deliberate since a rule-based approach like that adopted by the RC00 algorithm posits that all prefixes repel stress and also, that each prefix is likely to be pronounced in a single manner (e.g., 'be' is always pronounced as /bI/ irrespective of its context). In contrast, a statistical learning approach like that adopted by the CDP++ model predicts that the stress pattern and pronunciation of prefixes depends on the context in which they occur and the statistical properties of the English lexicon.

Our analyses of the stress data showed that the RC00 algorithm showed sensitivity to the spelling-to-stress consistency of the first and the second syllable in assigning stress. Yet we know that this sensitivity must be due to the model's implemented rule that prefixes and suffixes repel stress. In other words, prefixes are associated with spelling-to-stress consistencies that point toward 2nd-syllable stress while suffixes are associated with spelling-to-stress consistencies that point toward 1st-syllable stress. Could it be then that once people learn the associations between certain spellings that correspond to affixes and their stress patterns, they form the general rule that affixes repel stress? If that is the case we would expect that participants assign second-syllable stress to prefixed items in a very consistent manner. A rule of this kind may even be formed only in those cases where prefixes are strongly and almost always associated with second-syllable stress in the lexicon.

We examined this issue by calculating first the spelling-to-stress consistency of all of the prefix units used in our study. Similarly to the calculation of the spelling-to-stress consistency of the onset and rime units, prefix units with a spelling-to-stress consistency close to 0 pointed toward 2nd-syllable stress in the CELEX database, whereas prefixes with a spelling-to-stress consistency close to 1 pointed toward 1st-syllable stress. We identified 5 prefixes that very strongly pointed toward 2nd-syllable stress (Mean spelling-to-stress consistency = 0.04). These were the prefixes 'be', 'un', 'ap', 'em', and 're'. We then calculated the percentage of times that each participant assigned 1st- and 2nd-syllable stress to all of the nonwords in our study that contained these prefixes and we averaged these percentages for each prefix and type of stress (1st vs. 2nd). Across all five prefixes, 1st-syllable stress was assigned 47% of the time while 2nd-syllable was assigned 53% of the time. This result shows great inconsistency across participants in assigning stress to prefixed items. However, might it be that some readers use rules in assigning stress to prefixed items and others do not, depending on the type of reading instruction they received at school and their reading skills? If that were the case we should observe within-participant consistency in assigning second-syllable stress to prefixed items, at least on some occasions. This was not the case in our data; we could not find any participants who reliably assigned second-syllable stress to nonwords with particular prefixes across all items that contained those prefixes. Hence the results from this analysis provide strong evidence against a rule-based approach to stress assignment, at least the one adopted by the RC00 algorithm.

In terms of pronunciation, we took a quite similar approach to investigate whether sublexical spelling-to-sound knowledge is best characterized in terms of rules or statistics. In particular, we first calculated from CELEX the spelling-to-sound consistency of each of the prefix units used in our study and expressed this as an *H* value. Hence, a value of 0 corresponding to a certain prefix unit indicated that this prefix is always pronounced in the same way in the lexicon, whereas values greater than 0 indicated pronunciation variability of this prefix across the lexicon. We then identified seven prefixes with the lowest *H* values (mid, mis, out, im, in, fore, dis), indicating high pronunciation consistency (Mean $H = 0.06$).[7] We hypothesized that if people use rules to translate print to sound, and on the assumption that these rules might differ across readers depending on the type of reading instruction they received at school and/or their reading skills (Mousikou, Coltheart, Finkbeiner, & Saunders, 2010; Thompson, Connelly, Fletcher-Flinn, & Hodson, 2009), we may observe within-participant consistency in the pronunciation of prefixes that have very consistent spelling-to-sound mappings in the lexicon. This is because readers may be more likely to form a rule when there are strong associations between certain spellings and their corresponding sounds in the lexicon. For this reason, we further calculated the consistency with which participants pronounced the prefixes with the lowest identified *H* values. We expressed this consistency as an *H* value too. Even though participants were overall consistent

[7] The prefix 'by' had an *H* value of 0; hence it was one of the prefixes with the lowest *H* values. However, this prefix occurred only in a single item in our set, hence the corresponding analyses would not be informative and for this reason, we did not consider it.

in the way they pronounced these seven prefixes (Mean $H = 0.21$), only two of them pronounced these prefixes always in the same manner. Hence, even if we assume that different people may apply different rules to translate print to sound depending on their reading instruction and skills, the results from our analysis show that a rule-based approach to reading aloud is rather unlikely.

Our data further allowed us to investigate whether readers are sensitive to the statistics of the lexicon when they are translating printed letter strings into their corresponding sounds. In particular, in addition to identifying the prefixes with the lowest H values, we identified seven prefixes with the highest H values (pre, sur, com, ex, for, ar, ad), indicating low pronunciation consistency (Mean $H = 1.35$).[8] We then extracted all of the nonwords in our dataset that contained the identified prefixes with the lowest and highest H values ($N = 144$). A statistical learning approach would predict that items that contain prefixes with consistent spelling-to-sound mappings in the lexicon (prefixes with very low H values) may yield less pronunciation variability across participants than prefixes with inconsistent spelling-to-sound mappings in the lexicon (prefixes with high H values). Thus, we compared pronunciation variability across participants (H) in the items containing consistent prefixes and the items containing inconsistent prefixes. We observed that the difference between the two means (0.80 vs. 1.46) was highly significant ($t(65) = -4.91$, $p < .001$). Thus, this analysis provides support for a statistical learning approach to reading aloud.

Applied implications

There is now a broad consensus that the acquisition of sublexical knowledge is critical for the development of successful reading (Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001), and that most children with developmental reading disabilities are characterized by a phonologically-based deficit (Rack, Snowling, & Olson, 1992) associated with difficulty acquiring sublexical knowledge (Jackson & Coltheart, 2001). Thus, the assessment of this knowledge is central to diagnosing reading difficulties or disorders. The present study is the first to investigate extensively the nature of this sublexical knowledge in disyllabic reading, thus motivating the development of new nonword reading tests to assess reading deficits, and the implementation of a sound and effective literacy strategy.

Further, our work has important downstream clinical implications. For example, speech pathologists, clinical neuropsychologists, and professionals involved in special needs education all rely significantly on measures of nonword reading performance to conduct assessments of people with acquired and developmental reading disorders. However, a recognized problem in this area is that there is a lack of normative data on skilled nonword reading, thus making it very difficult to determine what constitutes 'impaired' or 'atypical' (Tree, 2008) and to design suitably targeted interventions (Colenbrander, Nickels, & Kohnen, 2011). Individuals with acquired phonological dyslexia, for example, have severe impairments at nonword reading (Coltheart, 1996), often producing real words in response to nonwords (i.e., 'zoo' or 'pool' in response to 'zool'). Such errors are typically judged against the 'modal' response that a small sample of typical skilled readers generates; for example, if the majority of a small group of skilled readers pronounce 'zool' to rhyme with 'pool' and none of the individuals with acquired dyslexia pronounces this item in the same way, it is considered an error. However, nonwords may elicit different responses across normal skilled readers, as our study demonstrates, and so having a distribution of nonword responses available at the item level will improve and facilitate the assessment of these individuals. Our dataset provides the first normative nonword corpus for British English and is the largest database of its kind, so we believe that it will be particularly useful to this group of users.

[8] The prefixes 'al' and 'cor' had among the highest H values, however, each of these prefixes occurred in a single item in our set, hence the corresponding analyses would not be informative and for this reason, we did not consider them.

Conclusion

The present paper reports a large-scale study in which 41 participants read aloud 915 disyllabic nonwords, yielding a total of around 37,000 responses. We investigated the cues to stress assignment and the factors that influence pronunciation variability and reading latencies in the English language. We also compared human reading performance to the reading performance of computational models of reading that adopt rule-based and statistical-learning approaches to explaining disyllabic reading. Our findings provided support for the latter approach, although we identified important deficiencies with the most fully developed model of this type. The findings from the present study make a critical theoretical contribution to our understanding of skilled adult reading. Further, this dataset provides the first normative nonword corpus for British English and is the largest database of its kind for any language, thus being critical for evaluating generalization in models of reading as they advance into the disyllabic domain. Finally, our findings have significant applied implications for the development of evidence-based strategies for literacy education and the clinical diagnosis and treatment of reading impairments.

Acknowledgments

This research was supported by a research grant from the Leverhulme Trust (RPG 2013-024) awarded to Kathleen Rastle, Max Coltheart, Jeremy Tree, and Petroula Mousikou. The authors thank Padraic Monaghan for running the 915 nonwords through the Ševa et al. (2009) network, Svetlana Rodinskaya for her invaluable help with the calculation of the consistency metrics, Amy Gatton for hand-marking participants' response latencies, and Eva Liu for labelling the acoustic boundaries of the vowels in a subset of nonword responses to extract vowel intensity. The study was conceived by P.M. and K.R.; stimuli were selected by P. M., J.S., and K.R.; the study was run by J.S. and R.L.; tran-

scriptions and stress judgments were conducted by R.L. and K.R.; analyses were conducted by P.M. and K.R.; and the manuscript was written by P.M. and K.R.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2016.09.003.

References

Andrews, S., & Scarratt, D. R. (1998). Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 1052–1088.

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

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

Baayen, H. R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database [CD-ROM]*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Baker, R. G., & Smith, P. T. (1976). A psycholinguistic study of English stress assignment rules. *Language and Speech, 19*, 9–27.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445–459.

Bates, T. C., Lind, P. A., Luciano, M., Montgomery, G. W., Martin, N. G., & Wright, M. J. (2010). Dyslexia and DYX1C1: Deficits in reading and spelling associated with a missense mutation. *Molecular Psychiatry, 15*, 1190–1196.

Besner, D., Twilley, R. S., McCann, R. S., & Seergobin, K. (1990). On the association between connectionism and data: Are a few words necessary? *Psychological Review, 97*, 432–446.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Burani, C., & Arduino, L. S. (2004). Stress regularity or consistency? Reading aloud Italian polysyllables with different stress patterns. *Brain and Language, 90*, 318–325.

Chateau, D., & Jared, D. (2003). Spelling-sound consistency effects in disyllabic word naming. *Journal of Memory & Language, 48*, 255–280.

Colenbrander, D., Nickels, L., & Kohnen, S. (2011). Nonword reading tests: A review of the available resources. *Australasian Journal of Special Education, 35*, 137–172.

Colombo, L. (1992). Lexical stress effect and its interaction with frequency in word pronunciation. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 987–1003.

Colombo, L., & Tabossi, P. (1992). Strategies and stress assignment: Evidence from a shallow orthography. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 319–340). Amsterdam, The Netherlands: Elsevier.

Colombo, L., & Zevin, J. (2009). Stress priming in reading and the selective modulation of lexical and sub-lexical pathways. *PLoS ONE, 4*, e7219.

Coltheart, M. (1996). Phonological dyslexia: Past and future issues. *Cognitive Neuropsychology, 13*, 749–762.

Coltheart, M. (2006). Acquired dyslexias and the computational modelling of reading. *Cognitive Neuropsychology, 23*, 96–109.

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.

Croissant, Y. (2013). mlogit: Multinomial logit model R package version 0.2-4 <http://CRAN.R-project.org/package=mlogit>.

Croot, K., Fletcher, J., & Harrington, J. (1992). Phonetic segmentation of the Australian National Database of Spoken Language. In *Proceedings of the 4th International Conference on Speech Science and Technology, Brisbane* (pp. 86–90).

Fletcher, T. D. (2012). QuantPsyc: Quantitative Psychology Tools R package version 1.5 <http://CRAN.R-project.org/package=QuantPsyc>.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers, 35*, 116–124.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>.

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

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.

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software, 17*, 1–27.

Guion, S. G., Clark, J. J., Harada, T., & Wayland, R. P. (2003). Factors affecting stress placement for English non-words include syllabic structure, lexical class, and stress patterns of phonologically similar words. *Language and Speech, 46*, 403–427.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*, 662–720.

Jackson, N., & Coltheart, M. (2001). *Routes to reading success and failure*. Hove: Psychology Press.

Jouravlev, O., & Lupker, S. J. (2015). Lexical stress assignment as a problem of probabilistic inference. *Psychonomic Bulletin & Review, 22*, 1174–1192.

Kahn, D. (1976). *Syllable-based generalizations in English phonology*. Bloomington, IN: Indiana University Linguistics Club.

Kello, C. T. (2006). Considering the junction model of lexical processing. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in lexical processing* (pp. 50–75). New York: Psychology Press.

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

Kelly, M. H., Morris, J., & Verrekia, L. (1998). Orthographic cues to lexical stress: Effects on naming and lexical decision. *Memory & Cognition, 26*, 822–832.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*, 627–633.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America, 118*, 1038–1054.

Ktori, M., Mousikou, P., & Rastle, K. (2016). Cues to stress assignment in reading aloud. In *UK orthography group annual meeting, Oxford, UK*. .

Ktori, M., Tree, J. T., Mousikou, P., Coltheart, M., & Rastle, K. (2016). Prefixes repel stress in reading aloud: Evidence from surface dyslexia. *Cortex, 74*, 191–205.

Kuhn, M. (2015). A short introduction to the caret package. http://cran.r-project.org/web/packages/caret/vignettes/caret.pdf.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy- Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE, 7*, e43230.

Masterson, J. (1985). On how we read nonwords: Data from different populations. In K. E. Patterson, J. C. Marshall, & M. Coltheart (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading* (pp. 289–299). London, UK: Lawrence Erlbaum Associates.

Mousikou, P., Coltheart, M., Finkbeiner, M., & Saunders, S. (2010). Can the dual-route cascaded computational model of reading offer a valid account of the masked onset priming effect? *The Quarterly Journal of Experimental Psychology, 63*, 984–1003.

Paap, K. R., & Noel, R. W. (1991). Dual route models of print to sound: Still a good horse race. *Psychological Research Psychologische Forschung, 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.

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology, 61*, 106–151.

Perry, C., Ziegler, J. C., & Zorzi, M. (2013). A computational and empirical investigation of graphemes in reading. *Cognitive Science, 37*, 800–828.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56–115.

Pritchard, S. C., Coltheart, M., Palethorpe, S., & Castles, A. (2012). Nonword reading: Comparing dual-route cascaded and connectionist dual-process models with human data. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 1268–1288.

Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods, 39*, 859–862.

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing <http://www.R-project.org/>.

Rack, J. P., Snowling, M. J., & Olson, R. K. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly, 27*, 29–53.

Rastle, K., & Coltheart, M. (1999). Serial and strategic effects in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 482–503.

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.

Rastle, K., Croot, K. P., Harrington, J. M., & Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 1083–1095.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*, 31–74.

Robidoux, S., & Pritchard, S. (2014). Hierarchical clustering analysis of reading aloud data: A new technique for evaluating the performance of computational models. *Frontiers in Psychology, 5*, 267.

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., Petersen, A. S., McClelland, J., & McRae, K. (1994). Nonword pronunciation and models of word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1177–1196.

Selkirk, E. O. (1982). The syllable. In H. Van der Hulst & N. Smith (Eds.), *The structure of phonological representations (Part II)*. Dordrecht: Foris.

Ševa, N., Monaghan, P., & Arciuli, J. (2009). Stressing what is important: Orthographic cues and lexical stress assignment. *Journal of Neurolinguistics, 22*, 237–249.

Shannon, C. E. (1949). The mathematical theory of communication. In C. E. Shannon & W. Weaver (Eds.), *The mathematical theory of communication* (pp. 29–125). Urbana: University of Illinois Press.

Sibley, D. E., Kello, C. T., Plaut, D. C., & Elman, J. L. (2008). Large-scale modeling of wordform learning and representations. *Cognitive Science, 32*, 741–754.

Sibley, D. E., Kello, C. T., & Seidenberg, M. S. (2010). Learning orthographic and phonological representations in models of monosyllabic and bisyllabic naming. *European Journal of Cognitive Psychology, 22*, 650–668.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8*, 411–416.

Sulpizio, S., Burani, C., & Colombo, L. (2015). The process of stress assignment in reading aloud: Critical issues from studies on Italian. *Scientific Studies of Reading, 19*, 5–20.

Sulpizio, S., & Colombo, L. (2013). Lexical stress, frequency and stress neighborhood effects in the early stages of Italian reading development. *Quarterly Journal of Experimental Psychology, 66*, 2073–2084.

Sulpizio, S., Job, R., & Burani, C. (2012). Priming lexical stress in reading Italian aloud. *Language and Cognitive Processes, 27*, 808–820.

Taylor, J. S. H., Rastle, K., & Davis, M. H. (2013). Can cognitive models explain brain activation during word and pseudoword reading? A meta-analysis of 36 neuroimaging studies. *Psychological Bulletin, 139*, 766–779.

Thompson, G. B., Connelly, V., Fletcher-Flinn, C. M., & Hodson, S. J. (2009). The nature of skilled adult reading varies with type of instruction in childhood. *Memory & Cognition, 37*, 223–234.

Tree, J. J. (2008). Two types of phonological dyslexia – A contemporary review. *Cortex, 44*, 698–706.

Treiman, R., & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language, 27*, 87–104.

Treiman, R., Mullennix, J., Bijelac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General, 124*, 107–136.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). 0-387-95457-0. New York: Springer.

Woollams, A. M., Lambon Ralph, M. A., Plaut, D. C., & Patterson, K. (2007). SD-squared: On the association between semantic dementia and surface dyslexia. *Psychological Review, 114*, 316–339.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60*, 502–529.

Zevin, J. D., & Seidenberg, M. S. (2006). Consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language, 54*, 145–160.