Behavior Research Methods, Instruments, & Computers
2002, 34 (3), 416-423

Similarity relations among spoken words: The special status of rimes in English

BRUNO DE CARA and USHA GOSWAMI
University College London, London, England

This paper presents an analysis of the distribution of phonological similarity relations among monosyllabic spoken words in English. It differs from classical analyses of phonological neighborhood density (e.g., Luce & Pisoni, 1998) by assuming that not all phonological neighbors are equal. Rather, it is assumed that the phonological lexicon has psycholinguistic structure. Accordingly, in addition to considering the *number* of phonological neighbors for any given word, it becomes important to consider the *nature* of these neighbors. If one type of neighbor is more dominant, neighborhood density effects may reflect levels of segmental representation other than the phoneme, particularly prior to literacy. Statistical analyses of the nature of phonological neighborhoods in terms of *rime* neighbors (e.g., *hat/cat*), *consonant* neighbors (e.g., *hat/hit*), and *lead* neighbors (e.g., *hat/ham*) were thus performed for all monosyllabic words in the Celex corpus (4,086 words). Our results show that most phonological neighbors are rime neighbors (e.g., *hat/cat*) in English. Similar patterns were found when a corpus of words for which age-of-acquisition ratings were available was analyzed. The resultant database can be used as a tool for controlling and selecting stimuli when the role of lexical neighborhoods in phonological development and speech processing is examined.

Recent theories of phonological development suggest that, as language is acquired, the growing number of similar-sounding words in the mental lexicon (*phonological neighborhood density*; henceforth, *N*) creates a pressure to represent words in a phonologically well specified manner to support efficient discrimination (Metsala & Walley, 1998). The proposal is that phonological awareness (the ability to manipulate components of spoken words in tasks including word segmentation and sound categorization) may emerge as the result of spoken vocabulary growth and associated changes in interitem phonological similarity relations (lexical restructuring theory, or LRT; see Metsala, 1999; Metsala & Walley, 1998). When vocabulary size is small, phonological similarity between words is thought unlikely to interfere with efficient access, and so it is assumed that there is no need to represent words in a phonologically detailed manner. Early word representations are thus claimed to be holistic (i.e., to represent global phonological characteristics; e.g., Ferguson, 1986; Jusczyk, 1986, 1993; Walley & Flege, 1999). As vocabulary grows, children need to distinguish between more and more words that sound similar to each other, and this eventually creates a developmental pressure to represent smaller segments of speech, such as syllables and, ultimately, phonemes (Fowler, 1991; Metsala & Walley, 1998; Walley, 1993). By adulthood, it is assumed that all words are represented as linear sequences of phonemes (e.g., *prince* is represented as /p/ /r/ /i/ /n/ /s/). Neighborhood density effects in speech-processing tasks in adults (typically, words from sparser neighborhoods are recognized more quickly) are usually taken as evidence for such phoneme-based representations.

The developmental lexical restructuring process postulated by Metsala and Walley (1998) is thought to be relatively word specific, depending on such factors as overall vocabulary size and the number of similar-sounding words in the lexicon. For example, words with many similar-sounding neighbors (words with *dense N*) are thought to experience more pressure for phoneme-level restructuring than do words with few similar-sounding neighbors (words with *sparse N*). Hence, early in development, words with dense *N* should be processed more accurately in speech-based tasks. Consistent with this prediction, Logan (1992) found that 2-year-olds were better at identifying (by pointing to pictures) familiar words from dense neighborhoods than those fro m sparse neighborhoods. This density effect had disappeared by age 4. In addition, Metsala (1999) found that 3- and 4-year-old children performed significantly better in a simple phoneme-blending task when the target words were from dense neighborhoods, rather than from sparse neighborhoods.

The broad picture of phonological development characterized by LRT is probably correct. However, there are two problems with LRT as a developmental hypothesis. First, the theory goes from syllable to phoneme without postulating a strong developmental role for intrasyllabic units like

Support for this research was partly provided by Fyssen Foundation and Marie Curie Foundation awards to B.D.C. Preparation of this paper was supported by ESRC small grant (RN000223153) to U.G. We thank Ronald Peereman for helpful discussions concerning our approach. Correspondence concerning this article should be addressed to U. Goswami, Behavioural and Brain Sciences Unit, Institute of Child Health, University College London, 30 Guilford St., London WC1N 1EH, England (e-mail: u.goswami@ich.ucl.ac.uk).

onset/rime (e.g., /k/–/æt/ for *cat*). This is surprising given the importance of onset-rime units in phonological development prior to literacy (Goswami & Bryant, 1990; Treiman, 1988). For example, preliterate children usually perform relatively well in *onset-rime* tasks (e.g., segmenting *cat* into/k/–/æt/) while doing rather poorly in *phonemic* awareness tasks (e.g., segmenting *cat* into/k/–/æ/–/t/; see Goswami & Bryant, 1990). The representation of phonemes in words is largely dependent on literacy acquisition, not on vocabulary acquisition (e.g., illiterate adults have poor phoneme awareness; Morais, Cary, Alegria, & Bertelson, 1979). Some authors have also reported a significant relationship between vocabulary development and rime-level phonological skills in young children (e.g., Maclean, Bryant, & Bradley, 1987), but not between vocabulary development and phoneme-level skills (e.g., Hulme et al., 2002).

Second, LRT does not allocate any special role to the subtypes of the neighbors that a word has. Metsala and Walley (1998) considered the overall *number* of phonological neighbors as a causal factor in LRT, but not the *nature* of these neighbors. This is important, because the lexicon of spoken word forms may have psycholinguistic structure at levels other than the phoneme. For example, given the psychological salience of the rime to young children, it seems possible that many phonological neighbors in English are rime neighbors. Traditional studies of speech processing describe some of these similarity relations in terms of phonotactic probabilities (i.e., possible combinations between phonemes; e.g., Bailey & Hahn, 2001; Vitevitch & Luce, 1999). The focus of these studies in the adult auditory-processing literature is always on the phoneme. We propose that such a focus is appropriate only for literate participants. It may not be appropriate for preliterate participants. For preliterate participants (usually young children), neighborhood relations may operate at linguistic levels other than the phoneme.

Furthermore, it is possible that the developmental salience of onsets and rimes demonstrated in behavioral work with children (Goswami & Bryant, 1990) may leave a "footprint" in the adult lexicon, leaving a preference for onset-rime organization in auditory-processing tasks even when phonemes are fully represented (e.g., Treiman, 1988). In order to investigate these hypotheses empirically, a description of phonological neighborhoods in terms of type of phonological neighbor is required. Since we could find no analyses of this nature in English, we analyzed the corpus of single-syllable words in the Celex database[1] (Baayen, Piepenbrock, & Gulikers, 1995) in terms of *rime* neighbors (RNs, sharing the rime as in *hat*/*cat*), *consonant* neighbors (CNs, sharing the consonant phonemes, as in *hat*/*hit*), and *lead* neighbors (LNs, sharing the onset–vowel sequence or lead, as in *hat*/*ham*; see Peereman & Content, 1997). Standard pronunciation in southern British English was used as a basis for the neighborhood calculations (hence, some rime neighbors will not rhyme in American English). The resultant database is the focus of this paper.[2]

STATISTICAL ANALYSIS

Presentation of the Database

The initial database corresponded to all monosyllabic words (lemmas only) found in Celex (7,256 words). However, we used some restrictions to reduce this database. First, we excluded 2,680 words that were both *homophones and homographs* of other words in the database. For example, *browse* as a verb and *browse* as a noun were reduced to only one entry. The lexical frequency assigned to this entry corresponded to the cumulated frequency of words with the same phonology and the same spelling. Second, we excluded 258 words that were *homographic nonhomophones* of other words in the database. For example, *lunch* pronounced with the fricative /ʃ/ or with the affricate /tʃ/ at the end were reduced to only one entry. The standard pronunciation in southern British English (found in Jones, Roach, & Hartman's, 1997, phonetic dictionary) was selected. Nevertheless, homographic nonhomophone words that differed on syntactic class (e.g., *live* as verb and *live* as an adjective), grammatical tense (*lead* as present and *lead* as past tense), or meaning (e.g., *bow* as bending and *bow* for shooting) were kept distinct. Third, we excluded 128 words that either corresponded to single contractions (e.g., *'s*), complex contractions (e.g., *how's*), or abbreviations (e.g., *sq*). Fourth, homophonic nonhomograph words (e.g., *pear*, *pair*) were kept distinct, except when they referred to the same morpheme (e.g., *disk* and *disc*; 104 words removed). This left 4,086 monosyllables (see the Appendix). The distribution of syllable types in the 4,086 database is shown in Table 1.

Metrics for Calculating Phonological Neighborhood Density

We used two definitions of phonological neighborhood. The first definition, based on models of speech recogni-

Table 1
Syllable Types With Their Occurrence in the 4,086 Database

Type			Subtype		
Syllable	Occurrence	%	Syllable	Occurrence	%
[C]V[C]	3,627	88.8			
			CVC	1,758	43.0
			CCVC	858	21.0
			CVCC	622	15.2
			CCVCC	233	5.7
			CCCVC	78	1.9
			CVCCC	44	1.1
			CCCVCC	19	0.5
			CCVCCC	15	0.4
[C]V	299	7.3			
			CV	184	4.5
			CCV	102	2.5
			CCCV	13	0.3
V[C]	146	3.6			
			VC	94	2.3
			VCC	50	1.2
			VCCC	2	0.0
V	14	0.3			

Note—C, consonant; V, vowel.

Table 2
Phonological Neighborhood for the Target Word *Hat*

OVC Metric ($N = 45$)			Ph±1 Metric ($N = 32$)		
RN = 24	*CN* = 10	*LN* = 11	*RN* = 14	*CN* = 10	*LN* = 8
vat	hut	haves	vat	hut	have
that	hurt	have	that	hurt	hatch
tat	hot	hatch	tat	hot	hash
sprat	hoot	hash	rat	hoot	hap
splat	hit	hap	pat	hit	hang
spat	height	hank	matt	height	ham
slat	heat	hang	mat	heat	hag
scat	heart	hand	gnat	heart	hack
rat	hate	ham	gat	hate	
prat	hart	hag	fat	hart	
plait		hack	chat		
pat			cat		
matt			bat		
mat			at		
gnat					
gat					
flat					
fat					
drat					
chat					
cat					
brat					
bat					
at					

Note—OVC represents all phonological neighbors that differ from a target word by one *onset*, *vowel*, or *coda* substitution, deletion, or addition. Ph±1 represents all phonological neighbors that differ from a target word by one *phoneme* substitution, deletion, or addition. *N* = number of overall neighbors, among which *RN* = number of rime neighbors (e.g., *hat*/*cat*), *CN* = number of consonant neighbors (e.g., *hat*/*hit*), and *LN* = number of lead neighbors (e.g., *hat*/*ham*).

tion, considers phonological neighborhood as a set of words that differ from a given target by one *phoneme* substitution, addition, or deletion (metric Ph±1; Charles-Luce & Luce, 1990; Landauer & Streeter, 1973; Luce, 1986; Luce & Pisoni, 1998). For example, according to the Ph±1 metric, the similarity neighborhood for the word *hat* would include *bat*, *hot*, *ham*, and *at*, among others. The second definition of phonological neighborhood is based on the linguistic coding of syllables in three dimensions: *onset* (initial consonant or consonant cluster), *vowel*, and *coda* (final consonant or consonant cluster; e.g., /pr/–/i/–/ns/ for *prince*). Words that differ only by the substitution, addition, or deletion of one component in this three-dimensional (3-D) coding of syllables can be viewed as phonological neighbors (onset–vowel–coda, or OVC, metric). According to this alternative definition of phonological neighborhood, words differing in onsets by more than one phoneme but sharing rimes (e.g., *hat*, *flat*, *splat*, *drat*) are phonological neighbors. Similarly, words differing in codas by more than one phoneme but sharing leads (e.g., *hat*, *hand*) are phonological neighbors. This metric was adopted on the basis of Treiman's work (e.g., Treiman, 1988) and also because word games for children and nursery rhymes do not restrict their rhyming patterns to Ph±1 neighbors (e.g., "Hickory Dickory Dock" rhymes *dock* with *clock*; "Twinkle Twinkle Little Star" rhymes *star* with *are*). The chief difference between these similarity metrics is that words like *flat* and *hand* would count as phonological neighbors of *hat* in the OVC metric, but not in the Ph±1 metric. This is because *flat* and *hand* differ from *hat* by more than one phoneme substitution, addition, or deletion (Table 2).

Neighbors were calculated by computerized routines.[3] For each similarity metric, we defined three subtypes of neighbor: RNs (e.g., *hat*/*fat*), CNs (e.g., *hat*/*hit*), and LNs (e.g., *hat*/*ham*). The phonological neighborhood related to each target word was divided into RNs, CNs, and LNs in both metrics. For example, according to the OVC metric, the target word *hat* has 45 neighbors, 24 of which are RNs (53%), 10 of which are CNs (22%), and 11 of which are LNs (25%). According to the Ph±1 metric, *hat* has 32 neighbors, 14 of which are RNs (44%), 10 of which are CNs (31%), and 8 of which are LNs (25%). The number of RNs and the total number of neighbors was highly correlated for both metrics (OVC, $r = .89$; Ph±1, $r = .92$; both $ps < .0001$).

Number of Phonological Neighbors

We then calculated *N* for each word in the 4,086 database by type and by token. The calculation by type was based on the absolute number of neighbors (it gave an equal weight to each neighbor). The calculation by token was based on the cumulated frequencies of neighbors (it gave a relative weight to each neighbor as a function of its lexical frequency). Lexical frequency corresponded to the Celex measure for spoken frequency of lemmas (Cob-

Table 3
Number of Phonological Neighbors (*N*) for English Monosyllabic Words (4,086 database)

		OVC Metric				Ph±1 Metric			
Calculation	Measure	*RN*	*CN*	*LN*	*N*	*RN*	*CN*	*LN*	*N*
By type	*M*	16.0	5.0	8.5	29.6	7.9	5.0	5.0	17.9
	SD	13.1	5.1	5.8	17.0	8.2	5.1	4.1	14.6
	%	54.2	17.0	28.9	100.0	44.1	28.0	27.8	100.0
By token	*M*	4,482	1,940	1,583	8,006	2,824	1,940	1,051	5,814
	SD	12,476	9,198	5,515	18,526	9,856	9,198	4,297	16,872
	%	56.0	24.2	19.8	100.0	48.6	33.4	18.1	100.0

Note—*RN*, number of rime neighbors (e.g., *hat*/*cat*); *CN*, number of consonant neighbors (e.g., *hat*/*hit*); *LN*, number of lead neighbors (e.g., *hat*/*ham*). The calculation by type is based on the absolute number of neighbors. The calculation by token is based on the cumulated frequencies of neighbors.

SMln; occurrence per million within a 17.9 million spoken word corpus). The mean number of phonological neighbors for the 4,086 database is shown in Table 3.

As can be seen from Table 3, the mean number of phonological neighbors is higher for the OVC metric (29.6) than for the Ph±1 metric (17.9). This is not surprising, since the Ph±1 metric is a subset of the OVC metric. The percentage of RNs is 54.2% for the OVC metric and 44.1% for the Ph±1 metric. It should be noted that similar patterns were found with the analyses by token: The percentage of RNs is 56.0% for the OVC metric and 48.6% for the Ph±1 metric. A token-based analysis reduces the percentage of LNs, however. These analyses demonstrate that the majority of similar-sounding words in English are RNs. Theoretically, this should have an impact on the development of phonological awareness, and it does (De Cara & Goswami, 2002).

Proportion of Rime Neighbors

The proportion of RNs (%RN) was then analyzed as a function of the number of overall neighbors (*N*), in order to measure variations in %RN as the total number of neighbors that a word has varies. For example, a word like *date* has 40 neighbors, 26 of which are RNs (65%; e.g., *rate*, *skate*, *late*). A word like *safe* has 16 neighbors, only 2 of which are RNs (12.5%; e.g., *chafe*). In order to compare dense versus sparse areas of the lexicon in terms of similarity structure, we selected two samples at the extremes of the 4,086 database: 1,226 words in a dense neighborhood ($N \geq 36$ for the OVC metric) and 1,271 words in a sparse neighborhood ($6 \geq N \geq 20$ for the OVC metric). The results showed that the proportion of RNs varies considerably as a function of overall neighborhood density (Table 4).

As can be seen from Table 4, the proportion of RNs (%RN) reaches 59.4% in dense *N* and drops to only 41.9% in sparse *N* (OVC metric). In sparse neighborhoods, the proportion of LNs is approximately equivalent to the proportion of RNs. Similar patterns were found for the Ph±1 metric (%RN = 48.0% in dense *N* and 35.4% in sparse *N*). These results show that RNs predominate in dense neighborhoods only. It is therefore possible that neighborhood density effects in speech processing may be caused by the high proportion of RNs, rather than by the overall density of neighbors per se. This possibility is illustrated in Figure 1 but awaits empirical test.

Age of Acquisition

There has been some dispute in the developmental literature concerning similarities and differences between the phonological neighborhoods of adults and children. Since children's lexical neighborhoods are smaller than those of adults and are constantly being updated (which means that neighborhood statistics are much more dynamic), estimates of neighborhood similarity based on adult data can only approximate the developmental picture (e.g., Charles-Luce & Luce, 1990, 1995; Dollaghan, 1994; Logan, 1992). The best method for calculating phonological neighborhood density in children would be to create a database of all early-acquired words. An approximation to this is possible via age-of-acquisition (AoA) norms. We therefore analyzed neighborhood statistics for the AoA data reported in Gilhooly and Logie (1980). These measures were based on adults' AoA ratings on a 7-point scale (1, *age of 0–2 years*; 7, *age 13 years and older*) for 1,944 words. A total of 632 of these (all monosyllabic) were found in the 4,086 database. Specific lexicons were then created for words acquired by the age of 3 (216 words), 4 (436 words), 5 (565 words), 6 (614 words), and 7 (632 words). The percentage of RNs among the words familiar to 3-year-olds was 49.8%, and corresponding percentages for 4-, 5-, 6-, and 7-year-olds were 54.8%, 56.2%, 56.7%, and 57.1%, respectively. The mean number of neighbors for each age is shown in Figure 2.

Roughly similar patterns were found using the AoA data reported by Morrison, Chappell, and Ellis (1997): RNs were 46.7% (41 words) for 3-year-olds, 40.6% (76 words) for 4-year-olds, 50.0% (116 words) for 5-year-olds, and 47.8% (126 words) for 6-year-olds. We can thus be fairly confident that the neighborhood relations for the adult monosyllabic lexicon are broadly characteristic of the child's constantly changing lexicon.

Discussion of the OVC and Ph±1 Metrics

As was previously stated, the OVC metric includes more words in a neighborhood than does the Ph±1 metric, because the OVC metric is based on the 3-D coding of syllables (onset–vowel–coda), whereas the Ph±1 metric is

Table 4
Distribution of Rime Neighbors (RNs), Consonant Neighbors (CNs), and Lead Neighbors (LNs) as a Function of Neighborhood Density (*N*, by type) From the 4,086 database

		OVC Metric				Ph±1 Metric			
Calculation	Measure	*RN*	*CN*	*LN*	*N*	*RN*	*CN*	*LN*	*N*
Dense *N*	*M*	30.0	9.9	10.6	50.5	16.4	9.9	7.8	34.1
	SD	13.2	5.3	6.0	13.1	9.3	5.3	4.3	14.2
	%	59.4	19.6	21.0	100.0	48.0	29.1	22.9	100.0
Sparse *N*	*M*	5.9	1.7	6.4	14.0	2.3	1.7	2.5	6.6
	SD	4.6	2.1	4.5	4.1	2.3	2.1	2.5	4.4
	%	41.9	12.4	45.8	100.0	35.4	26.5	38.2	100.0

Note—Dense *N* includes words whose $N \geq 36$ (OVC metric, 1,226 words). Sparse *N* includes words whose $6 \geq N \geq 20$ (OVC metric, 1,271 words).

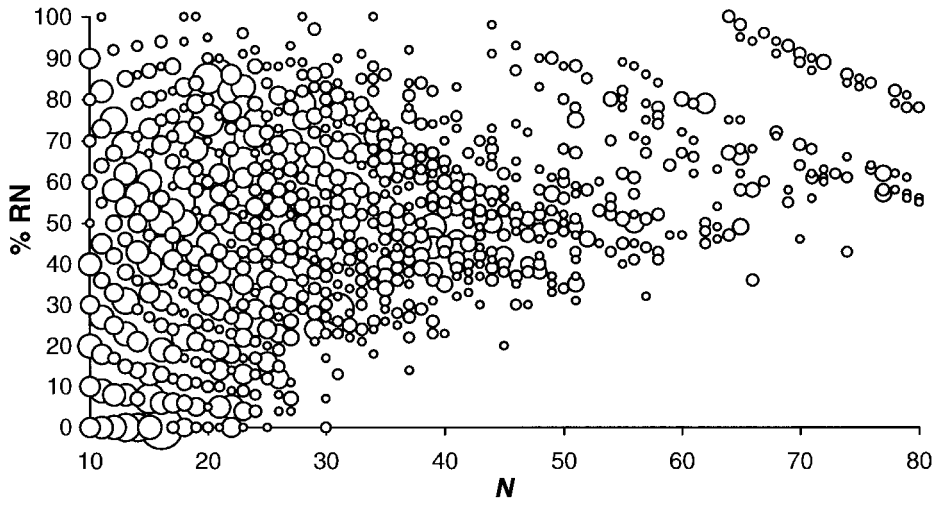

Figure 1. Percentage of rime neighbors (*%RN*) as a function of overall neighborhood density (*N*; OVC metric). The area of each circle is proportional to the number of words represented.

based on the phonemic coding of syllables. The mean *N* is accordingly 29.6 for the OVC metric and 17.9 for the Ph±1 metric in the 4,086 database. In addition, the absolute difference between OVC and Ph±1 metrics is higher for RNs (16.0 − 7.9 = 8.1) than for LNs (8.5 − 5.0 = 3.5). Thus, the OVC metric gives an advantage to RNs, rather than to LNs, as is shown in Table 3. Phonotactic constraints between lead/coda versus onset/rime and the greater differentiation of onsets probably account for this difference. It is known that phonotactic constraints are stronger within the rime (i.e., vowel–coda) than within the lead (i.e., onset–vowel). For example, Kessler and Treiman (1997) found a significant connection between the vowel and the coda (i.e., vowel–coda combinations being more frequent than would be expected by chance), whereas they did not find significant associations between the onset and the vowel in English monosyllabic words. Our analysis showed the same patterns. The constraints for lead/coda combinations were much stronger than those for onset/rime combinations. In the 4,086 database, there are 3,697 different phonological entries (i.e., 389 words are homophonic nonhomographs of other words). Regarding the lead/coda segmentation, there are 768 different leads and 102 different codas. So, the number of possible combinations between lead and coda is 768 × 102 = 78,336. Only 3,697 of these lead/coda combinations (4.7%) cor-

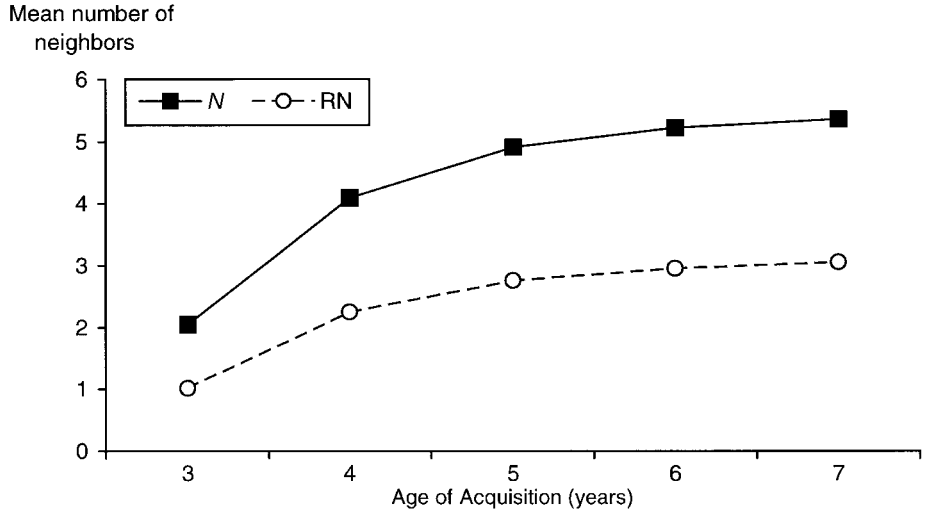

Figure 2. Phonological neighborhood density (*N*) for words acquired between 3 and 7 years of age (OVC metric), based on the age of acquisition data reported in Gilhooly and Logie (1980) for the ages of 3 (216 words), 4 (436 words), 5 (565 words), 6 (614 words), and 7 (632 words).

respond to real words. Regarding the onset/rime segmentation, there are 80 different onsets and 613 different rimes. So, the number of possible combinations between onset and rime is $80 \times 613 = 49{,}040$. Only 3,697 of these onset/rime combinations (7.5%) correspond to real words. Therefore, the constraints between lead and coda combinations are stronger (occupied space, 4.7% only) than the constraints between onset and rime combinations (occupied space, 7.5%).

Gupta and Dell (1999) proposed a psychological reason for the lead/rime asymmetry, suggesting that it could have a temporal basis. They argued that the need to retrieve the sounds of words in sequence favors the development of a vocabulary with an onset-rime organization. If too many competing words share initial sounds (e.g., *hat–ham*), interference delays retrieval. If vocabulary evolution instead favors rime neighbors (e.g., *hat–cat*), rapid retrieval is facilitated. Sevald and Dell (1994) verified this empirically. They showed that word production was slower for word pairs like *cat* and *cab* (shared lead) than for word pairs like *cat* and *bat* (shared rime). They argued that it was difficult to recite words with shared initial sounds, because of sequential interference, and that sequential interference should also affect word recognition. Hence, some vocabularies are better than others: "A good lexicon's words are not randomly distributed in phonological space" (Gupta & Dell, 1999, p. 452). On Gupta and Dell's view, onset-rime vocabulary structure is an emergent property of the sequential temporal processing of human speech.

This view makes two predictions. First, the same evolutionary argument should be applicable to other languages. If onset-rime structure is an emergent property of sequential processing, rime neighbors should predominate in the vocabularies of other languages as well. So far, statistical analyses comparable to those presented here have been run for French monosyllables by Ronald Peereman and for German monosyllables by Johannes Ziegler (see Goswami, 2002, for an overview). These analyses showed a preponderance of RNs in the monosyllabic lexicons of French and German. Second, similar tendencies should be apparent for multisyllabic words. For example, neighbors like *abort* and *about* (shared beginning sounds) should be less frequent than neighbors like *pocket* and *rocket* (shared end sounds). This also seems to be the case. The most frequent sound change in English bisyllables is to change the first consonant phoneme: Neighbors like *pocket* and *rocket* account for 43.5% of the neighbors of bisyllables in Celex, whereas neighbors like *abort* and *about* account for 8.4% of the neighbors.[4] Of course, there are a variety of linguistic syllabification rules, and creation of a database comparable to that for monosyllables is therefore more complicated.[5] Nevertheless, multisyllabic words will also affect lexical processing.

CONCLUSION

In English, similarity relations between spoken words are not equally distributed in phonological space. The similarity structure of the monosyllabic lexicon emphasizes the rime. Furthermore, RNs predominate in dense phonological neighborhoods. Therefore, neighborhood density effects in speech-processing tasks may be caused by the high proportion of rime neighbors, rather than by the overall density of neighbors per se. This certainly seems to be the case for children in phonological awareness tasks (De Cara & Goswami, 2002). Hence, a similarity-based analysis of spoken word forms provides an additional tool for designing experiments to investigate the role of lexical factors in phonological development and speech processing.

REFERENCES

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

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory & Language*, **44**, 568-591.

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighborhoods of words in young children's lexicons. *Journal of Child Language*, **17**, 205-215.

Charles-Luce, J., & Luce, P. A. (1995). An examination of similarity neighbourhoods in young children's receptive vocabularies. *Journal of Child Language*, **22**, 727-735.

De Cara, B., & Goswami, U. (2002). Vocabulary development and phonological neighbourhood density effects in 5-year-old children. Manuscript submitted for publication.

Dollaghan, C. A. (1994). Children's phonological neighbourhoods: Half empty or half full? *Journal of Child Language*, **21**, 257-271.

Ferguson, C. A. (1986). Discovering sound units and constructing sound systems: It's child's play. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 36-51). Hillsdale, NJ: Erlbaum.

Fowler, A. E. (1991). How early phonological development might set the stage for phoneme awareness. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological processes in literacy: A tribute to Isabelle Y. Liberman* (pp. 97-117). Hillsdale, NJ: Erlbaum.

Gilhooly, K. J., & Logie, R. H. (1980). Age of acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, **12**, 395-427.

Goswami, U. (2002). In the beginning was the rhyme? A reflection on Hulme, Hatcher, Nation, Brown, Adams, & Stuart. *Journal of Experimental Child Psychology*, **82**, 47-57.

Goswami, U., & Bryant, P. E. (1990). *Phonological skills and learning to read*. Hillsdale, NJ: Erlbaum.

Gupta, P., & Dell, G. S. (1999). The emergence of language from serial order and procedural memory. In B. MacWhinney (Ed.), *The emergence of language* (pp. 447-481). Hillsdale, NJ: Erlbaum.

Hulme, C., Hatcher, P. J., Nation, K., Brown, A., Adams, J., & Stuart, G. (2002). Phoneme awareness is a better predictor of early reading skills than onset-rime awareness. *Journal of Experimental Child Psychology*, **82**, 2-28.

Jones, D., Roach, P., & Hartman, J. (1997). *English pronouncing dictionary*. Cambridge: Cambridge University Press.

Jusczyk, P. W. (1986). Toward a model of the development of speech perception. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 1-33). Hillsdale, NJ: Erlbaum.

Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, **21**, 3-28.

Kessler, B., & Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory & Language*, **37**, 295-311.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning & Verbal Behaviour*, **12**, 119-131.

Logan, J. S. (1992). A computational analysis of young children's lexi-

cons. In *Research on Speech Perception* (Tech. Rep. No. 8). Bloomington: Indiana University, Department of Psychology.
Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon*. Unpublished doctoral dissertation, Indiana University.
Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, **19**, 1-36.
Maclean, M., Bryant, P. E., & Bradley, L. (1987). Rhymes, nursery rhymes and reading in early childhood. *Merrill-Palmer Quarterly*, **33**, 255-282.
Metsala, J. L. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *Journal of Educational Psychology*, **91**, 3-19.
Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89-120). Hillsdale, NJ: Erlbaum.
Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, **7**, 323-331.
Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology*, **50A**, 528-559.
Peereman, R., & Content, A. (1997). Orthographic and phonological neighbors in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory & Language*, **37**, 382-410.
Sevald, C. A., & Dell, G. S. (1994). The sequential cueing effect in speech production. *Cognition*, **53**, 91-127.
Treiman, R. (1988). The internal structure of the syllable. In G. Carlson & M. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 27-52). Dordrecht: Kluwer.
Treiman, R., & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory & Language*, **27**, 87-104.
Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory & Language*, **40**, 374-408.
Walley, A. (1993). The role of vocabulary development in children's spoken word recognition and segmentation ability. *Developmental Review*, **13**, 286-350.
Walley, A. C., & Flege, J. E. (1999). Effects of lexical status on native and non-native vowel perception: A developmental study. *Journal of Phonetics*, **27**, 307-332.

NOTES

1. http://www.kun.nl/celex/.

2. We have also manipulated rime neighborhood density in different phonological awareness tasks with children and have found effects on rhyme judgment and short-term memory tasks even when overall neighborhood/phonotactics are controlled (see De Cara & Goswami, 2002). Such behavioral evidence supports the psycholinguistic arguments concerning LRT made here.

3. Each monosyllabic word was phonetically coded on a nine-slot sequence (one phoneme per slot). Words were centered on the vowel slot (which is the only obligatory element in the syllable). From the vowel, prevocalic consonants (onset) were coded on four slots from right to left (o1, o2, o3, o4), and postvocalic consonants (coda) were coded on four slots from left to right (c1, c2, c3, c4). The o1 and c1 slots stand closer to the center of the syllable than do the o2 and c2 slots, and so on. Empty slots were coded with a dot. For example, a word like *skill* was coded as [..skil…] for the o4–o3–o2–o1–V–c1–c2–c3–c4 nine-slot sequence, and a word like *wind* was coded as […wind..]. For the OVC metric, *ND* represents the number of all phonological neighbors that differ from a target word by a one-onset, vowel, or coda substitution, deletion, or addition (i.e., one-slot change out of the three-slot sequence). Among the OVC neighbors only, *RN* represents the number of neighbors sharing the rime (i.e., the V–c1–c2–c3–c4 sequence), *CN* represents the number of neighbors sharing the consonants (both the o4–o3–o2–o1 and the c1–c2–c3–c4 sequences), and *LN* represents the number of neighbors sharing the lead (the o4–o3–o2–o1–V sequence). For the Ph±1 metric, *ND* represents the number of all phonological neighbors that differ from a target word by a one-phoneme substitution, deletion, or addition (i.e., one-slot change out of the nine-slot sequence). Among the Ph±1 neighbors only, *RN* represents the number of neighbors sharing the rime, *CN* represents the number of neighbors sharing the consonants, and *LN* represents the number of neighbors sharing the lead. For both metrics, the calculation by type is based on the absolute number of neighbors. The calculation by token is based on the cumulated lexical spoken frequencies of neighbors. Any target word is not a neighbor of itself.

4. For this analysis, we applied the OVC metric to bisyllabic words. Bisyllabic words were viewed as a sequence of two syllables. In the OVC metric, each syllable was coded on a three-slot sequence (O–V–C). Thus, each bisyllabic word was coded on a six-slot sequence, O1V1C1–O2V2C2. Bisyllabic words that differed by the substitution, addition, or deletion of one slot (out of six) were counted as phonological neighbors. We then ran a preliminary *N* calculation for all the bisyllabic words found in the Celex corpus (21,648 words; Baayen et al., 1995). We reduced this to 16,970 words, since we only kept distinct phonological entries (i.e., 4,678 words were removed because they were homophones of other words in the database). We then defined six subtypes of phonological neighborhood for bisyllabic words, one subtype per position of change on the six-slot sequence. Examples of the subtypes and their neighborhood distribution follows:

O1 change: pocket/rocket	43.5%
V1 change: racket/rocket	14.2%
C1 change: sector/center	5.3%
O2 change: hunger/hunter	18.7%
V2 change: abort/about	8.4%
C2 change: device/divide	9.8%

5. A number of possible principles of syllabification could be used as a basis for creating such a database (e.g., among others, cluster legality, stress position, and speaking rate), but linguists disagree, and behavioral syllabification does not always match linguistic syllabification rules (e.g., Treiman & Danis, 1988). The most appropriate metric for phonological neighborhood calculation is also unclear. Nevertheless, the change of C1 seems to predominate, since we also explored a five-slot coding metric, treating the middle consonant as one slot (this reduced the database to the 7,694 English bisyllables with a single intervocalic consonant). Using this five-slot coding, the percentage of super-rime neighbors (*rocket–pocket*) was slightly higher (47.5%), whereas the percentage of neighbors differing at the end was approximately the same (*abort–about*, 7.3%).

APPENDIX

The 4,086 database can be found at http://www.ich.ucl. ac.uk/ich/html/academicunits/behav_brain_sci/database/similarity_words.html. Both OVC and Ph±1 metrics are supplied. The subtypes of neighbors (RNs, CNs, and LNs) are also indicated. Words are sorted by rime, and rimes are sorted by rime neighborhood density (RND, decreasing order). Within RND, words are sorted by spoken lexical frequency. Computerized routines for calculating the neighbors of a target word or nonword are supplied on the same Web site.

Key to Phonetic Codes

Vowels				Consonants			
Phonetic Class	Code	Examples	Occ.	Phonetic Class	Code	Examples	Occ.
Short vowels	ï	sit	447	Glides	y	yes	108
	ë	bed, head	302		w	wet	301
	x	the, again	4				
	a	hot, what	281	Liquids	r	red	722
	^	cup, come	324		l	leg	918
	°	book, put	39				
	@	man, have	358	Nasals	m	man	451
					n	nod	657
Long vowels	i	sheet, teach	308		¿	sing	131
	ä	are, car	207				
	c	sort, walk	250	/h/	h	horn	159
	®	turn, heard	154				
	u	boot, who	261	Weak fricatives	f	fill	411
	%	thin	122				
Diphthongs	e	came, way	351		v	van	146
	¥	my, wine	257		∂	then	41
	ø	boy, voice	54				
	&	out, now	119	Strong fricatives	s	sit	1,085
	o	home, load	255		$	she	181
	ì	beer, dear	45		z	zone	246
	è	hair, care	55		§	azure	4
	ù	tour, cure	15				
				Affricates	©	chair	207
					J	joke	164
				Stops	p	pen	636
					t	to	966
					k	keep	852
					b	but	418
					d	day	564
					g	go	297

Note—Occ., phoneme occurrence in the 4,086 database.

(Manuscript received April 25, 2001;
revision accepted for publication March 21, 2002.)