Cross-linguistic evidence for probabilistic orthographic cues to lexical stress

Padraic Monaghan

Lancaster University, UK

Joanne Arciuli

University of Sydney, Australia

Nada Seva

Lancaster University, UK

January 2013

Running head: Cross-linguistic analyses of stress assignment

Please address correspondence to:

Padraic Monaghan, Centre for Research in Human Development and Learning,

Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK.

Tel: +44 1524 593813; Fax: +44 1524 593314; Email: p.monaghan@lancaster.ac.uk

Cross-linguistic evidence for probabilistic orthographic cues to lexical stress

Most of what we know about the process of converting orthography to phonology during reading aloud is based on data from monosyllabic words. Moreover, much of the research in this area is based on reading aloud in English. There has been increasing interest in the mechanisms underpinning the reading aloud of polysyllables. In many languages monosyllabic words represent a small proportion of the whole vocabulary, and so restriction to these items may mean that the model is not representative of the reading system – monosyllabic words may be a special case. For instance, in English though single syllable words account for 70.9% of tokens in the CELEX corpus (Baayen, Pipenbrock, & Gulikers, 1993) they only account for 15.5% of the word types. In other languages, the imbalance is even greater, for the Dutch CELEX database, monosyllables account for 63.3% and 7.9% of tokens and types, respectively, but for the German CELEX database, monosyllables account for 50.8% of tokens but only 3.8% of types. However, if models of reading can apply to bisyllabic and trisyllabic words as well, then this increases the coverage of the whole language in English up to 96.8% of tokens and 82.4% of types. For German and Dutch, the coverage is 91.9% and 94.4% for tokens, and 55.8% and 61.1% for types, respectively.

A comprehensive understanding of word naming must thus include knowledge of how both monosyllabic and polysyllabic words are read aloud – and knowledge of how this process operates in distinct languages, otherwise, as in the case of German, models of reading based on monosyllables apply effectively to only 1 in 26 words. There are substantial challenges that are introduced when one considers naming of polysyllables. One of the biggest challenges has been to determine how lexical stress is assigned when reading aloud polysyllabic words (Arciuli & Cupples, 2006; 2007; Arciuli, Monaghan & Seva, 2010; Burani & Arduino, 2004; Colombo, 1992; Colombo & Zevin, 2009; Levelt, Roelofs, & Meyer, 1999). In this chapter we report the results of a cross-linguistic examination of probabilistic orthographic cues to lexical stress in 6 languages: English, Dutch, German, Italian, Spanish, and Greek.

For Italian, Colombo (1992) and Burani and Arduino (2004) demonstrated that there were regularities between certain patterns of word endings and stress position in Italian trisyllabic words. Naming times for high-frequency words were not affected by whether the stress assignment associated with the ending was consistent or inconsistent with the word's stress position. However, for low-frequency words, words with inconsistent endings were named more slowly than words with consistent endings. Hence, the critically important frequency x regularity effect, observed for languages with deep orthographies for segmental phonology (Paap & Noel, 1991; Taraban & McClelland, 1987) which provides important constraints for the role of sublexical information in models of reading (Seidenberg & McClelland, 1989), could be demonstrated for languages with shallow orthographies, but which contained inconsistencies with regard to orthography-stress correspondences. For recent work on stress assignment in Italian see Sulpizio, Arduino, Paizi, and Burani, (2012) and Sulpizio, Job and Burani (2012).

Arciuli and colleagues have shown that the orthography contained at word beginnings and endings provides substantial information regarding stress position in English words. In their corpus analyses and behavioural studies, they examined the letters corresponding to the body of the first syllable, and the letters corresponding to the rime of the last syllable, as cues to stress position in bisyllabic (Arciuli & Cupples, 2006; 2007). They found a high degree of accuracy in determining stress position from just these beginning/ending cues, and also that ending cues appeared to be more helpful at determining stress position. Extending this work, Arciuli, Monaghan, and Seva (2010) measured the reliability of these beginning and ending cues in vocabularies that are appropriate for children of different ages learning to read. They found the same pattern of beginning/ending cues: both were reliable indicators of stress position, but endings were more accurate. Children tested on nonwords that varied the presence of beginning/ending cues to stress position demonstrated sensitivity to these cues, and increasing sensitivity to endings as reading development progressed.

Thus, for at least two languages, there are orthographic indicators of stress position that have profound influences on stress assignment in reading. In this study we examined whether orthographic cues at the very beginning and ending of words serve as reliable indicators of lexical stress across a broader range of languages, and if so, what level of granularity the orthographic cue is expressed – is it single letters, bigrams, or larger constructions such as syllables? There is reason to suspect that the granularity may vary across languages. For instance, English has complex syllables – there are consonant clusters in both onset and offset of syllables, so the beginnings and endings may be particularly useful due to their variability. In languages such as Italian or Spanish, with simpler syllable structures, the same patterns may not be observed because the variation in syllabic structure and the possibilities for indicating distinctions in the syllable are reduced. Of note, the Italian research mentioned above suggests that endings might be important. Additionally, in other orthographies, stress may be indicated by particular cues such as diacritics in Greek, or accents in Spanish (Gutiérrez-Palma, 2003; Gutiérrez-Palma & Palma-Reyes, 2008); therefore, stress position reflected in patterns of letters would be redundant. In Greek, stress position is indicated by a mark over the letter similar to an acute accent but only for writing in the lower-case alphabet. When capitalised, diacritics are generally not used, and the reader must rely on other information to determine stress position (Protopapas, 2006).

We investigated this issue by undertaking corpus analyses of beginning and ending cues to stress position in 6 languages that differ in terms of the regularity of orthography-phonology mappings, syllabic complexity, and the extent to which stress is marked or not marked explicitly in the orthography.

Method

Corpus Preparation

We selected 6 languages to analyse: English, Dutch, German, Italian, Spanish, and Greek. These languages were selected in order to represent variation in language family, variation in the transparency of orthography for segmental phonology, as well as variation in the way in which the languages marked stress (frequently, as for Greek, rarely, as for Italian and Spanish, and seldom, as for Dutch, English, and German).

The English, Dutch, and German corpora were taken from the CELEX database (Baayen et al., 1993), with frequency information for English from 18.6 million words, frequency information for Dutch from 40.2 million words, and frequency information for German from 5.0 million words. The Italian corpus was taken from the CoLFIS database (Bertinetto, Burani, Laudanna, Marconi, Ratti, Rolando, & Thornton, 2005), with frequency information derived from a corpus of 3 million words. Stress position information was taken from the De Mauro Italian Dictionary (De Mauro, 2000). The Spanish corpus was taken from LEXESP (Sebastian, Marti, Carreiras, & Cuetos, 2000), with frequency counts derived from a total count of 4.8 million words. The Greek corpus was taken from the Hellenic National Corpus (HNC; Hatzigeorgiu et al., 2000; http://hnc.ilsp.gr), and frequency counts were generated from 18.3 million words, as available in 2009, and described further in Protopapas and Vlahou (2009).

From each corpus, we selected all the bisyllabic and trisyllabic words with frequency greater than 0. We focused on predicting the primary stress position within words, as not all corpora indicated secondary stress. Characteristics of each corpus are shown in Table 1.

Insert Table 1

Cue preparation

We assessed each language corpus for whether orthographic beginnings and endings of words could predict the stress position of the word. We first replicated the analyses by Arciuli and colleagues by deriving the orthography corresponding to the body of the first syllable and the rime of the last syllable. We also generated cues representing the first or last consonant in the word, the first or last vowel in the word, the first or last one, two, three, four or five letters to determine whether different types of cue carried more or less information about stress position in different languages. Each word beginning or ending was taken as a cue in the analyses, and bisyllabic and trisyllabic words were assessed separately.

The effectiveness of each cue was determined using a variant of discriminant analysis, appropriate for binary cues that are discrete in their coverage of the corpus. The first stage of the method was to determine for each cue which stress position it was most closely associated with. So, for each cue individually, we determined the proportion of words containing that cue with stress on each syllable. To determine whether these proportions were significantly different from chance, we computed estimated proportions of classification using a random baseline. The baseline involved randomly reassigning the stress position and frequency for each word to another word in the corpus, and then determining the proportion of words containing the cue in question with stress on each syllable position. A chi-squared statistic determined the significance of the proportions of words with stress on each syllable containing the given cue, and then these proportions were interrogated to determine which stress position was most strongly associated with the cue. As an example, for the cue -*air* there are 9 words in English that end with these letters, 4 with first syllable stress and 5 with second syllable stress: *a'ffair, cor'sair, de'spair, 'funfair, 'horsehair, im'pair, 'mohair, 'pushchair,* and *re'pair*. The estimated type frequencies for this cue derived from the random baseline (hence dependent on the general distribution of first and second syllable stress words in the corpus) are that there would be 7.6 words ending in –*air* with first syllable stress and 1.4 with second syllable stress. So, -*air* is more strongly associated with second syllable stress than first syllable stress.

The final stage of the method was to assign stress position to the words in the corpus according to the association between the cue and stress position. In the example of –*air*, for instance, all words ending in –*air* would be classified as second syllable stress in the analyses. The final classification accuracy, once all cues had been considered, was determined by counting the correct classifications based on all the cues and measuring

Cramer's V (ϕ_c) on the classification. Cramer's V results in a value between 0 and 1 indicating the strength of the association between the stress position derived from the cue classifications and the actual stress position of words in the corpus. The value of ϕ_c is related to the number of observations, so comparisons between values of ϕ_c are only valid within each language. Both type and token analyses were performed, but for reasons of conciseness we report here only the token analyses as the type analyses were similar in the proportions of words correctly classified.

Results

Table 2 shows the results in terms of correct classification of bisyllabic and trisyllabic words for English, along with the ϕ_c value for the classification.

Insert Table 2

The results replicate previous studies showing that both the beginnings and endings (CV_/_VC) for words in English assessed by Arciuli and colleagues were highly significant in accurately classifying bisyllabic and trisyllabic words according to stress position (Arciuli & Cupples, 2006; 2007). Indeed, for classifications derived from each cue type, and in all languages, for the ϕ_c values, $p < .001$. Also replicating previous work focusing on the CV_/_VC analyses, we found that endings were more accurate at classifying words according to stress position than were beginnings in English (Arciuli et al., 2010).

In terms of the other types of cue for English, as with the CV_/_VC cue, both the initial and the final parts of words were highly significant in terms of classifying words according to their stress position, even when this was just a single letter. The effectiveness of first or last vowel in classifying words according to stress position is consistent with Yap and Balota's (2009) discovery that the first vowel provided information about naming times in polysyllabic reading. However, we found that the first or last consonant, or even just the first or last letter was even more effective in classifying words according to their stress position. Hence, just a single letter can provide substantial information about stress position, and the prominence of first and last letters in visual word recognition means that such information is available at an early stage to the reading system (Beech & Mayall, 2005; Shillcock & Monaghan, 2001). Unsurprisingly, more information from the word provides increasingly accurate classification, with corresponding increasing ϕ_c values. However, providing more orthographic information about the word also increases the information that has to be processed, an issue that will be addressed in the second study when *efficiency* as well as *accuracy* of encoding is considered.

Tables 3 and 4 show the results for the Germanic languages of Dutch and German. Both languages have similar distributions of stress across syllables (a dominant trochaic pattern) and also have complex onsets and offsets in syllables. The results for these languages indicate that stress is encoded in orthographic cues to a high degree of accuracy in Germanic languages other than English. For all cue types, for both bisyllabic and trisyllabic words, accuracy of classification was highly significant, and once again increased as the length of the cue increased. The advantage for word endings as indicating stress position for English, however, was not observed in Dutch and German. For bisyllabic words, the pattern was mixed with all cues except for 1^{st}/last 2, 3, 4, or 5 letters in German and 1^{st}/last 4 letters in Dutch, resulting in greater accuracy of classification based on word endings than beginnings, similar to the pattern discovered for English. However, for trisyllabic words, for all cues, word *beginnings* were better indicators than endings of stress position for both Dutch and German.

For English, there are a greater number of distinct endings than beginnings. For trisyllabic words with the CV_/_VC cue, there are 789 distinct beginnings and 1411 distinct endings. For Dutch, the relative frequencies are reversed – there are 825 distinct beginnings and 771 distinct endings. However, this cannot explain the reversed pattern between German and English, as German too has a larger number of distinct endings than beginnings for trisyllabic words: 618 beginnings and 783 endings.

Insert Tables 3 and 4

The results of the analyses of Italian and Spanish are shown in Tables 5 and 6, respectively. For Italian and Spanish, final-syllable stress is indicated in the orthography with an accent, and so the focus for these analyses is on distinguishing first from second syllable stress for the trisyllabic words. We nevertheless include the analysis for these bisyllabic words, disregarding the diacritic when present, hence, performance was not 100% accurate for these words. The results for both Italian and Spanish demonstrate reliable orthographic indicators of stress position, as with the Germanic languages in Tables 2, 3 and 4. For Italian and Spanish, both beginnings and endings were highly significant indicators of stress, despite these being languages with fewer opportunities to individuate stress position based on orthography due to a simpler syllabic structure than the Germanic languages tested.

For Spanish, comparing beginnings versus endings, the results were similar regardless of cue type: The word endings provided more accurate classification than did word beginnings, for all cue types. For Spanish, accurate classification was based on the 1^{st}/last three letters, with slight improvements resulting from extending this to 4 or 5 letters, though the last consonant in the word also proved highly accurate in determining stress position. For Italian, the benefit of beginnings and endings varied according to the cue type. The 1^{st}/last 2, 3, 4 or 5 letters or consonant resulted in an advantage for endings over beginnings (and also resulted in the most accurate classifications) but for the CV_/_VC analyses and the 1^{st}/last vowel or letter there was an advantage for beginnings over endings in terms of accurate classifications. The most accurate classification was again for the largest cue – the first or last three letters. The high degree of accuracy of the last three letters for determining stress position is consistent with the stress neighbourhoods analyses based on the final VCV sequences of words in Italian (Burani & Arduino, 2004; Colombo, 1992), which have been shown to affect visual word processing of words and nonwords in Italian.

Insert Tables 5 and 6

The final analysis, for Greek, demonstrated that letter identity in both beginnings and endings provides valuable information about stress position, even for a language that also indicates stress position using diacritics. Thus, there is redundant information in Greek orthography regarding stress, and this can be potentially used in situations where the diacritics are not available (as in capitalised words, Protopapas et al., 2007). For bisyllabic words, the general result is that word beginnings are more accurate than word endings in indicating stress position for all cue types, but for trisyllabic words, word endings tend to be more informative (except for the 1^{st}/last consonant cue type). As with the other languages, orthographic cues can prove to be highly accurate in reflecting stress position.

Discussion

One of the greatest challenges facing researchers of word reading is to extend what we know about reading aloud for English monosyllables to accommodate polysyllabic words and languages other than English. Here we undertook corpus analyses to assess one critical issue in polysyllabic word reading – to investigate sources of information in orthography that contribute to stress assignment in polysyllabic reading. To extend previous work on orthographic indicators of stress assignment reported for English and Italian, we systematically investigated probabilistic orthographic cues to lexical stress in polysyllabic words in 6 languages: English, Dutch, German, Italian, Spanish, and Greek.

Our analyses revealed that orthographic cues, which are not wholly morphologically derived, provide information about stress position in languages that vary in terms of syllabic complexity and transparency of stress in the orthography (e.g., diacritics). However, this cross-linguistic comparison also demonstrates that it is not the case for all languages that endings are more effective in indicating stress position than beginnings. This was the case for English and Spanish, but was reversed for Dutch and German trisyllables, and was mixed for Italian and Greek, according to which cue type, or granularity of the orthography, was assessed.

It is important to note that these analyses have focused only on accuracy of classification. They do not take into account the complexity of the information on which the classifications are based. Yet, comparing between different types of cue for their effectiveness in reflecting stress positions cannot be based purely on accuracies. This is because the more information that is available the more accurate the classification will be. So, a 3 letter cue provides more possibility for correctly classifying words than just a single letter, and a large set of cues provides more possibilities for accurate encoding than a smaller set of cues (as in comparisons of beginnings and endings, for instance). The extreme situation would be to take the whole word as a cue – the first or last 20 letters, for instance – to provide almost 100% accurate classification (only homographs such as *project,* which can be pronounced with first or second syllable stress, would still be misclassified). However, this would fail to capture useful generalisations in the orthography that may indicate stress position with less information. Thus, an important next step for future research is to assess the *optimal* granularity of the reading system for determining stress position by taking into account the complexity of the cue itself in contributing to indicating stress assignment.

There have been some attempts to describe the relative grain-size for letter to phoneme mappings that may be effective in different languages (Ziegler & Goswami, 2005) and experiments have tested the potential for varying grain-sizes within a single language according to orthographic properties of words (Rey, Ziegler, & Jacobs, 2000; Smith & Monaghan, 2012). Computational models of reading have suggested potential grain-sizes within a language for orthography to phonology mappings. These suggestions are either explicit in that they state the graphemes that may act as units of analysis in the orthography (Coltheart et al., 2001), or implicit in models that facilitate acquisition of the useful level of analysis and then interpret patterns of behaviour related to consistency or regularity of orthography to phonology mappings as indicators of multiple simultaneous grain-sizes discovered by the model (Harm & Seidenberg, 1999; Perry, Ziegler, & Zorzi, 2007, 2010; Seidenberg & McClelland, 1989).

Recently a precise computational statement of the most efficient grain-size for word naming for segmental phonology in English has been provided. Vousden et al. (2011) provided a minimum description length analysis of English orthography to phonology mappings to compare whether the most efficient mapping is based on whole words, head-code, onset-body, or graphemes. Minimum description length provides an analysis of the length of the code required to describe a theory about data together with the length of code required to describe the data using that theory, and is thus an implementation of the Simplicity Principle, whereby the complexity of a system is given by the code length required to describe the system, and the assumption that the solution discovered by the cognitive system for a given task will be the most efficient available, in other words, it will have the shortest code length. The precise implementation of the code (i.e., the program used to encode the theory and the data) is not critical, as each implementation varies by a constant (Chater & Vitányi, 2003). In the case of mappings from orthography to phonology, the theory is the grain-size of the mapping, and the data is the vocabulary.

Vousden et al. (2011) found that, for vocabularies that were greater in size than 100 words, grapheme to phoneme mappings were the most efficient description of the vocabulary, and that addition of multiple-grapheme chunks, that were often mispronounced using other grapheme-phoneme mappings (such as *dge*->/dʒ/, or *ce*->/s/), improved efficiency of the mapping further. In terms of grain-size for reading, the minimum description length approach demonstrated that optimal for English was a grapheme level of encoding of orthography with some larger multi-letter chunks.

Pronouncing a word in a language with varying stress position requires determining the orthography to stress position mappings in addition to forming orthography to phonology mappings, and this may require an altogether different grain-size than that for phoneme production. One possibility is that stress is stored at the lexical level (Daelemans, Gillis, & Durieux, 1994), thus the whole-word is a candidate for the grain-size for pronunciation of stress in English. Alternatively, Rastle and Coltheart (2000) proposed a set of sublexical rules for assigning stress, which required accessing a database of affixes with their stress position. Such a set of affixes can be interpreted as the proposed grain-size for stress assignment. At a finer granularity still, Arciuli and colleagues' analyses provided a demonstration that there was considerable information in orthography for indicating stress when the portion of the word up to the first vowel, or from the last vowel, was taken as the orthographic cue. Most of these cues were not related to morphological units within the word, in contrast to the approach by Rastle and Coltheart (2000), demonstrating that sublexical, and submorphological orthographic chunks are likely the effective grain size for reading for stress assignment in English. Though Arciuli and colleagues' analyses showed that sets of letters in the word provide substantial information about stress, it may be that sufficient information is contained in the first or last one or two or more letters, the first or last vowel (Yap & Balota, 2009), or the first or last consonant.

Furthermore, the region that is indicative of stress position may vary across different languages. We are currently undertaking further research to address this question by analysing different sequences at the beginning and endings of words for the different languages in terms of the accuracy of the classification based on these cues, as well as determining the description length of encoding each language using different grain-sizes for indicating stress position.

Our findings reported in this Chapter confirm that there is a rich source of probabilistic information pertaining to assignment of lexical stress present in orthography across a wide variety of languages. This information can be used to design behavioural experiments and undertake computational modelling to further advance our knowledge of the reading system. Potentially, this information can also assist in designing appropriate stimuli for educational and therapeutic purposes. For example, a recent study with typically developing children used orthographically biased stimuli to accelerate appropriate production of lexical stress (Van Rees et al., 2012).

References

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

Arciuli, J., & Cupples, L. (2007). Would you rather 'embert a cudsert' or 'cudsert an embert'? How spelling patterns at the beginning of English disyllables can cue grammatical category. In: *Mental States: Language and Cognitive Structure*. Schalley, A. & Khlentzos, D. (Eds.). pp. 213-237. John Benjamins Publishing.

Arciuli, J., & Monaghan, P. (2009). Probabilistic cues to grammatical category in English orthography and their influence during reading. *Scientific Studies of Reading, 13*(1), 73-93.

Arciuli, J., Monaghan, P., & Seva, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioural, and computational investigations. *Journal of Memory and Language, 63*, 180-196.

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

Beech, J.R. & Mayall, K.A. (2005). The word shape hypothesis re-examined: evidence for an external feature advantage in visual word recognition. *Journal of Research in Reading, 3*, 302-319.

Bertinetto, P.M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., & Thornton, A.M (2005). *Corpus e Lessico di frequenza dell'italiano scritto (CoLFIS)* [lexical database].

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.

Chater, N., & Vitányi, P. M. B. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences, 7*, 19–22.

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., & Zevin, J. (2009). Stress Priming in Reading and the Selective Modulation of Lexical and Sub-Lexical Pathways. *PLoS ONE, 4*, e7219. doi:10.1371/journal.pone.0007219

Daelemans, W., Gillis, S., & Durieux, G. (1994). The acquisition of stress: A data-oriented approach. *Computational Linguistics, 20*, 421-451.

De Mauro, T. (2000). *Il Dizionario della Lingua Italiana*, Paravia.

Ellison, T.M. (1992). *The Machine Learning of Phonological Structure*. Unpublished PhD thesis, University of Western Australia.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*, 153-189.

Grünwald, P. (1994). A minimum description length approach to grammar inference. In G. Scheler, S. Wermter, & E. Riloff (Eds.). *Connectionist, statistical and symbolic approaches to learning for natural language* (203-216). Berlin: Springer Verlag.

Goswami, U., Ziegler, J., Dalton, L., & Schneider, W. (2003). Nonword reading across orthographies: How flexible is the choice of reading units? *Applied Psycholinguistics, 24,* 235-247.

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

Jared, D. & Seidenberg, M. S. (1990). Naming multisyllabic words. *Journal of Experimental Psychology: Human Perception & Performance, 16*, 92–105.

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

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

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–75.

Paap, K. R., & Noel, R. W. (1991). Dual-route models of print to sound: Still a good horse race. *Psychological Research, 53*, 13-24.

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

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

Protopapas, A. (2006). On the usefulness of stress diacritics in reading Greek. *Reading and Writing, 19*, 171-198.

Protopapas, A., & Gerakaki, S. (2009). Development of processing stress diacritics in reading Greek. *Scientific Studies of Reading, 13*, 453-483.

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

Protopapas, A., Gerakaki, S., & Alexandri, S. (2007). Sources of information for stress assignment in reading Greek. *Applied Psycholinguistics, 28*, 695–720.

Protopapas, A., & Vlahou, E. L. (2009). A comparative quantitative analysis of Greek orthographic transparency. *Behavior Research Methods, 41*, 991–1008.

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.

Rey, A., Ziegler, J. C, & Jacobs, A. M. (2000). Graphemes are perceptual reading units. *Cognition, 75*, B1-B12.

Rissanen, J. (1982). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, *11*, 416-431.

Sebastian, N., Marti, M.A., Carreiras, M.F. & Cuetos, F. (2000). *LEXESP: Léxico informatizado del Español*. Barcelona, Universidad de Barcelona.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523-568.

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

Seymour, P. H. K., Aro, M., Erskine, J. M. and collaboration with COST Action A8 network (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143–174.

Shillcock, R.C. & Monaghan, P. (2001). The computational exploration of visual word recognition in a split model. *Neural Computation, 13*, 1171-1198.

Smith, A.C. & Monaghan, P. (2012). What are the functional units in reading? Evidence for statistical variation influencing word processing. In E. Davelaar (Ed.), *Connectionist models of behaviour and cognition, Volume 12*. Singapore: World Scientific.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana IL: University of Illinois Press.

Sulpizio, S., Arduino, L. S., Paizi, D., & Burani, C. (2012). Stress assignment in reading Italian polysyllabic pseudowords. *Journal of Experimental Psychology: Learning, Memory & Cognition.*

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

Taraban, R., & McClelland, J.L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language*, *25*, 608-631.

Van Rees, L., Ballard, K., McCabe, P., Macdonald-D'Silva, A., & Arciuli, J. (2012). Training production of lexical stress in typically developing children with orthographically biased stimuli and principles of motor learning. *American Journal of Speech-Language Pathology*, *21*, 197-206.

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

Ziegler, J.C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131,* 3–29.

Acknowledgements

This research was supported by a bilateral funding scheme: ESRC grant RES-000-22-1975 (UK) and ARC grant LX0775703 (Australia) awarded to the first two authors. Thanks to Cristina Burani, Lucia Colombo, Fernando Cuetos, Rob Davies, Nicolas Gutiérrez-Palma, and Athanassios Protopapas for helpful discussions and for providing access to corpora.

Table 1. Number of bisyllabic and trisyllabic word types from each corpus with stress position (parentheses show percentage of words with each stress position).

Corpus	Bisyllables		Trisyllables		
	Stress Position		Stress Position		
	1st	2nd	1st	2nd	3rd
English	18571 (85%)	3287 (15%)	8486 (58%)	5295 (36%)	857 (6%)
Dutch	10108 (74%)	3629 (26%)	11579 (58%)	6627 (33%)	1780 (9%)
German	16469 (83%)	3380 (17%)	15751 (69%)	6328 (28%)	859 (4%)
Italian	1885 (99%)	23 (1%)	968 (15%)	5295 (81%)	256 (4%)
Greek	5686 (42%)	7948 (58%)	11398 (28%)	15565 (38%)	14291 (35%)
Spanish	11672 (71%)	4835 (29%)	19 (0%)	32746 (83%)	6513 (17%)

Table 2. Classification accuracy of different cue types for English tokens, for reflecting stress position for bisyllabic and trisyllabic words.

	Bisyllabic			Trisyllabic			
	Stressed Syllable			Stressed Syllable			
	1st	2nd	ϕ_c	1st	2nd	3rd	ϕ_c
CV_ / _VC							
beg	.79	.93	.61	.57	.70	.73	.39
end	.90	.92	.74	.77	.80	.94	.59
1st/last C							
beg	.61	.86	.39	.53	.39	.52	.25
end	.73	.78	.42	.54	.52	.77	.32
1st/last V							
beg	.62	.70	.26	.30	.58	.60	.14
end	.88	.35	.24	.55	.74	.31	.34
1st/last letter							
beg	.74	.74	.41	.53	.58	.47	.23
end	.65	.80	.37	.62	.38	.67	.27
1st/last 2 letters							
beg	.78	.92	.60	.66	.69	.79	.42
end	.78	.85	.54	.62	.77	.62	.41
1st/last 3 letters							
beg	.87	.90	.82	.78	.75	.84	.61
end	.78	.90	.73	.71	.83	.91	.57
1st/last 4 letters							
beg	.92	.98	.92	.82	.84	.96	.89
end	.88	.96	.88	.80	.88	.98	.91
1st/last 5 letters							
beg	.99	.99	.97	.95	.95	.98	.85
end	.98	.99	.96	.96	.97	.99	.94

Table 3. Classification accuracy of different cue types for Dutch tokens, for reflecting stress position for bisyllabic and trisyllabic words.

	Bisyllabic Stressed Syllable			Trisyllabic Stressed Syllable			
	1st	2nd	ϕ_c	1st	2nd	3rd	ϕ_c
CV_ / _VC							
beg	.75	.90	.58	.60	.68	.81	.52
end	.88	.96	.78	.55	.63	.93	.52
1st/last C							
beg	.83	.57	.40	.39	.63	.70	.40
end	.85	.68	.51	.57	.48	.66	.33
1st/last V							
beg	.75	.64	.36	.45	.65	.53	.36
end	.81	.88	.63	.21	.74	.78	.34
1st/last letter							
beg	.73	.63	.32	.40	.60	.66	.36
end	.82	.68	.47	.53	.53	.69	.29
1st/last 2 letters							
beg	.81	.84	.59	.76	.66	.75	.54
end	.86	.84	.66	.46	.70	.84	.44
1st/last 3 letters							
beg	.91	.95	.81	.82	.84	.85	.70
end	.90	.97	.82	.58	.79	.93	.60
1st/last 4 letters							
beg	.95	.99	.89	.88	.89	.91	.79
end	.95	.99	.90	.77	.83	.95	.76
1st/last 5 letters							
beg	.99	.99	.97	.92	.94	.95	.85
end	.98	.99	.95	.84	.91	.98	.85

Table 4. Classification accuracy of different cue types for German tokens, for reflecting stress position for bisyllabic and trisyllabic words.

	Bisyllabic			Trisyllabic			
	Stressed Syllable			Stressed Syllable			
	1st	2nd	ϕ_c	1st	2nd	3rd	ϕ_c
CV_ / _VC							
beg	.90	.88	.65	.86	.80	.69	.65
end	.92	.95	.75	.60	.69	.95	.61
1st/last C							
beg	.86	.70	.48	.45	.60	.77	.43
end	.90	.69	.54	.56	.58	.63	.34
1st/last V							
beg	.75	.68	.33	.59	.79	.52	.48
end	.83	.85	.54	.88	.88	.82	.30
1st/last letter							
beg	.82	.70	.42	.64	.78	.59	.47
end	.85	.74	.50	.42	.55	.71	.26
1st/last 2 letters							
beg	.91	.85	.58	.87	.78	.75	.65
end	.89	.91	.52	.48	.77	.87	.50
1st/last 3 letters							
beg	.95	.95	.71	.90	.87	.90	.76
end	.92	.98	.62	.69	.75	.95	.64
1st/last 4 letters							
beg	.99	.98	.94	.94	.95	.95	.86
end	.97	1.00	.91	.78	.85	.98	.78
1st/last 5 letters							
beg	.99	1.00	.98	.96	.97	.98	.91
end	.99	1.00	.96	.86	.92	.99	.87

Table 5. Classification accuracy of different cue types for Italian tokens, for reflecting stress position for bisyllabic and trisyllabic words.

	Bisyllabic Stressed Syllable			Trisyllabic Stressed Syllable			
	1st	2nd	ϕ_c	1st	2nd	3rd	ϕ_c
CV_ / _VC							
beg	.89	.99	.19	.63	.58	.94	.28
end	1.0	1.0	.88	.77	.35	.90	.14
1st/last C							
beg	.81	.93	.13	.44	.32	.84	.11
end	.99	.18	.14	1.0	.00	.62	.14
1st/last V							
beg	.60	.97	.08	.66	.33	.45	.11
end	1.0	.96	.86	.73	.35	.49	.09
1st/last letter							
beg	.94	.92	.24	.57	.31	.73	.12
end	.92	1.0	.22	.70	.31	.85	.09
1st/last 2 letters							
beg	.92	.99	.23	.65	.60	.95	.26
end	1.0	1.0	.84	.83	.58	.90	.27
1st/last 3 letters							
beg	.97	1.0	.39	.87	.83	.99	.46
end	1.0	1.0	.98	.97	.88	.96	.56
1st/last 4 letters							
beg	.98	1.0	.47	.97	.95	.99	.66
end	1.0	1.0	1.0	.99	.96	.98	.71
1st/last 5 letters							
beg	1.0	1.0	1.0	1.0	.99	1.0	.79
end	1.0	1.0	1.0	1.0	.99	1.0	.81

Table 6. Classification accuracy of different cue types for Spanish tokens, for reflecting stress position for bisyllabic and trisyllabic words.

	Bisyllabic			Trisyllabic			
	Stressed Syllable			Stressed Syllable			
	1^{st}	2^{nd}	ϕ_c	1^{st}	2^{nd}	3^{rd}	ϕ_c
CV_ / _VC							
beg	.74	.66	.31	.98	45	70	.14
end	.97	.97	.95	.99	.97	.76	.57
1^{st}/last C							
beg	.47	.68	.11	.88	.37	.70	.10
end	.99	.71	.80	.84	1.0	.74	.63
1^{st}/last V							
beg	.56	.64	.15	.77	.45	.21	.04
end	.92	.31	.26	.68	.60	.55	.20
1^{st}/last letter							
beg	.68	.53	.15	.88	.45	.60	.09
end	.99	.79	.83	.99	.99	.75	.61
1^{st}/last 2 letters							
beg	.72	.68	.30	.51	.51	.68	.15
end	.94	.88	.75	.96	.96	.78	.57
1^{st}/last 3 letters							
beg	.84	.87	.58	.63	.63	.80	.24
end	.94	.94	.79	.94	.94	.93	.59
1^{st}/last 4 letters							
beg	.89	.95	.71	1.0	.74	.87	.33
end	.97	.98	.90	.99	.96	.96	.73
1^{st}/last 5 letters							
beg	.96	.98	.86	1.0	.83	.91	.46
end	.98	.99	.93	1.0	.98	.98	.83

Table 7. Classification accuracy of different cue types for Greek tokens, for reflecting stress position for bisyllabic and trisyllabic words.

	Bisyllabic			Trisyllabic			
	Stressed Syllable			Stressed Syllable			
	1st	2nd	ϕ_c	1st	2nd	3rd	ϕ_c
CV_ / _VC							
beg	.81	.85	.65	.57	.48	.64	.34
end	.79	.60	.37	.63	.58	.63	.43
1st/last C							
beg	.84	.33	.19	.40	.62	.27	.15
end	.74	.29	.04	.30	.14	.78	.14
1st/last V							
beg	.60	.82	.43	.50	.26	.54	.16
end	.65	.61	.26	.58	.55	.64	.41
1st/last letter							
beg	.69	.75	.43	.43	.35	.54	.17
end	.64	.62	.25	.55	.34	.57	.24
1st/last 2 letters							
beg	.81	.81	.60	.59	.51	.60	.35
end	.75	.78	.52	.67	.57	.78	.51
1st/last 3 letters							
beg	.94	.89	.81	.78	.66	.74	.59
end	.89	.89	.77	.84	.75	.86	.72
1st/last 4 letters							
beg	.98	.97	.95	.85	.83	.87	.73
end	.96	.96	.92	.93	.91	.93	.88
1st/last 5 letters							
beg	.95	.98	.97	.93	.81	.93	.84
end	.95	.99	.97	.98	.97	.98	.96