358,534 nonwords: The ARC Nonword Database

Kathleen Rastle, Jonathan Harrington, and Max Coltheart

Macquarie University, Sydney, Australia

The authors present a model of the phonotactic and orthographic constraints of Australian and Standard Southern British English monosyllables. This model is used as the basis for a web-based psycholinguistic resource, the ARC Nonword Database, which contains 358,534 monosyllabic nonwords—48,534 pseudohomophones and 310,000 non-pseudohomophonic nonwords. Items can be selected from the ARC Nonword Database on the basis of a wide variety of properties known or suspected to be of theoretical importance for the investigation of reading.

A great deal of experimental research has investigated relationships between various psycholinguistic properties of words and indices of reading behaviour such as reading aloud latency or visual lexical decision latency. Such work requires careful selection of words so that they vary appropriately on the property of interest and are matched appropriately on other, potentially confounding, variables. The MRC Psycholinguistic Database (Coltheart, 1981) was developed as a tool to facilitate such word selection. Words can be selected from that database on a variety of psycholinguistic criteria. A particularly convenient World-Wide-Web (WWW) version of this database is now available at http://www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm.

That database contains only words, however, and that is a limitation as the study of nonword reading and its impairment (cf., phonological dyslexia, e.g., Coltheart, 1996) has also proven useful in understanding the mechanisms that underlie visual word recognition and reading aloud. Given the influential view that the generation of phonology from print involves both addressed and assembly procedures (see Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001, for a review), studying the nature of phonological assembly in its purest form (nonword reading aloud) should pose important constraints on a model of reading. Such constraints have, of course, emerged: For example, the effect of neighbourhood size on the reading aloud of nonwords (e.g., McCann & Besner, 1987; Peereman & Content, 1995) may indicate that processing in a model of reading is cascaded rather than thresholded (Coltheart & Rastle, 1994;

Requests for reprints should be sent to Kathleen Rastle, Department of Psychology, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, UK. Email: kathy.rastle@rhul.ac.uk

We are grateful to Steve Harrison, Steve Saunders, and Breck Thomas for programming assistance, to Carl Windhorst for research assistance, and to Conrad Perry for helpful discussion.

http://www.tandf.co.uk/journals/pp/02724987.html DOI:10.1080/02724980244000099

Coltheart et al., 2001); the finding that pseudohomophones (nonwords that sound like words, e.g., "brane") are read aloud more quickly (e.g., McCann & Besner, 1987; Taft & Russell, 1992) and rejected in lexical decision more slowly (e.g., Coltheart, Davelaar, Jonasson, & Besner, 1977) than control nonwords (e.g., "brame") may indicate that processing in the reading system is interactive; and the finding that word length strongly affects nonword reading aloud latency (Weekes, 1997) may indicate that the phonological assembly procedure operates serially (Rastle & Coltheart, 1998). Moreover, the finding that the inclusion of nonword fillers in reading aloud experiments can change substantially the pattern of word-naming effects (e.g., Baluch & Besner, 1991; Monsell, Patterson, Graham, Hughes, & Milroy, 1992; Tabossi & Laghi, 1992) may indicate that the procedures by which phonology is derived are under strategic control (but see Lupker, Brown, & Colombo, 1997, for a different view). Finally, there is good evidence that the procedures by which lexical decisions are made can be influenced by the characteristics of nonword distracters (e.g., Forster & Veres, 1998; Pugh, Rexer, Peter, & Katz, 1994).

Given such findings, it is clear that the selection of nonword target, filler, and distractor experimental stimuli must be accomplished systematically and with great care. To this end, we have created a database of monosyllabic nonwords accessible via a WWW interface, which can be used to select nonwords and pseudohomophones on the basis of a number of psycholinguistic dimensions. The database URL is http://www.maccs.mq.edu.au/~nwdb. The interface was modelled after the WWW interface for the MRC Psycholinguistic Database referred to earlier. Users are asked to specify criteria for each nonword search in the form of minimum and maximum values on a range of psycholinguistic variables (e.g., number of letters, number of neighbours, bigram frequency). Users also specify which of these variables they would like displayed in the output. The database returns appropriate nonwords from its pool of 48,534 pseudohomophones and 310,000 non-pseudohomophonic nonwords.

Some type of random selection from the database is, of course, necessary in order to ensure that repeated searches using the same criteria result in differing output. However, given the size of the database, it would be prohibitively slow to randomly reorder the database on the initiation of each new search. Therefore, a method of pseudorandomization is employed: The database is randomly ordered, and that random order is fixed; the randomly ordered database is entered at some random point, a point computed independently each time a new search is initiated; from this random point, the first *N* items meeting the criteria specified are returned (where *N* is the number of items requested by the user). This method of randomization makes it unlikely that any two searches will generate overlapping output, unless the particular search is very constrained.

The database described here contains what we have deemed to be *legal* nonwords. It is commonplace for researchers to use in their experiments nonwords that they characterize as legal, either orthographically or phonotactically, and to use the notion of nonword legality in their theoretical explanations of psycholinguistic phenomena. However, to our knowledge, legality in this context has not previously been properly explored or defined, and, indeed, in exploring this notion many interesting issues come to light. The path that we followed in generating legal nonwords was twofold. We first developed a corpus of 68,497 phonological monosyllables based upon an original phonotactic grammar; we consider these to be legal syllables of both Australian English (Harrington, Cox, & Evans, 1997) and Standard Southern British English (Deterding, 1997; Harrington, Palethorpe, & Watson, 2000), to the extent that these two accents can be considered to be phonemically equivalent (Wells, 1982). Orthographic representations for these syllables, which we consider to be legal based upon the sound–spelling and spelling–sound relationships that characterize Australian and Standard Southern British English (see, e.g., Baayen, Piepenbrock, & van Rijn, 1993; Rastle & Coltheart, 1999b), were then derived.

PHONOTACTICALLY LEGAL SYLLABLES

Phonotactic constraints: General considerations

All languages have restrictions on the way in which phonemes can (legally) be sequentially arranged in a syllable known as *phonotactic constraints*. Although /plɪg/ is a nonword in English, its phonemes do not violate any phonotactic constraints: There are words that begin with /pl/ ("please"), words that end in /ɪg/ ("dig") and, although /plɪ/ is rare, it does occur ("plinth"). On the other hand, the sequence /mzoibf/, although entirely pronounceable, violates at least three phonotactic constraints of English: There are no words that begin with /mz/, none that end with /bf/, and no monomorphemic words that have syllable-internal /oib/ or /zoi/ (see Appendix A for our system of phonemic transcription).

Although phonotactic constraints vary across languages (e.g., German allows syllables to begin with the phonemes /kn/ whereas English does not), they are often strongly influenced by the sonority of segments (Jespersen, 1904; Saussure, 1916; Sievers, 1881) and their sequential arrangement, which gives rise to a *sonority profile* (Blevins, 1995; Clements, 1990; Hooper, 1972; Kiparsky, 1981; Lowenstamm, 1981; Rice, 1992; Selkirk, 1984; Zec, 1995; Zwicky, 1972). A segment's sonority can be defined in terms of the extent to which the vocal tract is constricted in speech production (Beckman, Edwards, & Fletcher, 1992; de Jong, 1995). For example, voiceless oral stops and vowels (especially open vowels) are at opposite ends of the sonority scale because in producing /t/, the vocal tract is tightly constricted at a certain stage of its production (low sonority) whereas in /[/, the vocal tract is wide open, and acoustic energy can radiate from the lips. An acoustic–perceptual correlate of sonority is also loudness: Sounds with high sonority are generally louder than those with low sonority (Lindblom & Sundberg, 1971). The sonority scale from least to greatest sonority is usually taken to be oral stops < fricatives < nasal stops < liquids < glides < vowels, where < denotes "has less sonority than".

There is a preference in languages for syllables to be made up of phonemes such that the sonority rises from the left syllable margin to a peak and falls from the peak to the right syllable margin. Therefore, languages are more likely to have syllables that begin with /fl/ than /lf/ because the former, but not the latter, conforms to the sonority profile. For the same reason, syllable-final /lf/ is preferred to /fl/ (or if word-final /fl/ occurs it is usually bisyllabic as in English "waffle"). It must be emphasized, however, that the relationship between sonority and phonotactics is a strong tendency, not a rule: For example, in English there are many clusters with /s/ (e.g., "sport", "task") and some oral stop clusters ("act") that do not conform to the sonority profile, although various phonologists (e.g., Fudge, 1969; Halle & Vergnaud, 1980; Rubach & Booij, 1990; Steriade, 1982) have argued that the sonority profile does not apply to consonants at the edges of a word: That is, those consonants are outside the domain of syllabification and sonority (Itô, 1986; Kaye, 1990; McCarthy & Prince, 1990).

In discussing the phonotactics of English, syllables are often subdivided into an *onset* that includes all prevocalic consonants and a *rhyme* that includes the vowel or diphthong and following consonants. The rhyme branches into a *nucleus*, which is the vowel or diphthong, and a *coda*, which includes any following consonants (Cairns & Feinstein, 1982; Halle & Vergnaud, 1980; Kenstowicz, 1994; Mohanan, 1985; Pike & Pike, 1947; Trubetzkoy, 1939/1958). This hierarchical division of the syllable is in part motivated by the strength of the phonotactic constraints within and across syllabic constituents. There are tighter restrictions *within* constituents than *across* constituents; that is, fewer constraints apply to the rhyme than to either the onset or the coda, and, indeed, there are perhaps no constraints that restrict the co-occurrence of onsets and rhymes.

A grammar for monosyllables

In this section, we present a grammar for English monosyllables, from which we derive monomorphemic forms and describe how these forms can be augmented with the addition of various inflectional and derivational suffixes. Our grammar for monosyllables follows some of the same principles, if not details, that are set out in Coleman (1998). The first principle is one of maximal generalization, in which the smallest number of rules can account for the maximum number of possible syllables. The second principle follows in part from the first: There should be a minimum number of negative constraints, or filters, which provide exceptions to these generalizations. Some of these exceptions are as follows.

1. This grammar is based on isolated word, citation-form speech (so we do not account for fast-speech forms such as /mri/ for “Marie” that are excluded in citation-form speech).
2. This grammar does not derive sequences such as /ʃn/ (“Schnapps”), /km/ (“Khmer”), and /oiŋ/ (“boing”) that are attested only in loan words, names, or onomatopoeic forms. Words that occur only in archaic forms such as “couldst” are also excluded.
3. Phonotactic oddities that violate phonotactic generalizations are also excluded. For example, /eips/ is recognized as phonotactically odd (and additionally, it is restricted to “traipse”), and only the derived forms “sixths” and “twelfths” have four consonant phonemes following the nucleus, of which the final two are /θs/. Our aim here is not to filter forms that occur only in a handful of words (for example, we derive final /lb/, even though this is restricted to “alb” and “bulb”), but rather to exclude those that are recognised as being exceptions to phonotactic principles. We list all exceptions to the rules at various stages of derivation.

The syllables that we generate are presented in terms of a binary categorization: A given syllable is or is not phonotactically legal. We recognize, however, that there is much recent research to show that native speakers rank the acceptability of phoneme sequences probabilistically based on frequency of occurrence in the lexicon (e.g., Coleman & Pierrehumbert, 1997; Hay, Pierrehumbert, & Beckman, in press; Pierrehumbert, 1994), rather than performing a binary all-or-none categorization. Much research needs to be carried out to determine whether, for example, acceptability judgements of the monosyllables with repetitive sequences like /skɪk/ (which do not occur in English, as discussed later) are predictable from the phonological statistics in the lexicon. Currently, we generate such sequences—that is, they are declared to be possible nonwords—although we do not generate monosyllables with rhymes like /eilm/ because they are considered to be in violation of a phonotactic constraint of English. In subsequent versions of our nonword model, it may be more appropriate to derive both these kinds of sequence and to match them to a probability model based on judgements of acceptability, rather than to assign them, as we currently do, to two separate categories (possible and impossible nonword).

Further aspects of grammar

Onsets

The set of possible syllable onsets is both well understood and extensively researched (e.g., Cairns & Feinstein, 1982; Clements, 1990; Clements & Keyser, 1983; Selkirk, 1984). The following consonants can occur in single-constituent onsets:[1]

1. voiceless stops: /p t k/
2. voiced stops: /b d g/
3. voiceless fricatives: /f θ s ʃ/
4. nasals: /m n/
5. voiced fricatives /v ð z/
6. affricates /t͡ʃ d͡ʒ/
7. /h/
8. /j/
9. glides: /w r l/

Categories 1–3 can combine with the glides in 9 to form 30 onsets with two constituents. A further 8 two-constituent onsets are formed by combining /s/ with Categories 1, 4, and 9. However, a filter that disallows any onset in which the two consonantal onset phonemes have the same place of articulation excludes /pw bw fw/ (both consonants are labial) and /tl dl θl/ both are dental/alveolar). Coleman (1998) suggested that a place of articulation constraint can also be applied to exclude /sr ʃl ʃw/. Finally, although /sf/ occurs in a small number of words ("sphere" and its morphological relatives, and "sphinx"), we have excluded it from the set of possible two-constituent onsets because (1) it would be the only C_1C_2 onset in which both consonants are fricatives, and (2) it violates the tendency for languages to build onsets with consonants that are maximally different on the sonority scale (Clements, 1990; Steriade, 1982).

Three-constituent onsets can be formed by combining /s/ with any legal two-constituent onsets that have voiceless stops in the first position, but only after the place of articulation filter has been applied (so /spw/ is not derived). Additionally, /stw/ does not occur in English and must be excluded (this is presumably an accidental gap) leaving six onsets with three constituents. Within these, /skl/ is restricted to the single morpheme "sclerosis" and its relatives and does not occur in English monosyllables; hence, it has been excluded.

[1] The motivation for these groupings is partly because these are natural phonetic classes, but also because these are needed to account for onsets with two and three constituents.

We have not so far considered sequences with /j/ in the second or third constituent of the onset. These are all restricted to sequences with an /u:/ nucleus ("few", "news", etc.) being derived historically from a close front rounded vowel in Old English and Norman French. /C_1ju:/ sequences occur where the initial consonant is an oral or nasal stop (although /bju: gju:/ are generally restricted to polysyllabic words) and where C_1 is one of the fricatives /f v θ s h/. The only approximant that can occur in /C_1j/ position is /l/ ("lewd"). In three-consonant /C_1C_2ju:/ sequences, the first two consonants are restricted to /sp st sk/ ("spew", "stew", "skew"). These constraints add a further 14 /C_1j/ and 3 /C_1C_2j/ sequences to the set of onsets that can precede an /u:/ nucleus (excluding loan words like "fjord").

Nuclei

We recognize three different types of nucleus. First, English has a set of lax vowels /ɪ e o ʌ [/ that can be defined on phonological distributional grounds: They never occur in open monosyllables (syllables that end in a vowel; so there are no words like /p[/ or /ko/). The *lax vowels* are in many English accents phonetically short (e.g., "brick", "back") and can be distinguished from the *tense vowels* /i: u: o: ɑ: ə: ai ei au ou/, which do occur in open syllables and which are often phonetically long (e.g., "brake", "bake"). The third category includes the *falling* or *centring diphthongs* /ɪe eə ʊə /. These can be distinguished from tense and lax vowels on phonological distributional grounds because, with a small number of exceptions ("beard", "fierce", "pierce", "scarce", "weird"), they only occur in open syllables (as discussed in Coleman, 1998, this is because they function as rhymes, rather than nuclei).

There are two vowels not included in these lists: /ʊ/ and /oi/. Although these vowels can be categorized as phonologically lax and tense, respectively, the distribution in their combination with the possible codae is sufficiently defective that we treat them separately in building the set of possible syllables.

Codae

Any consonant phoneme that can occur in onset position is also found as a single-constituent coda except /w j h/ and, in non-rhotic[2] accents, /r/. We also exclude /ʒ/, which is restricted to a very small number of loan words ("plage", "beige", "rouge"), and this means that /ʒ/ is excluded entirely from the grammar, as it does not occur in onset position.

Most phonemes that can occur in single-constituent coda position are also permitted as the second constituent. The exceptions include /l/ (so there are no monomorphemic syllables with rhymes like /VC_1l/), which is possibly excluded because there is no consonant with more sonority that could occupy the C_1 position (the exception in rhotic accents is the sequence /rl/ in words like "curl"). /ð/ is excluded in C_2 coda position, and we also exclude

[2] A rhotic English accent (e.g., Scots English or General American) is one that has the phoneme /r/ preceding consonants or before a pause. In a non-rhotic accent (e.g., Australian English or Southern British English), the phoneme /r/ can occur before vowels, but not before consonants or pauses. For example, in a rhotic accent, the word "bird" contains four phonemes (where the third phoneme is /r/); in a non-rhotic accent, the word "bird" contains only three phonemes, /b/, a vowel, and /d/. Similarly, an isolated production of the word "car" in a rhotic accent would end in /r/, whereas it would end in a long vowel in a non-rhotic accent.

/ʃ/ from C_2 position, which only occurs in the name "Welsh". This exclusion of /ʃ/ presupposes, however, that we represent the coda of words like "bench" and "conch" with the affricate /nt͡ʃ/ and that there is a variable deletion rule to convert /nt͡ʃ/ to /nʃ/ (the variable production of the oral stop following a nasal consonant and preceding a fricative is common—as a result of which "tents" and "tense" or "mints" and "mince" are homophonous for speakers of most accents of English).

The sonority profile exerts a strong influence on the remaining set of possible two-constituent codae. For example, most C_1C_2 codae consist of a nasal consonant or /l/ followed by another less sonorant consonant. /l/ may therefore precede oral stops ("alp", "bulb), fricatives and affricates ("elf", "turf", "mulch", "bulge") and nasal consonants ("helm", "kiln"). The sonority profile predicts that only fricatives, affricates, and oral stops should follow nasal consonants, but there is in addition a place of articulation constraint that requires C_1 and C_2 to have the same place of articulation if C_1 is a nasal. Such codae can be analysed by assuming that the nasal consonant is an *archiphoneme* that is unspecified for place of articulation, and which adopts the place of articulation of the following consonant. Therefore, "bank" would be /b[Nk/, where /N/ is an archiphoneme, and "sing" is /siNg/, requiring a subsequent deletion (in most English accents) of the final /g/. Consequently, sequences like /mp nd mf/ ("lamp", "band", "nymph") occur in English (being derived from /Np Nd Nf/), but there could never be /mg mk nk np ŋp ŋb/ precisely because the archiphoneme is always translated into a nasal consonant at the same place of articulation as the following consonant. Only a few two-constituent codae have an obstruent in C_1 position. These include /sp st sk/ ("cusp", "paste", "task"), /ft/ ("lift"), and /ps ks/ ("lapse", "axe").

As three-constituent codae are so rare in English, being restricted to perhaps no more than a dozen words, there are good reasons to list these as exceptions to the generalization that syllables can have maximally two consonants in the coda. Moreover, many of these could be reanalysed as two-consonant codae as they consist of a nasal + stop + obstruent cluster in which the stop is ambiguously produced. For example, the production of the codae in "glimpse", "tense", and "jinx" is variable along a continuum from /mps nts ŋks/ to /ms ns ŋs/, respectively. Apart from such clusters, the only other monomorphemic monosyllables (according to the CELEX database, Baayen et al., 1993) with three-consonant codae are "whilst", "next", and "text"; these have been excluded from our grammar.

Rhymes

Rhymes with no consonants in the coda can be formed from any tense vowel, falling diphthong, and /oi/.

All tense and lax vowels can be combined with any legal single-constituent coda. However, the following filters must also be applied:

1. Apart from "with", lax vowels cannot be combined with /ð/.
2. Tense vowels cannot combine with /ŋ/ (so there are no words like /si:ŋ/). The reason for this second constraint is that, at a more abstract level of phonological representation, syllable-final /ŋ/ in words like "song" consists of two consonants /Ng/ (where /N/ is an archiphoneme), and this is filtered by a nucleus–coda constraint (discussed later).
3. /au/ cannot combine with labials or velars.

4. /ai/ cannot combine with the palatal consonants: hence /aiʃ ait͡ʃ aid͡ʒ / are excluded.

Rhymes with /oi/ and /ʊ/ before single-constituent codae have a restricted distribution: /oi/ only occurs before alveolars and /ʊ/ can combine with /t d k ʃ l s t͡ʃ/ (which forms no clear phonological pattern).

The possible rhymes with two constituents in the codae are summarized in Table 1.

The gaps (empty cells) in this table arise both because of the constraints within the coda (discussed earlier) and because of the following further constraint on nucleus–coda combinations: Tense nuclei can only be combined with two-constituent codae in which both consonants are alveolar and in which C_2 is an oral stop (there are five such codae: /nt nd st lt ld/). Hence, there can be neither rhymes like /i:sp/ or /eisk/ (C_2 is non-alveolar) nor /oups/ /i:kt/ /eifs/ (C_1 is non-alveolar and/or C_2 is not an oral stop). There are a few exceptions to these constraints that we note as follows:

1. The combinations that are represented in Table 1 seem to fail to account for monosyllables like "ask" and "clasp" which, for speakers of Australian English and Southern British English, have a tense /ɑ:/ nucleus. Following Coleman (1998), we assume that such vowels are *phonetically* tense/long in some accents of English (such as British English Received Pronunciation) but are *phonologically* lax (i.e., "clasp" is represented as /kl[sp/): In this case, the nucleus–coda constraint on the occurrence of a tense nucleus with a non-alveolar in C_2 coda position is not violated.

TABLE 1
A summary of the possible rhymes with two consonants in the codae as a function of the legal consonants in first (vertical) and second (horizontal) constituent position

First constituent	*Second constituent position*														
position	p	t	k	b	d	g	m	n	f	v	θ	s	z	t͡ʃ	d͡ʒ
p		L										L			
t															
k		L										L			
b															
d															
g															
m	L								L						
n		TL			TL						L	L	L	L	L
ŋ			L												
f		L													
v															
θ															
ð															
s	L	TL	L												
z															
ʃ															
t͡ʃ															
d͡ʒ															
l	L	TL	L	L	TL		L	L	L	L	L	L		L	L

Note: L, T, and TL denote: A Coda only occurs before Lax, Tense, or both Tense and Lax vowels, respectively (e.g., row 1, column 2 stands for /ɪpt ept opt [pt ʌpt/). A gap denotes that the rhyme does not occur.

2. A small number of tense vowels are possible when C_2 is something other than an alveolar stop. These include /ei/ and /au/ combined with /nd͡ʒ/ (“range”, “lounge”); /o:/ combined with /nt͡ʃ/ (“paunch”, “launch”); /ou/ combined with /ks/ (“coax”, “hoax”); and /au/ combined with /ns/ (“flounce”, “bounce”, “ounce”).

3. Combinations of vowels with /lz lg/ do not occur in monomorphemes (except in place names like “Wales” or loan words).

As far as /oi/ and /ʊ/ are concerned, /oi/ can only combine with /st/ and /nd/. With the exception of “wolf”, /ʊ/ does not combine with any two-constituent codae.

As stated earlier, there are a small number of three-constituent codae, and, apart from “whilst”, these only ever combine with lax vowels: They include /mps/ (“glimpse”), /ŋks/ (“lynx”), /mpt/ (“prompt”), and /kst/ (“next”).

Syllables

Syllables are formed by combining the sets of possible onsets and rhymes. We consider here briefly two types of constraint that might apply to these onset–rhyme combinations.

Apart from the restriction that /Cj/ and /CCj/ onsets can only combine with rhymes that have /u:/ nuclei, there are no constraints on onset–nucleus combinations in the same way that there are on the phonemes of the constituents considered so far (Fudge, 1969). There are certainly many gaps in the set of theoretically possible onset–nucleus combinations, of which over 30 are listed in Coleman (1998). The analysis of phoneme sequences in a large lexicon by Harrington, Johnson, and Cooper (1987) showed, for example, that the following onset–nucleus sequences do not occur: /rə:/, /θ ð v/ followed by /u:/ (with the exception of “voodoo”), and /wɑ:/, which is restricted to the single occurrence of “qualm”.

With respect to onset–coda constraints, there is an interesting pattern that involves the repetition of consonants in the onset and coda: For example, /ClVl/ does not occur (where C is any consonant) except in the word “flail”; and many /sCVC/ sequences are excluded when both consonants are identical (e.g., /spVp/, /skVk/) or share many features (e.g., /smVp/, /smVb/, Clements & Keyser, 1983; Davis, 1987; Fudge, 1969; Harrington et al., 1987). As discussed in Coleman (1998), sequences beginning with /sC/ and in which the first constituent of the coda is /s/ (e.g., /stest skɪsp spl[st/) are also excluded. Cairns (1988) interprets these in terms of the obligatory contour principle (McCarthy, 1986), which has been invoked to explain the absence of the repetition of some phonological features (it applies in particular to sequences of identical tones in tone languages), and he notes that this principle may also explain why some polysyllabic words that have consonants repeated across two syllables are excluded or sound odd (e.g., the nonword “spirrup”). In a study of large lexicons, Harrington et al. (1987) and Harrington, Watson, and Cooper (1989) showed that many of these repetitive sequences are excluded not only in monosyllables but also across syllable boundaries in polysyllabic words (see also, Davis, 1987).

Morpheme suffixes

The set of monomorphemic syllables can be augmented by adding three kinds of regular morphemic suffix. First, a regular past-tense morpheme can be added to syllables except those that have a final /t/ or /d/: This morpheme surfaces as /t/ for syllables ending in voiceless consonants (e.g., “frothed”, “packed”, “stamped”) and as /d/ for syllables with a final voiced segment (“feared”, “filled”, “filmed”, “grabbed”). Second, there are various regular morpheme suffixes (possessive, plural, third-person singular) that can all surface as /s/ after voiceless non-sibilant consonants (“paths”, “packs”, “stamps”) and as /z/ after voiced non-sibilant segments (“fills”, “grabs”, “films”, “plays”). Third, a very small number of syllables can be formed from the nominalizing /θ/ suffix. Only a handful of words that all end in a single constituent coda /p t m n ŋ f/ (e.g., “depth”, “width”, “warmth”, “ninth”, “strength”, “fifth”), and only two words (“sixth”, “twelfth”) that have two consonants in the coda are formed in this way. The resulting sequences could be further augmented with a morpheme /s/ suffix following /θ/ to create sequences with three consonants (“depths”, “widths”) or four consonants (“sixths”, “twelfths”) following the nucleus.

Monosyllables can also be augmented with some irregular past-tense morpheme suffixes, which we do not generate in this grammar. These include /n/, which are restricted to a small number of words such as “drawn”, “flown”, and “strewn”; and /t/, which can exceptionally be suffixed to /n/ and /l/ in a very small number of words (e.g., “burnt”, “dealt”, “meant”, “spilled” which can be /spɪlt/ or /spɪld/, and “spoiled” which can be /spoild/ or /spoilt/). These words are certainly generated in our system—but they are not marked as “polymorphemic”.

Generating nonword monosyllables

Having specified the rules and constraints on generating legal monosyllables of English, we describe in this section how these were matched at various stages against the monosyllables of the CELEX database (Baayen et al., 1993), both in order to check that the grammar had generated all the real-word monosyllables that we had expected it to and to create the final list of monosyllabic nonwords.

We combined 53 onsets with 495 rhymes, as well as 17 onsets that had /j/ in second or third constituent position in the onset with 24 rhymes containing an /u:/ nucleus. The total number of generated monomorphemic syllables (henceforth *Gm*) obtained in this way was 27,138. We then matched *Gm* against the phonemic forms of 3,970 monomorphemic monosyllables in the CELEX database (henceforth *Cm*) in order to assess the extent to which all real monosyllables were included in *Gm*. The results of this match showed that there were 201 words in *Cm* that were not in *Gm*. However, of these 201, the majority were forms that we had not intended to generate: They included reduced forms (e.g., /wɪ/ for “we’; the filler “huh”), onomatopoeic words (e.g., “oink”, “psst”), contractions (e.g., “maths”, “turps”), polymorphemic forms (e.g., “ninth”, “shears”), loan-words or names (e.g., “fjord”, “Hertz”), and forms with an alternative pronunciation (e.g., *Cm*: /klɑ:sp/ vs. *Gm* /kl[sp/; *Cm*: /benʃ/ vs. *Gm* /bent͡ʃ/). When we removed these forms that we had not expected to generate, 16 words remained (0.4% of *Cm*), which were erroneously not in *Gm* (see Appendix B). As shown in Appendix B, these are all rare exceptions to the principles for combining phonemes into syllables.

By augmenting our list of monomorphemic forms with regular morphemic suffixes, we generated 47,870 polymorphemic monosyllables. Within this list, 6,511 syllables were also present in the list of monomorphemic syllables (e.g., final /ks/ in monomorphemic “axe” and polymorphemic "backs"). We therefore removed such duplicate forms, leaving 41,359 syllables that we declare to be *necessarily* polymorphemic (hereafter, *Gp*).

The 6,511 syllables that we removed were marked as "either monomorphemic or polymorphemic". To this list we added the 690 forms that, as a result of variable citation-form pronunciations, could be either monomorphemic or polymorphemic. For example, final /ns/ ("mince") is restricted to monomorphemic words, and final /nts/ can be polymorphemic as a result of augmenting /nt/ with an /s/ suffix ("pint"/"pints"). However, as there is a complete phonetic overlap in the realization of /ns/ and /nts/ (as a result of which, "tense" and "tents" are homophonous for almost all speakers), we declared both final /ns/ and final /nts/ as belonging to the category of "either monomorphemic or polymorphemic".

We then determined whether any members of *Gp* occurred in *Cm*. If the generated forms in *Gp* are supposed to be polymorphemic only, then none of these should occur in the set of monomorphemic monosyllables in the CELEX database. A total of 80 *Gp* forms were found to occur in *Cm*. However, most of these could be discounted because they were polymorphemic ("boards"), loan words ("angst", "blitz"), or names ("Hertz"). After removing these, six words remained that erroneously occurred in *Gp*: These are all a subset of those in Appendix B ("beard", "corpse", "mulct", "sculpt", "traipse", "weird"). As a result, there will be some nonwords that have the same rhyme as any of these six words, which will erroneously be marked as "polymorphemic only". For example, the nonword /splɪəd/ will be marked "polymorphemic only", even though /bɪəd/ ("beard") exists as a monomorpheme, and /krɔːps/ will be declared "polymorphemic only" in spite of the existence of /kɔːps/ ("corpse").

As a final check on the extent to which real monosyllabic words were included in the output of our syllable grammar, we matched all generated forms (*Gm* and *Gp* combined) against all monosyllables in the CELEX database (*Cm* and *Cp* combined). The results of this match showed that there were 199 real words in CELEX that we had not generated. Of these, 74 were considered to be archaic words, contractions, exclamations, loan words, onomatopoeic words, and reductions that are beyond the scope of our grammar. A further 96 CELEX words were not included in the output of our grammar because their CELEX phonemic forms listed an alternative citation-form pronunciation such as /benʃ/ for "bench" and /klɑːsp/ for "clasp". Finally, there were 29 words that we did not generate because they were phonotactic oddities similar to those in Appendix B. If we exclude those forms in Appendix B that we have already discussed, as well as their morphological relatives ("cure", "cured", "cures", etc.), we are left with "dreamt", "midst", "sixth", "sixths", "spoilt", "twelfth", "twelfths", which are not output from our grammar. We can explain these omissions as follows: "dreamt" and "spoilt" have variant irregular past tense forms with /t/ as well as with (regular) /d/; "sixth/sixths/twelfth/twelfths" are the only two-constituent codae words to which the nominalizing /θ/ suffix is attached; and "midst" is the only word with final /dst/.

The total number of generated monomorphemic (27,138) and polymorphemic (47,870) syllables minus the duplicated forms (6,511) was 68,497; of these, 7,113 were real-word syllables that appeared in the CELEX database of monosyllabic forms. Although we decided not to filter any of these further by applying the onset–nucleus or onset–codae constraints discussed earlier, we did identify all those nonwords for which no real word was produced by substitution with any other vowel phoneme (e.g., /skok/ was identified because no vowel replacement produces a real word). About 62% (38,186) of the nonwords were marked in this way, which points to the very large number of theoretically possible syllables that do not occur in English because of gaps in onset–nucleus and onset–coda combinations.

PHONOLOGY TO ORTHOGRAPHY CONVERSION

Having derived these sets of phonotactically legal syllables, we next considered how they might be expressed orthographically. The issues surrounding this conversion are numerous, due largely to low level of systematicity in the phonology–orthography mapping (see, e.g., Kreiner, 1992; Ziegler, Stone, & Jacobs, 1997). Our purpose in this work was not, however, to provide a comprehensive analysis of English sound–spelling relationships. Rather, we sought to generate as many spellings as possible for each of the legal syllables, which would be read by the majority of users as the syllables from which they were derived.

Orthographic strings were generated from the legal syllables based upon the following general approach: Phoneme–grapheme correspondences (PGCs) were extracted from the monosyllabic words listed in the CELEX English database (Baayen et al., 1993) and used to derive possible spellings for the legal syllables; these spellings were then converted back to phonologies using the set of grapheme–phoneme correspondence (GPC) rules that forms a part of the Dual-Route Cascaded (DRC) Model of reading (Coltheart et al., 2001; Rastle & Coltheart, 1999b); those spellings that were not translated to their original base phonologies during this procedure were excluded from the database.

We begin with a discussion of how the phonology–orthography conversion was accomplished for monomorphemic syllables, then discuss the way in which this procedure was augmented in order to create orthographies for the polymorphemic syllables. Syllables marked "either monomorphemic or polymorphemic" were treated within both monomorphemic procedures and polymorphemic procedures; these morphologically ambiguous syllables are thus spelled as monomorphemic forms *and* as polymorphemic forms in the database.

Phoneme–grapheme correspondences

In order to convert all legal syllables to orthography, we derived a set of monomorphemic PGC correspondences, the complete set of which is contained in Appendix C.[3] PGCs were developed based upon *all* existing phoneme–grapheme relationships contained within the set of monomorphemic monosyllables in the CELEX English database (Baayen et al., 1993), except those contained in loan words (e.g., "Khmer", "czar"), proper names (e.g., "George"), reduced forms/contractions (e.g., "cant", "huh"), onomatopoeic words (e.g., "psst", "ahem"), archaic forms ("couldst", "didst"), and pseudo-morphological monomorphemic words (e.g., "cracked", "tongs"). For the purposes of PGC development, those monosyllabic orthographies contained in the CELEX database that consisted of an irregular morphological construction (e.g., "caught", "shelve", "flew") were considered to be monomorphemic.

[3] Because these correspondences are based upon British English (as recorded in the CELEX database of English), North American users may find themselves at a disadvantage. They may find correspondences that do not occur in North American English, thus producing some pseudohomophones within the database that are not pseudohomophonic with any North American English word (for example, "taunn" as a pseudohomophone of "torn") and some nonwords that do not reflect the sound–spelling structure of North American English.

Because PGC correspondences comprised *all* sound–spelling relationships in the monomorphemic/monosyllabic lexicon, irrespective of frequency or consistency of correspondence, the initial set of PGCs included a number of relationships that are infrequently used and would be construed as irregular (/ou/→ ew, as in "sew"). Nonwords formed from such correspondences, however, were generally removed when reconverted to phonology (see later).

Any descriptive inventory of phoneme–grapheme relationships, or indeed grapheme–phoneme relationships, must reflect particular commitments regarding the issue of graphemic parsing. Because English sound–spelling relationships are not entirely systematic, defining which letters in a given word represent particular phonemes, although fundamental, is not always straightforward (see, e.g., Berndt, D'Autrechy, & Reggia, 1994, for a discussion of this issue). For example, it is not clear whether the U in "build" should be included as a member of the head grapheme (/b/ → BU) or the vowel grapheme (/ɪ/→ UI); both parsings yield relatively uncommon correspondences, and each preserves the consistency of another phoneme–grapheme pair (i.e., including U in the onset preserves the consistent relationship /ɪ/→ I, and including the U in the vowel preserves the consistent relationship /b/→ B). There is, as yet, no universally accepted solution to this problem (see, e.g., Andrews & Scarratt, 1998; Berndt et al., 1994; Coltheart, Curtis, Atkins, & Haller, 1993; Lange, 2000; Ziegler et al., 1997, for a number of possible segmentation strategies).

Graphemic parsing in our system of PGCs was guided by two general constraints. First, orthographic vowels were never represented as part of the consonantal onset (thus, the correspondence /b/→ BU would never occur in our system), and they were represented only under two circumstances in the consonantal coda. The first of these circumstances concerned silent E, which, if separated from the orthographic vowel nucleus by more than one letter, was parsed as part of the coda. For example, the item "blonde" was parsed B, L, O, N, DE (not B, L, O..E, N, D), and the item "waste" was parsed W, A, S, TE (not W, A..E, S, T), even though the phonological vowel nucleus in this latter instance is tense.[4] Orthographic vowels were also considered part of the coda segment for the correspondences /g/→ GUE (e.g., "morgue") and /k/→ QUE (e.g., "cheque").

The second general constraint that guided parsing in our system was the following: When a parsing ambiguity occurred in the rhyme–body relationship, the letters in question were represented as part of the vowel (e.g., /frait/→ FRIGHT was parsed F, R, IGH, T and not F, R, I, GHT; similarly /tʃo:k/→ CHALK was parsed CH, AL, K and not CH, A, LK). This system of graphemic parsing is broadly consistent with that used in the GPC rules of the DRC model (Coltheart et al., 2001; Rastle & Coltheart, 1999b).

PGCs were position specific in the sense that different PGCs defined sound–spelling relationships for word beginnings, middles, and ends. For example, the correspondence /n/→ KN occurs at the beginnings of words, but this rule is absent in the middle and end positions; similarly, consonant doubling occurs frequently at the ends of monomorphemic words (e.g., "staff") but does not occur in beginning or middle positions. At this stage of the orthographic-translation process, context sensitivity was not allowed. Thus, the correspondence /s/→ C

[4] In these cases, orthographic R was not considered consonantal if immediately preceded by A, E, I, or U (so the parsing of SOURCE was S, OUR.E, C).

was not restricted by the fact that this relation generally holds only when the letter C is followed by E, I, or Y.

Legal monomorphemic syllables were subjected to an orthographic translation procedure in which all possible grapheme combinations (based upon the PGCs discussed earlier) were derived for each syllable. Only one constraint operated upon the grapheme combination procedures: Geminate (i.e., double) consonants could occur only at the end of an orthographic morpheme boundary and immediately following a vowel (thus, neither "paltt" nor "fappe" could occur in our corpus of nonwords). This initial translation process was applied to the legal syllables marked "monomorphemic only" and "either monomorphemic or polymorphemic" in our corpus of legal syllables.

Orthography–phonology re-conversion

As described, PGC correspondences were not restricted by context sensitivity or by the frequency/consistency of sound–spelling mapping. Moreover, the procedures by which graphemes were combined with other graphemes were virtually unrestricted. For these reasons, the first stage of nonword generation resulted in many instances in which a spelling may not have been read as its parent phonological form. For example, the pseudohomophone /weit/ was spelled "watee" as a result of the unconstrained combination of the correspondences /w/→ W, /ei/→ A.E, /t/→ TE. Furthermore, the initial nonword generation procedures resulted in many occurrences of a single spelling mapping onto a number of legal syllables; for example, because of the correspondences /ei/→ E.E in "crepe", and /i:/→ E.E in "these", the letter string "trepe" was generated for both /treip/ and /tri:p/. Our aim was to generate letter strings that would be pronounced by most readers as the syllables from which they were derived—which by definition entails that, whereas a legal syllable can be spelled in a number of ways, a single spelling cannot be used to represent more than one legal syllable. Thus, we re-converted all letter strings back to phonology using the GPC rules of the DRC model (Coltheart et al., 2001; see Rastle & Coltheart, 1999b, for a full listing and description of these rules). Letter strings that did not map back onto their original legal syllables through this conversion procedure were eliminated from the database (following the example earlier, the spelling "trepe" for /treip/ would be eliminated).

The reasons for using the GPC rules of the DRC model as a test of the adequacy of each spelling are numerous. In general, there is much evidence to suggest that readers' pronunciation of novel words relies overwhelmingly on GPC correspondences, even when items contain bodies with very high-frequency irregular neighbours (e.g., "pook" is pronounced with respect to typical GPC correspondences, i.e., /pu:k/ by most readers, despite the existence of "cook", "look", "hook", and so on, see Andrews & Scarratt, 1998). More specifically, we chose to use the particular rules of the DRC model for this purpose because they have been studied extensively both in relation to nonword pronunciations by human readers and in relation to reaction times in speeded nonword reading aloud tasks (see, e.g., Andrews & Scarratt, 1998; Coltheart et al., 2001; Rastle & Coltheart, 1998, 1999a, 2000). These investigations have suggested that the GPC rules of the DRC model provide a good description of the rules that people might use in reading aloud nonwords.

One might wonder why PGC rules were derived separately from the GPC rules of the DRC model, as those GPC rules ultimately determined which nonword letter strings were included in the database. Our reasoning was as follows: Although the GPC rules of the DRC model are much more constrained than our PGC correspondences, they do not constitute a description of the position-specific correspondences *existing* in the lexicon. Rather, they constitute a set of hypotheses about the spelling–sound generalizations that people use in reading nonwords (see Rastle & Coltheart, 1999b, for a discussion of this point). For example, the GPC rules of the DRC model would translate the spelling "ahm" to /a:m/ using the correspondence rules AH →/ɑ:/ and M→/m/. But "ahm" will never be generated from our system of PGC rules because AH does not occur at the beginning of any monomorphemic monosyllable in the CELEX database (with the exception of the onomatopaeic word "ahem"). We used this two-step procedure of phoneme-to-grapheme conversion, then grapheme-to-phoneme re-conversion in order to ensure that the database was comprised of nonwords reflecting *existing* spelling–sound relationships, which would also be read as the syllables from which they were derived.

Generating polymorphemic spellings

The translation procedures described earlier were applied to monomorphemic syllables and to syllables that can legally occur in either monomorphemic or polymorphemic contexts. A similar set of procedures—one with only two general differences from the monomorphemic procedures—was used to generate those syllables that could occur in a polymorphemic context (i.e., those marked "polymorphemic only" or "either monomorphemic and polymorphemic"). One difference was that PGC correspondences for these legal syllables were derived from the *polymorphemic* monosyllables—not the monomorphemic syllables—listed in the CELEX database (Baayen et al., 1993). The other difference was that position specificity in this PGC inventory operated with respect to the morpheme boundary; for example, in the word "banned", the correspondence /n/→ NN was marked as an END rule. No adjustments were made to the GPC re-conversion procedures already discussed because those GPC rules are based on the entire corpus of monosyllabic words—both monomorphemic and polymorphemic.

By treating monomorphemic and polymorphemic syllables separately, we attempted to preserve the regularities that characterize morphological combination in English orthography. For example, the regular past-tense morphemes /t/ and /d/ are always spelled –D or –ED (e.g., so polymorphemic /beikt/ would never be spelled "baict" in our system); and the regular plural morphemes /s/ and /z/ are always spelled –S (so polymorphemic /ti:mz/ would never be spelled "teamz" in our system).

Comparisons to the English lexicon

All orthographies generated from real-word phonologies in the database were checked against the monosyllabic forms contained in the CELEX English database (Baayen et al., 1993) in order to ensure that we had generated all orthographic forms that we had intended to derive. This matching process revealed 1,216 monosyllabic orthographies contained in the CELEX database that did not appear in the set of orthographies generated from real-word phonologies. Of course, we did not want to generate many of these (1,132 words), as they are deemed by the GPC rules of the DRC model to be irregular (e.g., although the orthography "does" would be generated from the phonology /dʌz/ in the first stage of PGC application, it would be excluded when re-translated by the GPC rules of the DRC model). A further 44 of these spellings were not derived because their parent phonological forms were not originally generated (for reasons discussed in the previous section). A remaining 36 spellings were not generated because they are a subset of the words excluded in the initial design of PGC correspondences (loan words, proper names, archaic forms, etc). Four monosyllabic words—"ewe", "ewes", "braille", and "belle"—were erroneously excluded by our orthographic procedures. The words "ewe" and "ewes" were not generated because of difficulty with graphemic parsing of these forms; the words "braille" and "belle" were not generated because they violated our constraint that geminate consonants could occur only at the end of a morpheme.

Subsequent to this matching procedure, all existing monosyllables were removed from the database, leaving 358,534 novel orthographic forms. Of these, 48,534 are spellings that correspond to a real-word syllable (pseudohomophones, e.g., "staik"); 310,000 spellings are not pseudohomophonic with any Australian or Standard Southern British English monosyllable. Of the (non-pseudohomophonic) nonwords, 99,046 were formed from the "monomorphemic only" syllables, 158,376 were formed from the "polymorphemic syllables" and 52,578 were formed from the morphologically ambiguous syllables. Within the database of pseudohomophones, 17,488 were formed from "monomorphemic only" syllables, 14,153 from "polymorphemic only" syllables, and 16,893 from morphologically ambiguous syllables.

NONWORD SELECTION CRITERIA

Users may extract nonwords and pseudohomophones from this database based on a number of selection criteria. Users must specify the search database (nonword or pseudohomophone) and how many items are to be returned. Users are also invited to indicate whether they would like to restrict their search to items with existing orthographic onsets (so "ghrost" would excluded), existing orthographic bodies (so "tuve" would be excluded), and/or legal bigrams (so "slighd" would be excluded), and they are invited to select items from any of the three morphological subsets (monomorphemic, polymorphemic, ambiguous). Table 2 shows the remaining selection criteria available in the database, along with statistics that describe the distribution of these variables in the nonword and pseudohomophone pools. Zero can be entered as a maximum value in any of these fields (in order, for example, to extract nonwords with non-existing bodies). Only the specification of nonword/pseudohomophone and the number to be returned are required fields. If any subsequent field is left blank, the corresponding variable is not considered in the search.

It should be noted that the validity of a number of these variables hinges upon the truth or falsity of our hypotheses regarding the correspondence between orthographic and phonological forms. For example, whether "book", "look", and "cook" count as friends or enemies of "vook" depends upon whether "vook" is pronounced /vu:k/ or /vʊk/. We have discussed why we relied upon the GPC rules of the DRC model to derive the correspondence between orthographic and phonological forms. However, these rules represent only a set of hypotheses about spelling–sound relations, and we acknowledge that the nonwords contained in this database may be pronounced differently from the way in which we have proposed by at least some of the users at least some of the time.

TABLE 2
Selection criteria and descriptive statistics for the nonwords/pseudohomophones included in the database

Variable	*Mean*	*Standard deviation*	*Range*
# Letters	6.64	1.39	2-13
# Neighbours	.56	1.61	0-26
SF neighbours	36.79	741.82	0-100,727
# Body neighbours	1.25	3.22	0-29
SF Body neighbours	95.01	1279.83	0-81,439
# Body friends	1.08	3.01	0-28
SF Body friends	62.89	1037.19	0-81,439
# Body enemies	.17	.83	0-16
SF Body enemies	32.12	681.34	0-30,475
# Onset neighbours	132.11	210.56	0-1428
SF Onset neighbours	8798.21	23,882.50	0-147,271
# Phonological neighbours	5.75	8.09	0-74
SF Phonological neighbours	404.28	2701.04	0-159,955
Bigram frequency (position nonspecific), type measure	1075.39	706.71	0-4339
Bigram frequency (position nonspecific), token measure	931,369	918,608	0-8,171,363
Trigram frequency (position nonspecific), type measure	84.23	73.64	0-574
Trigram frequency (position nonspecific), token measure	57,903	177,191	0-2,821,479
Bigram frequency (position specific), type measure	120.37	170.31	0-842
Bigram frequency (position specific), token measure	25668.20	37595.70	0-1,154,614
Trigram frequency (position specific), type measure	11.68	18.18	0-127
Trigram frequency (position specific), token measure	2577.01	6907.80	0-235,885
# Phonemes	4.63	.99	1-8

Note: # = number, SF = summed frequency.

A FINAL WORD ON THESE NONWORDS

The resulting databases include a number of nonwords and pseudohomophones that might strike the reader as unusual, in various ways. For example, we did not apply any length restrictions, even though the longest monosyllable in English is only nine letters ("strengths"). Thus, some of the nonwords might seem unusually long (e.g., "sckweazzed" for /skwi:zd/). Moreover, as we did not exclude items with non- existing orthographic onsets (e.g., "ghrost") or bodies (e.g., "tuve"), or those items containing illegal bigrams (e.g., "slighd"), some of the nonwords contained in the database might be described as rather "unwordlike".

We retained these items for two reasons. First, these items are important objects of study in their own right. Are readers affected when a nonword, though legal according to our database, contains an orthographic body that no English monosyllable contains? That is a matter that deserves empirical investigation rather than being treated by fiat (i.e., by excluding such items from the database). Perhaps more importantly, however, we wished to exclude, as much as possible, syllables from the database on the basis of *lawful principles*; unlike the phonotactic principles used to exclude phonological rhymes such as /iːmp/ (that is, C_1 is non-alveolar), analogous orthographic principles could not be determined. For example, what orthographic principle would define the onset PHR as legal (as in "phrase") but exclude GHR? And might it be an accident of English orthography that the body –UVE does not occur, whereas the coda V can combine with almost all other vowel–space–E combinations ("rave", "breve", "dive", "love"). What general principle could exclude the body –UVE? The exclusion of such items would indeed appear to be unprincipled.

To the user who finds such nonwords unacceptable, however, two courses of action are available. Firstly, users could simply reject items that they find undesirable. In our view, however, that would not be methodologically acceptable if the rejection were done on intuitive grounds, and even less acceptable if it were done without acknowledgement. These rejections should be done explicitly and objectively, not covertly and intuitively (see, e.g., Forster, 2000, who showed that experimenter intuition about item difficulty is strongly related to participant performance in lexical decision tasks). Thus, if the reason for rejecting "pawde" is that the orthographic body –AWDE occurs in no English monosyllable, then this criterion should be applied uniformly (and so "tuve" would have to be rejected also).

Of course, achieving uniformity in nonword selection and excluding "unwordlike" nonwords is easily done by using criteria that minimize the frequency with which orthographically unusual items occur. Items with non-existing onsets and/or bodies, as well as items with illegal bigrams can be excluded from the search automatically; alternatively, users can specify minimum values for properties such as neighbourhood size, bigram frequency, and trigram frequency to remove undesirable items. Naturally, if this course of action is adopted, its adoption should also be acknowledged, and it should be applied consistently.

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*, 1052–1086.

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

Baluch, B., & Besner, D. (1991). Visual word recognition: Evidence for strategic control of lexical and nonlexical routines in oral reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 644–652.

Beckman, M.E., Edwards J., & Fletcher J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G.J. Docherty & D.R. Ladd (Eds.), *Papers in laboratory phonology II: Gesture, segment, prosody* (pp. 68–86). Cambridge: Cambridge University Press.

Berndt, R.S., D'Autrechy, C.L., & Reggia, J.A. (1994). Functional pronunciation units in English words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 977–991.

Blevins, J. (1995). The syllable in phonological theory. In J. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 206–244). Cambridge, MA: Blackwell.

Cairns, C. (1988). Phonotactics, markedness and lexical representation. *Phonology*, *5*, 209–236.

Cairns, C., & Feinstein, M. (1982). Markedness and the theory of syllable structure. *Linguistic Inquiry*, *13*, 193–225.

Clements, G.N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M.E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 283–333). Cambridge: Cambridge University Press.

Clements, N., & Keyser S. (1983). *CV phonology: A generative theory of the syllable*. Cambridge, MA: MIT Press.

Coleman, J. (1998). *Phonological representations: Their names, forms, and powers*. Cambridge: Cambridge University Press.

Coleman, J., & Pierrehumbert, J. (1997). *Stochastic phonological grammars and acceptability*. In *3rd meeting of the ACL Special Interest Group in computational phonology: Proceedings of the workshop* (pp. 49–56). Somerset, NJ: Association for Computational Linguistics.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505.

Coltheart, M. (Ed.) (1996). *Phonological dyslexia*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589–608.

Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Coltheart, M., & Rastle, K. (1994). Serial processing and reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1197–1211.

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.

Davis, S. (1987). Some observations concerning English stress and phonotactics using a computerized lexicon. In *Research on speech perception progress report*, *13* (pp. 225–237). Indiana University.

de Jong, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, *97*, 491–504.

Deterding, D. (1997). The formants of monophthong vowels in Standard Southern British English pronunciation. *Journal of the International Phonetic Association*, *27*, 47– 55.

Forster, K.I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory and Cognition*, *28*, 1109–1115.

Forster, K.I., & Veres, C. (1998). The prime lexicality effect: Form-priming as a function of prime awareness, lexical status, and discrimination difficulty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 498–514.

Fudge, E. (1969). Syllables. *Journal of Linguistics*, *5*, 253–287.

Halle, M., & Vergnaud, J. (1980). Three-dimensional phonology. *Journal of Linguistic Research*, *1*, 83–105.

Harrington J., Cox, F., & Evans Z. (1997). An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Australian Journal of Linguistics*, *17*, 155–184.

Harrington, J., Johnson, I., & Cooper, M. (1987). The application of phoneme sequence constraints to word boundary identification in automatic continuous speech recognition. In J. Laver & M. Jack (Eds.), *European Conference on Speech Technology* (pp. 163–166). Edinburgh, UK: CEP Consultants.

Harrington, J., Palethorpe, S., & Watson, C. (2000). Does the Queen speak the Queen's English? *Nature*, *408*, 927–928.

Harrington, J., Watson, G., & Cooper, M. (1989). Word boundary detection in broad class and phoneme strings. *Computer Speech and Language*, *3*, 367–382.

Hay, J., Pierrehumbert, J., & Beckman, M. (in press). Speech perception, well- formedness, and the statistics of the lexicon, (Postscript). *Papers in laboratory phonology VI*. Cambridge: Cambridge University Press.

Hooper, J.B. (1972). The syllable in phonological theory. *Language*, *48*, 525–540.

Itô, J. (1986). *Syllable theory in prosodic phonology*. PhD dissertation, University of Massachusetts, Amherst, MA, USA.

Jespersen, O. (1904). *Phonetische Grundfragen*. Leipzig, Germany: B.G. Teubner.

Kaye, J. (1990). "Coda" licensing. *Phonology*, *7*, 301–330.

Kenstowicz, M. (1994). *Phonology in generative grammar*. Cambridge, MA: Blackwell.

Kiparsky, P. (1981). Remarks on the metrical structure of the syllable. In W. Dressler, O. Pfeiffer, & J. Rennison (Eds.), *Innsbrucker Beiträge zur Sprachwissenschaft* (pp. 245–256). Innsbruck: Ernst Becvar.

Kreiner, D.S. (1992). Reaction time measures of spelling: Testing a two-strategy model of skilled spelling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 765–776.

Lange, M. (2000). *De l'orthographe à la prononciation: Nature des processus de conversion graphème-phonème dans la reconnaissance des mots écrits*. Unpublished PhD thesis, Université Libre de Bruxelles, Bruxelles, Belgium.

Lindblom, B., & Sundberg, J. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, *50*, 1166–1179.

Lowenstamm, J. (1981). On the maximal cluster approach to syllable structure. *Linguistic Inquiry*, *12*, 575–604.

Lupker, S.J., Brown, P., & Colombo, L. (1997). Strategic control in a naming task: Changing routes or changing deadlines? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 570–590.

McCann, R.S., & Besner, D. (1987). Reading pseudohomophones: Implications for models of pronunciation assembly and the locus of word-frequency effects in naming. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 14–24.

McCarthy, J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, *17*, 207–263.

McCarthy, J., & Prince, A. (1990). Prosodic morphology and templatic morphology. In M. Eid & J. McCarthy (Eds.), *Perspectives on arabic linguistics II* (pp. 1–54). Amsterdam: Benjamins.

Mohanan, K.P. (1985). Syllable structure and lexical strata in English. *Phonology Yearbook*, *2*, 139–155.

Monsell, S., Patterson, K.E., Graham, A., Hughes, C.H., & Milroy, R. (1992). Lexical and sub-lexical translation of spelling to sound: Strategic anticipation of lexical status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 452– 467.

Peereman, R., & Content, A. (1995). Neighborhood size effect in naming: Lexical activation or sublexical correspondences? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 409–421.

Pierrehumbert, J. (1994). Syllable structure and word structure: A study of triconsonantal clusters in English. In P. Keating (Ed.), *Phonological structure and phonetic form. Papers in laboratory phonology III* (pp. 90–117). Cambridge: Cambridge University Press.

Pike, K., & Pike, E. (1947). Immediate constituents of Mazateco syllables. *International Journal of American Linguistics*, *13*, 78–91.

Pugh, K.R., Rexer, K., Peter, M., & Katz, L. (1994). Neighborhood effects in visual word recognition: Effects of letter delay and nonword context difficulty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 639–648.

Rastle, K., & Coltheart, M. (1998). Whammies and double whammies: The effect of length on nonword reading. *Psychonomic Bulletin and Review*, *5*, 277–282.

Rastle, K., & Coltheart, M. (1999a). Lexical and nonlexical phonological priming. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 461–481.

Rastle, K., & Coltheart, M. (1999b). 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.

Rice, K. (1992). On deriving sonority: A structural account of sonority relationships. *Phonology*, *9*, 61–99.

Rubach, J., & Booij, G. (1990). Syllable structure assignment in Polish. *Phonology*, *7*, 121–158.

Saussure, F. de (1916). *Cours de Linguistique Générale*. Lausanne: Payot.

Selkirk, E. (1984). *Phonology and syntax*: *The relation between sound and structure*. Cambridge, MA: MIT Press.

Sievers, E. (1881). *Grundzüge der Phonetik*. Leipzig: Breitkopf and Hartel.

Steriade, D. (1982). *Greek prosodies and the nature of syllabification*. Unpublished PhD dissertation, MIT, USA.

Tabossi, P., & Laghi, L. (1992). Semantic priming in the pronunciation of words in two writing systems: Italian and English. *Memory and Cognition*, *20*, 303–313.

Taft, M., & Russell, B. (1992). Pseudohomophone naming and the word frequency effect. *Quarterly Journal of Experimental Psychology*, *45A*, 51–71.

Trubetzkoy, N. (1958). *Grundzüge der Phonologie. Travaux du Cercle Linguistique de Prague.* (Original work published 1939.) Göttingen: Vandenhoeck and Ruprecht.

Weekes, B.S. (1997). Differential effects of number of letters on word and nonword naming latency. *Quarterly Journal of Experimental Psychology*, *50A*, 439–456.

Wells, J.C. (1982). *Accents of English.* Cambridge: Cambridge University Press.

Zec, D. (1995). Sonority constraints on syllable structure. *Phonology*, *12*, 85–129.

Ziegler, J.C., Stone, G.O., & Jacobs, A.M. (1997). What is the pronunciation for -ough and the spelling for /u/? A database for computing feedforward and feedback consistency in English. *Behavior Research Methods, Instruments, and Computers*, *29*, 600–618.

Zwicky, A. (1972). Note on a phonological hierarchy in English. In R. Stockwell & R. Macaulay (Eds.), *Linguistic change in generative theory* (pp. 275–301). Bloomington: Indiana University Press.

Original manuscript received 15 May 2001
Accepted revision received 17 September 2001

APPENDIX A

The system for transcribing lax vowels, tense vowels, falling diphthongs, and consonants used in this paper is shown as follows. This transcription system is applicable to Australian English (AE) and other non-rhotic accents (such as Received Pronunciation) that are considered to be systemically equivalent to AE. (/ʊ/ and /oi/, although phonologically lax and tense, respectively, as shown here, are not grouped with these categories in generating syllables.)

Lax	
ɪ	hid
ʊ	hood
e	head
o	hod
ʌ	thud
[	had

Tense	
i:	heed
u:	who'd
ei	hay
oi	boy
au	how
ai	high
ou	hoe
ə:	heard
o:	hoard
ɑ:	hard

Falling diphthongs	
ɪə	here
ʊə	tour
eə	hair

Consonants	
b	bat
s	sat
d	did
f	fat
ɡ	gas
h	hat
k	cat
l	lamp
m	mat
n	not
p	pat
r	rat
t	tap
v	vat
w	won
j	young
z	zip
ŋ	sing
θ	thank
ð	that
ʃ	shut
d͡ʒ	jest
t͡ʃ	chat

APPENDIX B

The following monomorphemic monosyllables (included in the CELEX database, Baayen et al., 1993) that are not included in the generated phonological forms are all phonotactic exceptions for the following reasons: (1) Falling diphthongs are restricted to open syllables; (2) tense vowels require C_2 to be an alveolar stop; (3) after /Cj/ onsets, the nucleus must be /u:/; (4) these are three-constituent coda that occur only in these words and that form no pattern with three-constituent codae in any other word (additionally, "whilst" is odd because only lax vowels combine with three-constituent codae); (5) there are no other onsets in which two constituents are fricatives; (6) only tense vowels precede /ð/; (7) this is the only word that has a two-constituent coda after /ʊ/.

1. beard (bɪəd), fierce (fɪəs), pierce (pɪəs), weird (wɪəd), scarce (skeəs)
2. corpse (ko:ps), traipse (treips)
3. pure (pjʊə), cure (kjʊə)
4. whilst (wailst), sculpt (skʌlpt), mulct (mʌlkt)
5. sphinx (sfɪŋks), sphere (sfɪə)
6. with (wɪð)
7. wolf (wʊlf)

APPENDIX C

The following sound–spelling correspondences were used in the phonology–orthography translation of legal syllables. Graphemes were mapped onto phonemes in a position-specific manner, according to whether the relevant phoneme occurred at the beginning, end, or middle of a nonword syllable. Example words are provided in parentheses. Phon. = phoneme.

Phon.	*Beginning*	*Middle*	*End*
ɪ	i (it)	i (hit)	
		y (sync)	
		i.e (give)	
		ui (build)	
ʊ		oo (book)	
		u (put)	
		oul (would)	
		o (wolf)	
e	e (egg)	e (bet)	
		ea (head)	
		ue (guess)	
		ai (said)	
		ie (friend)	
o	o (odd)	o (hot)	
		a (watt)	
		ou (trough)	
ʌ	u (up)	oo (blood)	
		o.e (done)	
		u (hut)	
		o (front)	
		oe (does)	
		ou (touch)	
[	a (at)	a (hat)	
		a.e (have)	
		ai (plaid)	
i:	ea (east)	ea (lead)	ee (scree)
	ee (eel)	ea.e (grease)	ea (sea)
	e.e (eve)	ee (reed)	ey (key)
		e.e (these)	i (ski)
		ee.e (peeve)	ay (quay)
		ie (priest)	e (we)
u:	oo.e (ooze)	u.e (prude)	ew (dew)
	oo (oops)	eu (sleuth)	ue (true)
		oo (room)	u (flu)
		ui (suit)	oo (too)
		ou (group)	oe (shoe)
		ew (shrewd)	o (to)
		oo.e (booze)	wo (two)
		ui.e (juice)	
		ou.e (route)	
		o (tomb)	
ei	ai (aid)	ai (raid)	ay (stay)
	a.e (ate)	a.e (sale)	ey (they)
	eigh (eight)	ei (feign)	ae (brae)
		ai.e (praise)	eigh (sleigh)
		eigh (weight)	
		ea (break)	
		e.e (crepe)	
		aigh (straight)	
		a (waste)	
oi	oi (oil)	oi (groin)	oy (boy)
		oi.e (voice)	
au	ou (ouch)	ou (couch)	ow (how)
	ow (owl)	ow (fowl)	ough (bough)
		ow.e (browse)	ou (thou)
		ou.e (douse)	
ai	i.e (ice)	i.e (kite)	y (sky)
	eye (eye)	igh (tight)	ye (rye)
	ais (aisle)	ei (stein)	ie (pie)
	is (isle)	y (style)	igh (high)
		eigh (height)	uy (buy)
		ie (pied)	i (hi)
		i (pint)	
		ui.e (guide)	
		y.e (type)	
ou	oa (oat)	o.e (hole)	oe (hoe)
	o.e (ode)	oa (road)	ough (though)
	ow (own)	o (told)	ow (know)
	oh (ohm)	ow (bowl)	ew (sew)
	owe (owe)	ol (folk)	o (fro)
		ou (soul)	
		au.e (mauve)	
ə:	ear (earn)	ur (burn)	ir (sir)
	ir (irk)	er (fern)	ur (blur)
	ur (urn)	ir (chirp)	er (her)
	er (erg)	ear (heard)	urr (burr)
	ur.e (urge)	or (work)	irr (whirr)
	err (err)	ur.e (purge)	
		er.e (merge)	
o:	augh (aught)	or (lord)	oor (poor)
	aw (awl)	au (gaunt)	our (pour)

Phon.	*Beginning*	*Middle*	*End*
	awe (awe)	aw (scrawl)	ure (lure)
	oar (oar)	augh (caught)	aw (saw)
	ore (ore)	ough (sought)	oar (roar)
	ough (ought)	our (court)	or (for)
	or (orb)	ar (ward)	ar (war)
	au (auk)	al (talk)	ore (bore)
		au.e (cause)	augh (faugh)
		oa (broad)	
		or.e (force)	
		oar (board)	
		our.e (source)	
		oar.e (hoarse)	
		a (mall)	
ɑ:	ar (art)	ar (dart)	a (bra)
	a (aft)	a (cask)	ar (far)
	au (aunt)	al (balm)	arr (parr)
	are (are)	ar.e (carve)	
		a.e (vase)	
		au (laugh)	
		ear (heart)	
		uar (guard)	
		er (clerk)	
ɪə	ear (ear)	ier (fierce)	ere (here)
		ear (beard)	ear (near)
			eer (peer)
			ier (pier)
ʊə			ure (cure)
			our (tour)
			ewer (sewer)
eə	air (air)	air (cairn)	eir (their)
	heir (heir)	ar.e (scarce)	ere (there)
			air (chair)
			are (dare)
			ear (pear)
b	b (beg)		b (rob)
			bb (ebb)
s	s (sail)	s (last)	ss (lass)
	ps (psalm)		ce (sconce)
	c (cell)		se (sense)
	sw (sword)		z (waltz)
	sc (scene)		s (yes)
			c (brace)
d	d (dog)		d (pad)
			dd (odd)
			de (blonde)

Phon.	*Beginning*	*Middle*	*End*
f	f (fair)	f (aft)	f (spoof)
	ph (phone)	ph (sphere)	ph (graph)
			ff (gruff)
			gh (rough)
ɡ	g (god)		g (fog)
	gh (ghost)		gue (morgue)
			gg (egg)
h	h (hot)		
	wh (whole)		
k	c (cat)	c (scud)	k (bake)
	ch (chasm)	q (squirt)	ck (back)
	k (kill)	k (skill)	que (pique)
	q (quick)	ch (scheme)	c (sync)
			che (ache)
l	l (lamb)	l (slime)	l (bale)
			ll (roll)
m	m (mince)	m (smile)	m (ham)
			mn (hymn)
			gm (phlegm)
			mb (dumb)
n	n (nab)	n (snout)	n (ran)
	gn (gnu)		nn (inn)
	kn (know)		gn (sign)
p	p (pick)	p (spring)	p (nap)
			ppe (steppe)
r	r (rot)	r (string)	
	wr (wring)		
	rh (rhyme)		
t	t (tap)	t (start)	t (rat)
			tt (putt)
			te (waste)
v	v (vent)		v (clove)
			ve (delve)
w	w (will)	u (quilt)	
	wh (which)	w (sway)	
ks		x (axe)	x (ox)
j	y (young)		
z	z (zero)		z (blaze)
			zz (buzz)

Phon.	*Beginning*	*Middle*	*End*
			ze (adze)
			s (close)
ŋ		n (sink)	ng (sing)
θ	th (thick)		th (north)
ð	th (that)		the (loathe)
			th (booth)
ʃ	sh (shut)		sh (hush)
	ch (chef)		che (douche)
	sch (schwa)		ch (tench)
	s (sure)		

Phon.	*Beginning*	*Middle*	*End*
d͡ʒ	g (gell)		ge (bilge)
	j (jam)		dge (badge)
			g (barge)
t͡ʃ	ch (church)		tch (blotch)
			ch (punch)
			che (niche)
juː		u.e (cure)	ew (few)
		ew (hewn)	iew (view)
		u (fugue)	ieu (lieu)
		eu (deuce)	ue (cue)