A developmental analysis of similarity neighborhoods in European Portuguese

SELENE VICENTE
SÃO LUÍS CASTRO
AMANDA WALLEY

Abstract

We present a developmental analysis of the structural organization of young children's and adults' lexicons for European Portuguese. The production lexicons of 3-, 4-, and 5-year-olds, a receptive lexicon for 12- to 19-month-olds, and an adult lexicon were compared using the similarity neighborhood paradigm (e.g., Charles-Luce & Luce, 1990). For each lexicon, similarity neighborhoods were computed for words with 3 to 8 phonemes, and phonological neighborhood sizes were compared. A phonological neighbor was defined as any word in one of the lexicons that differed from a given target by one phoneme substitution, deletion, or addition. Results showed structural differences between shorter (3-, 4- and 5-phoneme) and longer (6- to 8-phoneme) words. There was no age effect for longer words, of which ca. 92% had no neighbors. Shorter words, in contrast, had more neighbors: in the children's lexicons, ca. 58% of shorter words had one to four neighbors, and 8% had five to seven neighbors; only ca. 36% had no neighbors. An age effect was found, whereby similarity neighborhoods become increasingly dense over the course of childhood. The results are discussed in light of previous findings for English-speaking children and adults, and their implications for the development of spoken word recognition by Portuguese listeners are considered.

1. Introduction

One salient aspect of language acquisition is the substantial increase in vocabulary that occurs in early through middle childhood. How does this increase in vocabulary size affect the structural relations that are established among words in the mental lexicon, and the very nature of the child's lexical representations? In particular, is there a developmental trend toward relating words in terms of their phonemic similarity to one another? And what impli-

***J*ournal of *P*ortuguese *L*inguistics, 2 (2003), 115-133** ISSN 1645-4537

cations might such a trend have for the process of spoken word recognition? To date, answers to these questions have come almost exclusively from studies conducted with English-speaking children. In this paper, we report structural analyses of young children's and adults' lexicons in European Portuguese. The results should be informative regarding the process of language acquisition in general, and at the same time contribute to our understanding of possible language-specific characteristics of the developing lexicon.

The question of age-related structural changes in the lexicon is important not only from a descriptive point of view, but also because these changes may help to explain differences between children and adults in spoken word recognition. It is well documented that adult listeners can quickly and accurately identify words on the basis of only partial acoustic-phonetic information, and this has been taken as evidence in favor of the theoretical claim that adult lexical representations are segmentally structured (see Luce & Pisoni, 1998). However, preschool and school children need more speech input than adults in order to recognize even simple, monosyllabic familiar words, and the amount of input needed to recognize words decreases significantly with age (see Walley, 1993). Further, the results from a variety of experimental paradigms, such as similarity judgment (e.g., Treiman & Baron, 1981; Treiman & Breaux, 1982; Walley, Smith, & Jusczyk, 1986) and mispronunciation detection (e.g., Cole & Perfetti, 1980; Walley, 1987; Walley & Metsala, 1990), indicate that segmental information is either not used by young children in the recognition process or that it is not very salient to them. Overall then, these observations suggest that early lexical representations may be more holistic or undifferentiated than they are for older listeners (for reviews, see Walley, 1993; Metsala & Walley, 1998).

A developmental trend toward more fine-grained, segmental lexical representations that arises from the demands of a rapidly growing vocabulary has recently been proposed. According to the Lexical Restructuring Model, LRM, (Walley, 1993; Metsala & Walley, 1998; Garlock, Walley, & Metsala, 2001; Walley, Metsala, & Garlock, 2003), increasing vocabulary size is a key factor in the shift away from holistic toward more segmental representations of spoken words. As the number of words in a child's lexicon grows, there should be a corresponding increase in the acoustic-phonetic overlap among words that are known, and consequently higher confusability among words in memory. That is, from one to two years of age, the child's lexicon is quite small, consisting of about 50 words that are acquired slowly, one at a time. At this point, there would be little need to represent words in a detailed, fine--grained manner because they can be discriminated by overall acoustic shape, number of syllables, and/or prosodic features. However, typically around 18 months, there is a sudden and large increase in the words that children can comprehend and produce (e.g., Reznick & Goldfield, 1992), and such vocabulary growth continues to be substantial through middle and late childhood. For example, it has been estimated that first-graders (aged 7)

understand as many as 7,000-10,000 root words, and that fifth-graders (age 11) know 39,000-46,000 words (Anglin, 1989). In the face of such expansion, presumably the need arises to distinguish among words of heightened phonological similarity; this may force the child to add phonological information to lexical representations and to restructure them into segment sized units (see also Charles-Luce & Luce, 1990, 1995; Logan, 1992; Storkel, 2002).

Analyses of the phonological relationships among words in children's lexicons at different points in development are important to evaluate the hypothesis that vocabulary growth might contribute to the heightened specificity of the phonological representations, namely their segmental nature. Charles-Luce and Luce (1990) were the first to provide such an analysis. They examined phonological similarity neighborhoods for the production lexicons of 5- and 7-year-old children (based on Wepman & Hass, 1969; N = 670 and 943) and an adult lexicon (Webster's Pocket Dictionary; N = 20,000). A similarity neighborhood was defined as a set of words that differed from a particular lexical target by a single phoneme addition, deletion or substitution. Similarity neighborhoods were computed for words of 3, 4 and 5 phonemes, because words of these lengths accounted for about 75% of each child lexicon (and for ca. 40% of the adult lexicon). They found that 3-phoneme words were the subset of words from denser neighborhoods (more phonological neighbors) in the children's lexicons (6 and 5% had 0 neighbors, 84 and 78% had 1 to 7 neighbors, and 10 and 17% had 8 to 12 neighbors, respectively, for 5- and 7-year-olds). Neighborhood density decreased sharply for the other word lengths. Developmental comparisons showed that neighborhoods in the adult lexicon were much more densely populated than children's neighborhoods. In addition, comparisons between the lexicons of 5- and 7-year-olds revealed a shift toward denser similarity neighborhoods. These findings were interpreted as indicating that spoken word recognition by young children may be supported by fairly global processes or strategies; however, as the lexicon grows and an increasing number of words exhibit structural overlap, recognition would have to become more segmentally based.

Subsequently, Dollaghan (1994) examined phonological similarity neighborhoods for even younger children. Specifically, she constructed a lexicon to represent the production vocabularies of 1- to 3-year-olds that was composed of monosyllabic wordforms with 3 and 4 phonemes (N = 407). In this lexicon, more than 80% of the words had at least one phonological neighbor, and nearly 20% had six or more neighbors. On the basis of these and other results, Dollaghan argued that toddlers and preschoolers must have access to fairly fine-grained lexical representations in order to distinguish among phonologically similar entries from the earliest stages of lexical acquisition, and called into question the claim that young children use more global recognition strategies than do older children and adults. Further, she suggested that age-related increases in neighborhood density are simply due to

increases in vocabulary size, and that production databases likely underestimate the number of words that children have stored in memory.

In a response to these criticisms, Charles-Luce and Luce (1995) analyzed the similarity neighborhoods of a receptive database (N = 743) extracted from a corpus of child-directed speech from mothers to 1- and 2-year-old children (Bernstein-Ratner, 1987). This analysis replicated the findings of the original study. In particular, the lexicon of older infants/toddlers was found to be less dense than that of adults, and more skewed toward sparse neighborhoods. This was the case even when only words that would be highly familiar to children were used in computing similarity neighborhoods. Further, although neighborhood density increases with vocabulary growth, such growth is not sufficient to explain the differences that were found. As Charles-Luce and Luce pointed out earlier (1990, p. 213) "… neighborhood density and the number of POSSIBLE overlapping words are not completely dependent." Indeed, word length is an important variable that must be taken into account. For example, in the adult lexicon, there are more words with 4 and 5 phonemes, but they have fewer neighbors than 3-phoneme words (i.e., the latter are less numerous, but have denser neighborhoods).

In sum, we have begun to learn more about changes in the structure of the lexicon over the course of childhood, and about the implications of these changes for the development of spoken word recognition. However, many, if not all of these recent findings come from work done with English-speaking children and adults. How generalizable are these findings? The present study will provide much-needed information from another language, European Portuguese (EP), and thus help to show how general or language-specific these findings are. We might expect a similar developmental trend in similarity neighborhoods. However, it is possible that there are language-specific differences. One reason to expect such differences is that many early-acquired words that are known by young Portuguese children are polysyllabic, whereas English children know many monosyllabic words. Therefore, we might not expect to find exactly the same pattern of neighborhood density. Increases in neighborhood density as a function of age might not be so marked for EP as for English, a result that could have important implications for spoken word recognition.

In the present study, we analyzed changes in the structural organization of European Portuguese young children's lexicons across age using the similarity neighborhood paradigm. The following lexical databases were compared: the Viana lexicon consisting of the verbal productions of 3-, 4-, and 5-year-old children (Vicente, 2002), the Ramos-Pereira receptive lexicon (age of child: 12 to 29 months; Ramos-Pereira, 1992; see also Vicente, 2002), and the Porlex adult database (Gomes, 2001; Gomes & Castro, 2003). Our main goal was to assess whether words in the young Portuguese child's lexicon tend to be more discriminable overall, on a structural basis, than those in the older child's and adult's lexicons, as has been reported for English-speaking

children. Analyses of computerized lexical databases can provide detailed characterizations of the structural properties of words in a language, and such information is lacking for European Portuguese. Once this groundwork has been laid, it will be possible to begin formulating and assessing principled hypotheses concerning actual spoken word recognition performance.

2. Method

2.1. Databases

The Viana lexicon (Vicente, 2002) is based on two main *corpora*: the *corpus* Viana Bilan de Langage, which contains the productions of 3- and 4-year-old children in a semi-structured play situation lasting about 30 minutes (Le Normand, 1989); and the *corpus* Viana Frog Stories, which contains the productions of 5-year-old children who were asked to tell the story from Mayer's (1969) picture book "Frog where are you?" (Castro, Delgado-Martins, Gomes, Amorim, & Pimenta, 1996; Delgado-Martins, Castro, Gomes, Amorim, & Pimenta, 1998). Both these *corpora* resulted from a longitudinal study with 3- to 5-year-old children. The mean ages of the three groups of children (N = 95) and the number of speech samples used to compile the Viana lexicon can be seen in Table 1. A speech sample is a collection of words produced in one observation session. It was obtained from the corresponding transcriptions with the CHAT coding system by using the program FREQ of the Child Language Data Exchange System (CHILDES; MacWhinney & Snow, 1990; MacWhinney, 1995).

Table 1. Mean age (M), standard deviation (SD), range, and number of speech samples (N = 173) in the three age groups used to construct the Viana Lexicon.

Groups	*M*	*SD*	Range	# speech samples
3 years	3.6	.27	2.96 – 4.12	78
4 years	4.8	.33	4.13 – 5.39	63
5 years	5.9	.30	5.42 – 6.46	32

The Viana lexicon is organized into two types of databases for each age group: wordform and lemma databases. The wordform databases include inflected content and function words; the lemma databases contain only uninflected nouns, adjectives, verbs, adverbs, and numerals. For example, the lemma

(1) *caixa* [kajSA]
'box'

includes the wordforms

(2) *caixa* [kajSA]; *caixas* [kajSAS]; caixinha [kajSiNA].
'box' 'boxes' 'little box'

Proper names (including names of animals and places) and onomatopoeias were excluded. In the present study, only the lemma databases, one for each age group, were used. Each entry for a word includes its corresponding orthographic form, a phonetic transcription, a grammatical classification, and a frequency count. Examples of entries are shown in Appendix A. The phonetic transcriptions were based on Porlex (Gomes, 2001; Gomes & Castro, 2003), and correspond to a careful pronunciation of a given word in isolation. Frequency counts were obtained by adding the number of occurrences of corresponding wordforms.

Table 2 shows the number (and percentage) of entries in the Viana production lexicon for 3-, 4- and 5-year-olds as a function of grammatical class.

Table 2. Number (and %) of entries per grammatical class in the Viana, Ramos-Pereira (R-P) and Porlex lexicons/databases.

	Viana lexicon			R-P	Porlex
Class[a]	3-years (*N* = 694)	4-years (*N* = 995)	5-years (*N* = 1091)	1-2 years (*N* = 1374)	Adult (*N* = 27,063)
No	373 *(53.7)*	801 *(42.4)*	606 *(55.5)*	742 *(54.0)*	16,313 *(60.3)*
Aj	85 *(12.2)*	256 *(13.6)*	161 *(14.8)*	218 *(15.9)*	4230 *(15.6)*
Vb	185 *(26.7)*	754 *(40.0)*	258 *(23.6)*	319 *(23.2)*	5449 *(20.1)*
Av	39 *(5.6)*	57 *(3.0)*	48 *(4.4)*	56 *(4.1)*	1019 *(3.8)*
Nu	12 *(1.7)*	19 *(1.0)*	18 *(1.6)*	39 *(2.8)*	52 *(0.2)*

Note. [a]no = noun, aj = adjective, vb = verb, av = adverb, nu = numeral.

The number of entries in the 4-year-old lexicon is the sum of those for the 3- and 4-year-old lexicons; similarly the 5-year-old lexicon includes the entries for younger children. Also shown are the entries in a receptive lexicon for 1- to 2-year-olds, the Ramos-Pereira database (see Vicente, 2002). This lexicon contains 1,374 lemmas and was constructed based on the Ramos-Pereira *corpus* of speech directed to 12- to 29-month-olds (*N* = 93,880; Ramos-Pereira, 1992). Comparisons between these child lexicons and the adult lexicon were made by reference to the content words of the Porlex database (*n* = 27,063; Gomes & Castro, ibd.).

2.2. Procedure

The structural organization of these lexicons was analyzed using the similarity neighborhood paradigm (e.g., Charles-Luce & Luce, 1990). A similarity neighborhood was defined as a set of words that differed from a particular target by a one phoneme substitution, addition, or deletion. For example, the phonological similarity neighborhood for the Portuguese word

(3) *gato* [gatu]
'cat'

includes, among others, the words

(4) ***p****ato* [patu]; ***f****ato* [fatu]; *ga****l****o* [galu]; g**r**atu [gratu]; acto [atu]
'duck' 'suit' 'rooster' 'grateful' 'act'

For each lexicon/database, we analyzed the number and percentage of words as a function of word length in phonemes, and the number and percentage of neighbors that words of a given length have. Only the neighborhoods of 3- to 8-phoneme words were analyzed, because words of this length account for the vast majority of those in the children's lexicons. Furthermore, for the words in the children's lexicons (i.e., Viana and R-P lexicons), similarity neighborhoods were also analyzed with reference to the adult Porlex database.

3. Results and Discussion

3.1. Word length

Figure 1 shows the percentage of words in the 3-, 4- and 5-year-olds' production lexicons (Viana database) as a function of length. As one can see, these lexicons consist primarily of words ranging from 3 to 8 phonemes; words of this length comprise over 93% of all the items in each of the three lexicons. The peak in these distributions occurs for 5-phoneme words, which represent about 25% of the lexicon for each age group. Very short and very long words represent about 7% of all the words in these lexicons: i.e., words with up to 2 phonemes represent less than 2% of the words in the lexicons, and words with 9 to 14 phonemes represent about 5% only. Despite the similarities in the pattern of word length distribution for the three age groups, the 4- and 5-year-olds' lexicons, when compared to that of 3-year-olds, contain more words with 7 to 8 phonemes (ca. 43% vs. 38%, respectively) and fewer words with 3 to 5 phonemes (ca. 50% vs. 55%, respectively). These developmental gains and losses thus occur for a fairly small portion of the lexicon for these particular age groups.

In the receptive lexicon for 1- to 2-year-olds (Ramos-Pereira database), the distribution of words as a function of length is very similar (see also Figure 1). There are more words with 6 phonemes (ca. 22% of the lexicon), and 92% of this lexicon consists of words ranging from 3 to 8 phonemes in length. Very short and very long words represent a small percentage of the lexicon (ca 2% and 8%, respectively).

The picture is quite different, however, for the adult lexicon (see Figure 1). The peak in the adults' distribution occurs for words with 7 to 8 phonemes (each length representing 16% of the lexicon), and the most frequent words range from 4 to 12 phonemes, which comprise over 95% of all words. The percentage of 3-phoneme words drops to less than 1% (cf. 5 and 4% in the Viana and R-P databases, respectively), and 4% of the words are 13 to 20 phonemes in length. Words with 3 to 5 phonemes account for about 50% of the children's lexicons (50 and 44% in the Viana and R-P databases, respectively), but only for 14% of the adult lexicon.

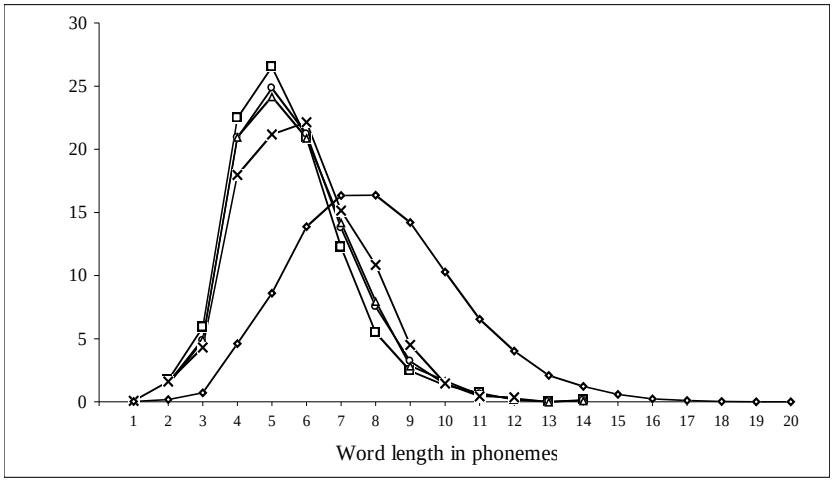


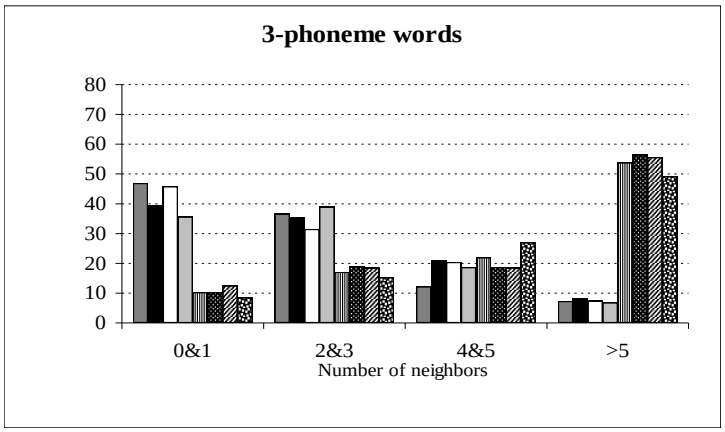
Figure 1. Percentage of words as a function of length in the Viana (3-, 4-, and 5-year-old), Ramos-Pereira (R-P; 1- to 2-year-old), and Porlex (adult) lexicons/databases.

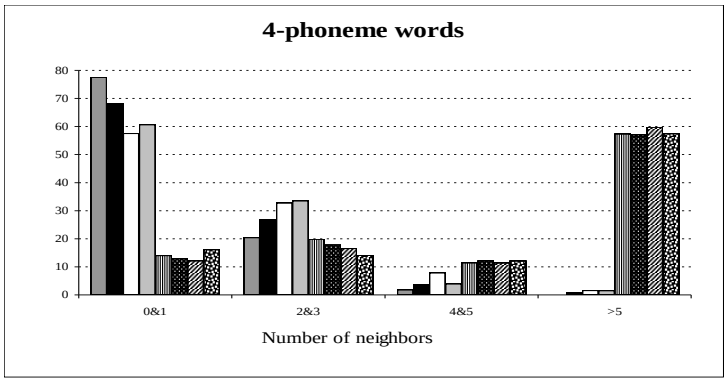
3.2. Similarity neighborhoods

Figure 2 shows the results of the similarity neighborhood analyses for words with 3 to 5 phonemes, and also for words with 8 phonemes. As one can see, there is a pronounced decrease in neighborhood size as a function of word length. On the whole, short words have more neighbors (come from denser neighborhoods) than do longer words (i.e., 3-, 4- and 5-phoneme words vs. words with 6 to 8 phonemes). Furthermore, there is a striking difference in neighborhood size when this is evaluated with reference to the adult vs. child lexicons. Specifically, in each analysis, there is a larger percentage of words with more neighbors – that is, adult neighborhoods are denser than are children's. The results are described in more detail below. Examples of the similarity neighborhoods computed for the child and adult databases are shown in Appendix C.

3.2.1. Three-phoneme words

Words with 3 phonemes comprise only a small part of children's lexicons: 5 to 6% in the Viana lexicon, and 4% in the R-P lexicon (see Figure 1). As shown in the top panel of Figure 2, neighborhood density for words of this length ranges from 0 to 6 neighbors; when an analysis is done for the same words with reference to Porlex, the adult lexicon, the range increases to 13 neighbors. The peak of the distributions in the children's lexicons occurs for words with 1 neighbor (ca. 27, 25 and 32% of the Viana lexicon, respectively; 24% of the R-P lexicon); most of the words have between 1 to 3 neighbors (63, 60 and 63% in the Viana lexicon, respectively; 63% in the R-P lexicon). Age effects are apparent in sparse neighborhoods with 0 or 1 neighbors. There is a decrease of about 8% in the percentage of words with no neighbors between 3 to 5 years of age (from 17 to 9%). For words with 2 and 3 neighbors, there are only small age-related variations. However, neighborhoods with 4 residents show an increase of 6% with age (from 3 to 9%).





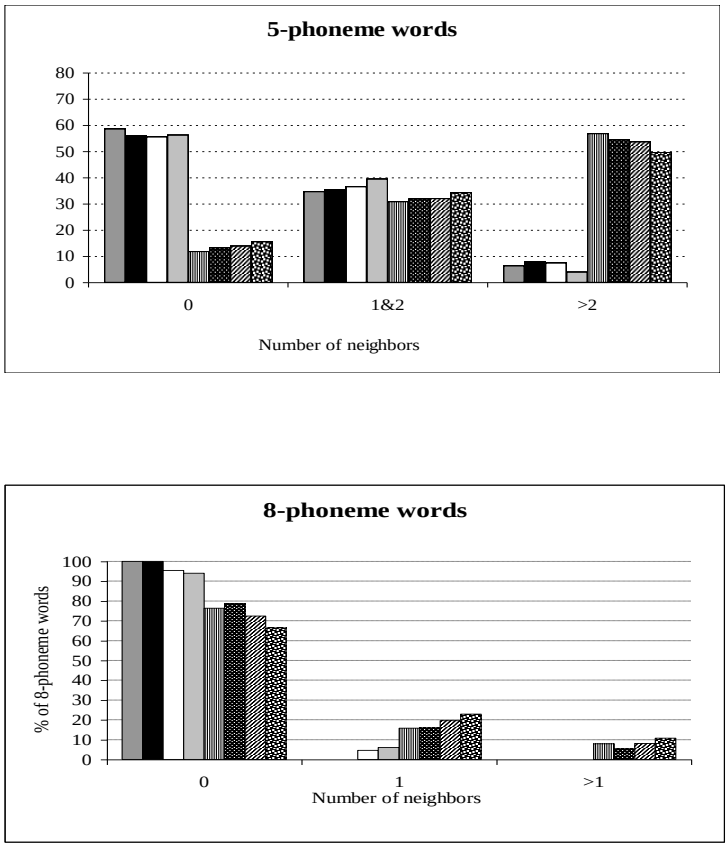


Figure 2. Neighborhood density for words with 3-, 4-, 5-, and 8-phonemes in the Viana (3-, 4-, and 5-year-old) and Ramos-Pereira (R-P; 1-2 year old) lexicons. We also present neighborhood density analyzed with reference to the adult Porlex database.

When neighborhood size is computed for the same words in the adult lexicon, the peak in the distributions moves to words with 7 neighbors for the Viana lexicon comparisons (ca. 17, 17, and 15%, respectively), and to words with 4 neighbors for the R-P lexicon (17%). Words with 0 to 1 neighbors account for less than 8% of the lexicons. Conversely, more than 50% of the words analyzed in the adult lexicon have 6 or more neighbors. For example, the word

(5) *pau* [paw]
'stick'

has 3 neighbors in the 3-year-old and R-P lexicons,

(6) ***m****au* [maw]; *pa**i*** [paj]; *pá* [pa]
'bad' 'father' 'shovel'

4 neighbors in the 4- and 5-year-old lexicons (the same plus)

(7) *pa**z*** [paS]
'peace'

and 7 neighbors in the adult database (the same plus)

(8) ***n****au* [naw]; ***v****au* [vaw]; *par* **[**par**]**
'ship' 'wade' 'pair'

3.2.2. Four- and five-phoneme words

Words with 4 or 5 phonemes comprise about 47% of the Viana lexicon (21 to 23% for 4-phoneme words; 24 to 27% for 5-phoneme words) and slightly less in the R-P lexicon (18 and 21%, respectively). As shown in the middle panels of Figure 2, neighborhood density ranges from 0 to 7 in the children's lexicons, but the majority of the 4- and 5-phoneme words have 0 to 1 neighbors. On the whole, at least 60% of 4-phoneme words (78, 68, and 58% for 3-, 4-, and 5-year-olds respectively; 61% for 1- to 2-year-olds) and 80% of 5-phoneme words (85, 81, and 82%, for 3-, 4-, and 5-year-olds respectively; 86% for 1- to 2-year-olds) have 0 to 1 neighbors. However, the percentage of sparse neighborhoods is higher for 5-phoneme words, 60% of which have no neighbors at all, in each age group. For 4-phoneme words, the peak of the distribution in the older children's lexicons occurs for words with 1 neighbor (ca. 35 and 30% of 4-phoneme words, for 4- and 5-year-olds, respectively). Age effects are especially apparent for 4-phoneme words. Indeed, there is a decrease of about 18% in words with 0 neighbors between ages 3 to 5 (from 46 to 28%). Also, there are increases of about 5 and 7% in words having 2 and 3 neighbors, respectively (from 16 to 21%, and from 4 to 11%). Although neighborhoods with 4 or more residents are relatively rare, they increase from 1% to 7% between 3 and 5 years of age.

The neighborhood density pattern in the R-P lexicon for 4- and 5-phoneme words is similar to that obtained for 5-year-olds in the Viana lexicon. When neighborhood density is calculated in Porlex (see Figure 2), the percentage of words having 0 or 1 neighbors is much smaller (12 to 14% of the 4-phoneme words; 28 to 31% of the 5-phoneme words); in contrast, the percentage of

words with five or more neighbors is much higher (ca. 60% of 4-phoneme words; at least 30% of the 5-phoneme words). For example, the word

(9) *bota* [bOtA]
'boot'

has 2 neighbors in the R-P lexicon

(10) *bo**l**a* [bO**l**A]; ***m**ota* [mOtA]
'ball' 'motorbike'

3 neighbors in the 3-year-old lexicon, the same plus

(11) ***n**ota* [nOtA])
'note'

4 neighbors in the 4- and 5-year-old lexicon, the same plus

(12) *ba**t**a* [batA]
'gown'

and 14 neighbors in the adult database, the same plus

(13) *bot**e*** [bOt6]; ***c**ota* [kOtA]; **j**ota [ZotA]; ***l**ota* [lOtA];
'boat' 'share' 'j' (name of letter) 'fish market'

***r**ota* [RotA]; *bo**g**a* [bOGA]; *bo**ss**a*[bOsA]; *bo**s**ta*[bOStA] *bo**i**a* [bOjA];
'route' 'bream' 'bump' 'crap' 'buoy'

More examples are presented in Appendix C.

3.2.3. Six- to eight-phoneme words

Words with 6 to 8 phonemes account for about 39 to 43% of the Viana lexicon, and 48% of the R-P lexicon. There is a sharp decrease in the percentage of such words as length increases (e.g., ca. 21, 14 and 8% for 6-, 7-, and 8-phoneme words in the 5-year-old lexicon). In the children's lexicons, neighborhood density for words of this length drops sharply (see bottom panel of Figure 2; only 8-phoneme words are shown because the pattern for 6- and 7-phoneme words is very similar); many of these words have no neighbors at all. For example, none of the 8-phoneme words in the 3- and 4-year-old lexicons has neighbors, and in the 5-year-old lexicon only 5% of them have 1 neighbor. The peak in the distributions occurs for words having 0 neighbors: about 86 to 90% of the 6-phoneme words; 92 to 93% of the 7-phoneme words; and 95 to 100% of the 8-phoneme words. Age effects

are small (ca. 3 to 4%) and they are evident only for 6-phoneme words: from the younger to the older children's lexicons, there is a decrease in the percentage of words with 0 neighbors (90 to 86%, respectively), and an increase in the percentage of words with 1 neighbor (10 to 12%, respectively).

Developmental comparisons with reference to Porlex show a consistent pattern of denser neighborhoods for all word lengths. For example, for the 8-phoneme words, although the great majority of words in the child lexicons have no neighbors (ca. 100% at ages 3 and 4), 16 to 20% of the same words analyzed in the adult lexicon possess at least 1 neighbor, and a small percentage of words has 2 to 5 neighbors (ca. 8%). An illustration of this is the word

(14) *professor* [pruf6sor]
'teacher'

which has 0 neighbors in the children's lexicons and 1 neighbor in Porlex,

(15) *profess**a**r* [pruf6sar]
'to profess'

In summary, the similarity neighborhoods for the Portuguese lexicons vary according to word length and age. They are very sparse for words with 6 phonemes or longer; no age-related effects were observed for these words. However, for shorter words, that come from denser neighborhoods, age--related effects were observed.

4. Conclusion

Analyses of word length revealed that words ranging from 3 to 8 phonemes represent the vast majority of European Portuguese children's productive and receptive lexicons (about 93% in the Viana lexicon, and 92% in the R-P lexicon). There is a shift towards longer words in the adult lexicon, where words from 4 to 12 phonemes account for 95% of the lexicon. Accordingly, the peak in the distributions of word length varies in the child and adult lexicons: from 5 phonemes in children's productive lexicons and 6 phonemes in the receptive lexicon, that account each for ca. one quarter of all words in the corresponding lexicons, to 7 and 8 phonemes in the adult lexicon, both with 16% of all words. It is worth mentioning that in the adult lexicon the next frequent word lengths include not only shorter words with 6 phonemes (14% of all words), but also longer 9- and 10-phoneme words (with 14% and 10% of the lexicon, respectively). Thus, it is clear that the increase in vocabulary from early and middle childhood into adulthood, in European Portuguese vs. American English (see below), occurs to a great extent through

the acquisition of longer words. These, because of their very length, cannot be phonemic neighbors of words residing in the younger children's vocabulary.

A different picture emerges from Charles-Luce & Luce's studies (1990; 1995) with 5- and 7-year-old English speaking children and adults. The English lexicons have a relatively higher percentage of shorter words. Words with 3, 4 and 5 phonemes are quite frequent not only in the child, but also the adult lexicons (ca. 75% and 40% of the words, respectively; the corresponding values for the Portuguese lexicons examined here are 50% and 14%). Whereas in Portuguese the peak for the adult lexicon occurs at 7- and 8-phoneme words in English there is a peak for 5-phoneme words, which alone comprise 15% of the adult lexicon. In the Portuguese children's lexicons, the peak occurs for words with 5 phonemes, with ca. 24% of the words at 5 years, whereas in English the peak occurs for 3-phoneme words, with approximately 35% of the words even for 7-year-old children.

Do these differences have an impact on the development of similarity neighborhoods? Similarity neighborhood analyses for European Portuguese, with 3- to 8-phoneme words, revealed that words in the children's lexicons have fewer phonological neighbors than the same words analyzed in the adult lexicon. This finding replicates what has been found for English, and is consistent with the theoretical hypothesis that representation and recognition processes may be more global or holistic in younger children than in adults. However, in European Portuguese there are structural differences between shorter (3-, 4-, and 5-phoneme) and longer (6-, 7- and 8-phoneme) words, in that for the latter there is no developmental increase in neighborhood density; when they are acquired in childhood, they have few neighbors, and they continue to reside in sparse neighborhoods through/into adulthood. In contrast, more than half of the short words in the children's lexicons have 1 to 4 neighbors. Four-phoneme words, which represent about 20% of the children's lexicons (but only 5% of the adult lexicon) are the densest. For these words, developmental gains in density are clearly visible: a decrease of 18% in the percentage of words with 0 neighbors from 3- to 5-years-old and a 6% increase in the percentage of words with 4 or more neighbors. However, for longer words, that in Portuguese are almost as frequent, no clear developmental increase in neighborhood density was observed. This is linked to the fact that neighborhood density decreases sharply with word length, and this is true both for children's, as well as for the adult's, lexicons. Indeed, the majority of words with 6- to 8-phonemes have 0 or 1 neighbor; they are lexical hermits with almost no structural overlap with other words. These structural differences between shorter and longer words in Portuguese are a further indication that vocabulary growth and neighborhood density are not completely dependent. Therefore, age-related neighborhood density changes can not simply be attributed to increases in the overall size of the lexicon.

In English, the developmental increases in neighborhood density reported by Charles-Luce and Luce (1990; 1995) are more marked than in Portuguese,

in particular for the numerous group of the (English) 3-phoneme words. In European Portuguese, there is also a similar developmental shift from sparser to denser neighborhoods, but this shift is clearly dependent on word length. How these structural differences between languages might have an effect on recognition remains to be seen. Future work integrating structural analyses with empirical evidence on children's auditory perceptual skills during lexical processing will help clarify the role of neighborhood similarity in spoken word recognition.

Acknowledgements

This research was supported by a FCT grant to the Center for Psychology at the University of Porto (Language Group). We thank Inês Gomes for her help in performing the structural analyses of the children's lexicons, and for making some of the analyses concerning the adult database Porlex available to us. We also thank two anonymous reviewers for their comments on a previous version of this paper. Correspondence should be addressed to: São Luís Castro, Faculdade de Psicologia e de Ciências da Educação, Universidade do Porto, rua do Campo Alegre, 1021, P 4169 – 004 Porto, Portugal (slcastro@psi.up.pt).

References

Anglin, J. M. (1989) Vocabulary growth and the knowing-learning distinction, *Reading Canada*, **7**, 142-146.

Bernstein-Ratner, N. (1987) The phonology of parent child speech. In *Children's language* (K. Nelson & A.V. Kleeck, editors), Vol. 6. Hillsdale, NJ: Lawrence Erlbaum Associates.

Castro, S. L., Delgado Martins, M. R., Gomes, I., Amorim, E. & Pimenta, F. (1996) Training language skills in pre-school children with computer word games. In *Actas do IV congresso português de engenharia biomédica, BIOENG'96* (B. S. Santos, J. A. Rafael, A. M. Tomé & F. Vaz, editors), pp. III.6.1-III.6.5. Aveiro: Sociedade Portuguesa de Engenharia Biomédica.

Castro, S. L., & Gomes, I. (2000). *Dificuldades de aprendizagem da língua materna*. Lisboa: Universidade Aberta. (pp. 227-228)

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.

Cole, R. A. & Perfetti, C. A. (1980) Listening for mispronunciations in a children's story: the use of context by children and adults, *Journal of Verbal Learning and Verbal Behavior*, **19**, 297-315.

Delgado-Martins, M. R., Castro, S. L., Gomes, I., Amorim, E. & Pimenta, F. (1998) Estimular a linguagem e prevenir o insucesso com meios multimédia: o projecto de Viana do Castelo. In *Actas do encontro linguística e educação* (R. V. Castro & M.

L. Sousa, editors), pp. 91-98. Lisboa: Edições Colibri e Associação Portuguesa de Linguística.

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

Garlock, V. M., Walley, A. C. & Metsala, J. L. (2001) Age-of-acquisition, word frequency and neighborhood density effects on spoken word recognition by children and adults, *Journal of Memory and Language*, **45**, 468-492.

Gomes, I. (2001). *Ler e escrever em Português Europeu.* Unpublished PhD thesis, Universidade do Porto, Porto.

Gomes, I. & Castro, S. L. (2003) Porlex, a lexical database in European Portuguese, *Psychologica*, **32**, 91-108.

Le Normand, M. T. & Chevrie-Muller, C. (1989) Exploration de la production lexicale chez six enfants dysphasiques, *Réducation Orthophonique*, **159**, 345-361.

Logan, J. S. (1992) A computational analysis of young children's lexicons, *Research on Speech Perception, Report No. 8*. Bloomington, Indiana: Department of Psychology, Speech Research Laboratory.

Luce, P. A. & Pisoni, D. B. (1998) Recognizing spoken words: the neighborhood activation model, *Ear & Hearing*, **19**, 1-36.

MacWhinney, B. (1995) *The CHILDES project: tools for analysing talk*. New Jersey: Lawrence Erlbaum Associates.

MacWhinney, B. & Snow, C. (1990) The Child Language Data Exchange System: an update, *Journal of Child Language*, **17**, 457-472.

Mayer, M. (1969) *Frog where are you?* New York: Dial Press.

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 *Word recognition in beginning literacy* (J. L. Metsala & L. C. Ehri, editors), pp. 89-120. Mahwah, NJ: Erlbaum.

Ramos-Pereira, D. (1992). *A linguagem dirigida à criança em fases iniciais da aquisição do Português Europeu como língua materna: aspectos lexicais e enunciativos*. Unpublished Masters thesis, Universidade de Lisboa, Lisboa.

Reznick, J.S. & Goldfield, B.A. (1992) Rapid change in lexical development in comprehension and production, *Developmental Psychology*, **28**, 406-413.

Storkel, H. L. (2002) Restructuring of similarity neighborhoods in the developing mental lexicon, *Journal of Child Language*, **29**, 251-274.

Treiman, R. & Baron, J. (1981) Segmental analysis ability: development and relation to reading ability. In *Reading research: advances in theory and practice* (G. E. MacKinnon & T. G. Waller, editors), Vol. 3 pp. 159-197. New York: Academic Press.

Treiman, R. & Breaux, M. (1982) Common phoneme and overall similarity relations among spoken syllables: their use by children and adults, *Journal of Psycholinguistic Research*, **11**, 569-598.

Vicente, S. (2002). *Reconhecimento de palavras faladas: abordagem desenvolvimental em Português Europeu*. Unpublished PhD thesis, Universidade do Porto, Porto.

Walley, A. C. (1987) Young children's detections of word-initial and final mispronunciations in constrained and unconstrained contexts, *Cognitive Development*, **2**, 145--167.

Walley, A. C. (1993) The role of vocabulary growth in children's spoken word recognition and segmentation ability, *Developmental Review*, **13**, 286-350.

Walley, A. C. & Metsala (1990) The growth of lexical constraints on spoken word recognition, *Perception & Psychophysics*, **47**, 267-280.

Walley, A. C., Metsala, J. L. & Garlock V. M. (2003) Spoken vocabulary growth: its role in the development of phoneme awareness and early reading ability, *Reading and Writing*, **16**, 5-20.

Walley, A. C., Smith, L. B. & Jusczyk, P. W. (1986) The role of phonemes and syllables in the perceived similarity of speech sounds for children, *Memory & Cognition*, **14**, 220-229.

Wepman, J. M. & Hass, W. (1969) *A spoken word count: children-ages 5, 6, and 7*. Chicago: Language Research Associates.

Selene Vicente
Universidade do Porto
Portugal
svicente@psi.up.pt

São Luís Castro
Universidade do Porto
Portugal
slcastro@psi.up.pt

Amanda Walley
Univeristy of Alabama
at Birmingham
awalley@uab.edu

Appendix A

Examples of entries, ordered by grammatical class, in the 3-year-old database of the Viana lexicon (*N* = 694)

#	Orthography	Phonetic Transcription[a]	Grammatical class[b]	Frequency
82	boca	`bo.kA	no	2
424	menino	m6`ni.nu	no	265
463	obrigado	u.Bri`Ga.Du	aj	11
24	alto	`a9.tu	aj	1
29	andar	1`dar	vb	61
186	comer	Ku`mer	vb	188
1	abaixo	A`Baj.Su	av	14
170	cinco	`s3.ku	nu	3

Note. [a]Phonetic transcriptions were imported from the Porlex database (Gomes, 2001; Gomes & Castro, 2003), and used the Unibet system (Castro & Gomes, 2000; 6 = ∂, 1 = ɐ, 3 = ĩ).
[b]no = noun, aj = adjective, vb = verb, av = adverb, nu = numeral.

Appendix B

Number (and percentage) of words per word length in phonemes in the Viana, Ramos-Pereira (R-P), and Porlex lexicons. The peak for each distribution is indicated in bold.

# Phoneme	*Databases*				
	3-years (*N* = 694)	4-years (*N* = 995)	5-years (*N* = 1092)	R-P (*N* = 1374)	Porlex (*N* = 27,063)
1	–	–	–	1 *(0.1)*	5 *(0.0)*
2	12 *(1.7)*	14 *(1.4)*	16 *(1.5)*	22 *(1.6)*	50 *(0.2)*
3	41 *(5.9)*	48 *(4.8)*	54 *(4.9)*	59 *(4.3)*	192 *(0.7)*
4	156 *(22.5)*	208 *(20.9)*	229 *(21.0)*	247 *(18.0)*	1248 *(4.6)*
5	**184 *(26.5)***	**247 *(24.8)***	**264 *(24.2)***	291 *(21.2)*	2329 *(8.6)*
6	145 *(20.9)*	211 *(21.2)*	228 *(20.9)*	**304 *(22.1)***	3750 *(13.9)*
7	85 *(12.2)*	137 *(13.8)*	155 *(14.2)*	208 *(15.1)*	**4419 *(16.3)***
8	38 *(5.5)*	75 *(7.5)*	87 *(8.0)*	149 *(10.8)*	**4429 *(16.4)***
9	17 *(2.4)*	32 *(3.2)*	31 *(2.8)*	62 *(4.5)*	3841 *(14.2)*
10	9 *(1.3)*	14 *(1.4)*	18 *(1.6)*	20 *(1.5)*	2783 *(10.3)*
11	5 *(0.7)*	6 *(0.6)*	6 *(0.5)*	6 *(0.4)*	1769 *(6.5)*
12	1 *(0.1)*	2 *(0.2)*	3 *(0.3)*	5 *(0.4)*	1091 *(4.0)*
13-20	1 *(0.1)*	1 *(0.1)*	1 *(0.1)*	–	1157 *(4.3)*

Note. In the three databases of the Viana lexicon, *N* = 694, 995 and 1091, respectively for each age group. In the R-P database *N* = 1,374 and in Porlex *N* = 27,063. The peak in each distribution is indicated by the characters in bold.

Appendix C

Examples of words (sun, key, bench, room, face, sand, window, chair, to write), and neighbors, computed in the 3- and 5-year-old databases of the Viana lexicon, as well as in the Porlex adult database. The number of neighbors is also presented.

Word	Transcription[a]	3-years-old	5-years-old	Adult
sol	SO9	1; sO	1; sO	4; + RO9, sa9, su9
chave	Sav6	0	1; kav6	3; + av6, nav6
banco	b1ku	1; br1ku	4; + b1kA, biku, m1ku	9; + b1bu, b1du, b1Zu, b1zu, beku
sala	salA	2; falA, malA	3; + sakA	14; + alA, balA, sElA, silA, galA, palA, saGA, sajA, sOlA, talA, valA
cara	karA	3; kazA, kaBrA, karGA	4; + kartA	19; + arA, kakA, kasA, kaLA, kapA, karDA, karu, karpA, katA, kavA, klarA, kOrA, kurA, tarA, varA
areia	arAjA	0	0	4; AmAjA, ArANA, AvAjA, tArAjA
janela	ZAnElA	1; pAnElA	1; pAnElA	2; + kAnElA
cadeira	kADAjrA	1; mADAjrA	1; mADAjrA	6; + lADAjrA, kArAjrA, kADAjA, kASAjrA, KAjAjrA
escrever	6Skr6ver	0	0	2; d6Skr6ver, 3Skr6ver

Note. [a] Phonetic transcriptions used the Unibet system for European Portuguese (Castro & Gomes, 2001; 6 =∂, S = ʃ, Z = z, L = λ).

Similarity neighborhoods computed in the lemma databases of the Viana lexicon are available in Vicente (2002). Results from developmental comparisons with the adult are available upon request.