RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Sound Similarity Relations in the Mental Lexicon: Modeling the Lexicon as a Complex Network[1]

Vsevolod Kapatsinski

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

---

[1] Preparation of this chapter was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012). Many thanks to Luis Hernandez for designing the program used to create neighbor pairs from a list of words, as well as to Katy Börner, Tom Gruenenfelder, and David Pisoni for helpful feedback on this project.

Sound Similarity Relations in the Mental Lexicon: Modeling the Lexicon as a Complex Network

Abstract. The standard definition of neighborhood density defines two words to be neighbors if they differ by one and only one segment (Luce & Pisoni, 1998). This definition assumes that the length of the shared part is irrelevant to sound similarity. However, confusability of non-linguistic sound sequences (Fallon Coble & Robinson, 1992; Kidd & Watson, 1992) as well as judgments of sound similarity between spoken pseudowords (Kapatsinski, in press, b) depend on the proportion of total duration of the word or sound sequence that is mismatched, and not on the absolute duration of the mismatch. To bring the definition of neighborhood into alignment with these results, this paper defines words to be neighbors if they share at least two thirds of their total duration, measured in segments. This simple change reduces the proportion of words with no neighbors to all words in the lexicon from over 58 percent to just 7 percent, increasing the applicability of the Neighborhood Activation Model (Luce & Pisoni, 1998). Lexical decision, naming, and familiarity judgment data indicate that speaker/hearers are sensitive to the more distant neighbors brought in by the new neighborhood definition. The large-scale properties of the network are discussed and future directions are indicated.

Introduction

The perceptual similarity relations of spoken words have attracted the attention of researchers since the seminal study by Greenberg and Jenkins (1964). Examination of sound similarity is fundamental for studying the structure of the mental lexicon and the nature of the linguistic system.

For instance, in the debate on whether rules are required to explain linguistic productivity, proponents of the rule-based models have claimed that in each inflectional domain there is one default, rule-based morpheme that is more productive than its competitors and bears little regard to how similar a novel word is to existing ones (Pinker & Prince, 1988). Kapatsinski (2005a, b, in press a) has demonstrated that there are dissociations between productivity and sensitivity to similarity effects, a conclusion that crucially depends on a psychologically real measure of phonological similarity.

A psychologically real measure of phonological similarity is also necessary for creating successful models of analogical extension of linguistic patterns, as in morphological productivity or the lexical diffusion of sound change. For instance, Albright and Hayes (2003) have shown that a model that weighs mismatches in segments adjacent to the affix whose behavior is to be predicted more heavily than more distant mismatches outperforms a model that weighs mismatches in all positions equally in predicting the past tenses of novel verbs produced by native English speakers.

Phonological similarity interacts with word recognition. For instance, Marslen-Wilson's (1990) Cohort Model of word recognition predicts that initial mismatches should lead to more between-word inhibition than later mismatches, as found by Radeau et al. (1995).

Finally, phonological similarity has been argued to influence the direction of sound change and phonological alternations. Steriade (2001) analyzed regressive and progressive place assimilation in CC clusters. She hypothesized that if /anpa/ is perceived to be more similar to /ampa/ than to /anta/, it will be realized as /ampa/. Fujimura et al. (1978) found that when a pause is preceded by transitions indicating one consonant and followed by transitions indicating another, listeners make the judgment of what the consonant is based on the CV transition, not the VC one in both English and Japanese. On the other hand, retroflection is primarily determined by VC transitions (Ladefoged & Maddieson 1986). Thus, assimilation in retroflection should affect C_2. This is precisely what has been found by Steriade (2001) in a cross-linguistic study. Steriade (2004) argued that phonological similarity can also influence loanword adaptation. For instance, in deciding to simplify a CVC_1C_2 input as CVC_2 the speaker judges C_1 to be less perceptible than C_2 and thus that C_1C_2 is more similar to C_2 than to C_1, and that C_1C_2 is more similar to C_2 than to C_1VC_2.

Recently, researchers have begun to explore the large-scale structure of the phonological mental lexicon using graph-theoretic tools. Vitevitch (2004) and Gruenenfelder and Pisoni (2005) modeled the lexicon as a network in which nodes are words and where two words are connected to each other if they differ only by the addition, deletion or substitution of one segment. This "one phoneme deletion, addition, substitution metric" has been the standard criterion used to determine whether or not two words are lexical neighbors (Luce & Pisoni, 1998).

Unfortunately, this metric has limited applicability since even among the 20000 most common English words more than half have no neighbors (Gruenenfelder & Pisoni, 2005). In addition, the metric contradicts results of confusability studies that have found that the confusability of two sounds depends on the *proportion* of the total duration that is mismatched and not on the absolute duration of the mismatched parts (Fallon Coble & Robinson, 1992; Kidd & Watson, 1992). It is also at odds with the finding that judged sound similarity of two words depends on how many segments they share as well as on how many segments they differ by (Kapatsinski, 2005b, in press b).

The aim of the present study was to examine the large-scale structure of the mental lexicon using a more psychologically plausible definition of neighbors based on the "proportion-of-total-duration rule" derived from confusability studies (Kidd & Watson, 1992). We define a word B to be a neighbor of word A if and only if it shares at least two thirds of A's segments. That is, if A is six segments long its neighbors can be derived from it by at most two phoneme changes (deletions additions, or substitutions), while if A is nine segments long its neighbors can differ from it by at most three segments. Under this metric, the proportion of hermit words (i.e., words with no neighbors) decreased from 58% of the lexicon to 7%. We show that the new metric outperforms the old metric in modeling reaction times and accuracy in the lexical decision and visual word naming tasks as well as in predicting familiarity judgments. Finally, we discuss the present limitations of the metric and ways to further improve and test it.

Another aim of this paper is to help resolve the debate on whether the mental lexicon has scale-free structure. Vitevitch (2004) has claimed that the histogram of number of neighbors per word (i.e., the degree distribution) follows a power law and thus, the lexicon is a scale-free network. Gruenenfelder and Pisoni (2005) have argued that this result is simply due to the relationship between length and the number of neighbors, which is an artifact of the one-phoneme deletion, addition, substitution metric. They have examined the set of monosyllabic words and found that the degree distribution did not follow a power law. In fact, they argued that it resembles much more a Poisson distribution. By eliminating the length bias with the new metric, we show that longer words still tend to have fewer neighbors. In addition, through fitting a number of curves to the data, we find that the best fit to the lexicon's degree distribution is provided by an exponential equation rather than a power law. While the power law accounts for 83% of the variance (85% under the old metric, Gruenenfelder & Pisoni, 2005), the exponential distribution accounts for 97% of the variance.

Finally, we will argue that neighborhood density should not be modeled as the degree of a word, i.e., the number of links connecting the word to other words, but rather as the sum of strengths of those links.

Methods

In this paper, we will be modeling the lexicon as a network in which words are nodes. A link is drawn from node A to node B if at least 2/3 of the segments that occur in the word represented by A also occur in the word represented by B.

The database analyzed was the Hoosier Mental Lexicon (Nusbaum et al., 1984). The phonologically transcribed form of each word in the lexicon was subjected to the new metric. The resulting set of nodes and links, which excluded nodes that had no links, was analyzed using Pajek (Batagelj & Mrvar, 2003). Random networks used for comparison with the actual networks were created using the Erdos-Renyi method in Pajek and had the same number of nodes and links as the actual networks. They differed from the actual networks in that nodes were connected randomly rather than based on the similarity metric.

The network created was "directed" because a long word can have a short word as a neighbor without the short word having the long word as a neighbor. For instance, 'moat' is a neighbor of 'demote' but 'demote' is not a neighbor or 'moat'. This is because 'demote' differs from 'moat' by two segments, which is 1/3 of the duration of 'demote' but more than 1/3 of the duration of 'moat.' The reason the ratio of 1/3 was used is because the traditional metric has performed well in predicting reaction times and familiarity ratings in experiments that mostly used CVC words, which consist of 3 segments (Luce & Pisoni, 1998). Ideally, the **ratio** of mismatch to total length should be derived empirically for each task by seeing which ratio provides the best fit to the data, which is a direction for future research.

For word lengths that are not divisible by three, rounding was used. Thus the maximum number of segments that neighbors could differ by was 1 for 2-segment, 3-segment, and 4-segment words, 2 for 5-, 6-, and 7-segment words, and 3 for 8-, 9-, and 10-segment words. Words that were longer than 10 segments (n=955) or shorter than 2 segments (n=5) were excluded from being heads of links. That is, they could be neighbors of other words but could not have neighbors themselves or, in graph-theoretic terms, all such words have an **output degree** (our operational definition of **neighborhood density**) of 0 and thus are only included in the network if they are pointed to by other words. We will call the word for which we are searching for neighbors the **base word** from now on**.** The reason we are using the output degree, that is, number of links pointing from the base word to other words, rather than input degree (i.e., number of links pointing to the base word) as a measure of neighborhood density is that we take neighborhood density of a word to be the number of words activated by the base word when the base word is presented. That is, we take neighborhood density to be a postlexical variable, which is in agreement with findings that neighborhood density starts to affect processing at a stage indexed by a later ERP component than sublexical variables, such as phonotactic probability (Pylkkanen et al., 2002). The total number of base words (size calculated using MonoConc Pro) was 18360.

Figure 1 shows the unfortunate side effect of rounding up the neighborhood radius for words that are not divisible by three. There is a large jump in average degree as the radius increases from 1 to 2 and then from 2 to 3 segments. There is no evidence that these bumps in the distribution are psychologically real.

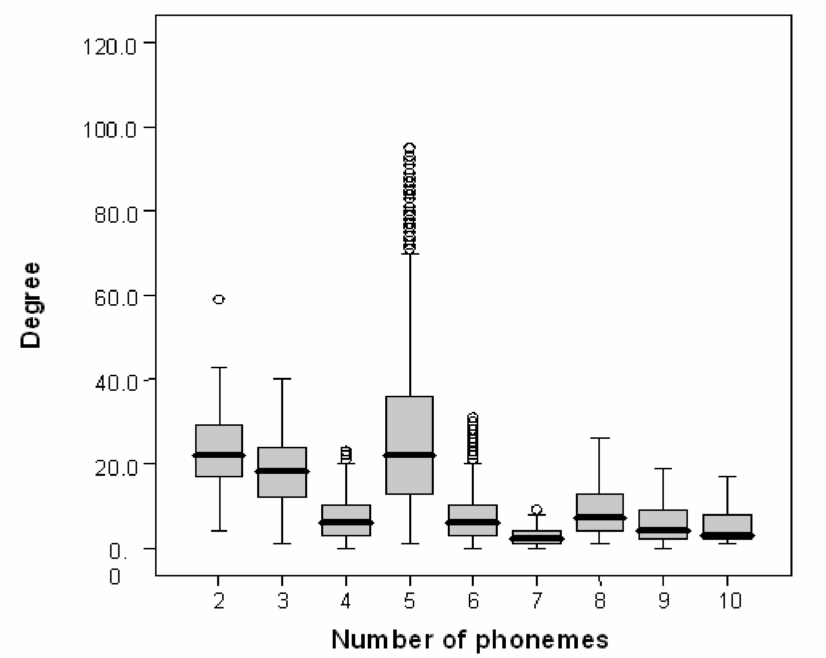

Figure 1: The relationship between base word length in segments and output degree (neighborhood density).[2]

One way to deal with this issue would be to make sure that the number of links between two words always represents the ratio of shared segments to the length of the base word. That means that the maximum number of links between two words would have to be divisible by all possible numbers of segments a base word may have, or, in our case, 2-10. This turns out to be 2520 with two-segment base words being connected to their two-segment neighbors by 1260 links pointing in each direction and ten-segment words being connected to neighbors that differ from them by 3 segments by 252 links. Clearly, this solution is extremely computationally expensive.

The alternative is to use link weights instead of numbers of links to express connection strength. With this approach, the degree of a node does not reflect how strongly it is connected to other nodes. Thus, degree stops being a theoretically justified predictor of behavioral and electrophysiological data. Rather, behavioral and electrophysiological dependent variables should be influenced by the sum of strengths of all links the node has. Future work should examine the characteristics of the link-strength-sum distribution in addition to the degree distribution for the lexicon.

We have used the English Lexicon Project (Balota et al., 2002), a repository of reaction times from the lexical decision task and the naming task for over 40,000 words collected from 1200 subjects, each of whom responded to all the words. The overlap between the 18,360 base words used for network creation and the English Lexicon Project consisted of 13,458 words. These are the words for which we have neighborhood density estimates as well as lexical decision and naming reaction times and familiarity ratings.

---

[2] The middle line on the rectangle indicates mean output degree, rectangle sides index the 25th and 75th percentile. Vertical lines indicate cases within 3 lengths of box length from the upper or lower edge of the box. Points indicate cases with values more than three box lengths removed from the nearest edge of the box.

Perhaps the biggest drawback of the English Lexicon Project for testing the influence of neighborhood density is that the words were presented to subjects visually. Facilitatory density effects are usually found for visually presented words (Andrews, 1997) while inhibitory effects are found for auditorily presented words (Luce et al., 2000). It is likely that the facilitatory effects found with orthographic presentation are sublexical in nature since high-density words contain high-frequency grapheme chunks whose high frequency can facilitate the orthography-to-phonology mapping (cf. Plaut et al., 1996; Pylkkanen et al., 2002). Therefore, while the new metric is shown to be better at predicting reaction times from the English Lexicon Project, this result should be taken with caution since the neighborhood density effect is facilitatory in this database. A definitive test would come from ERP studies where sublexical and lexical effects can be disambiguated and studies using auditorily presented stimuli, which eliminate the extra processing stage involved in converting orthographic representations into phonological ones.

Results

Small-world Characteristics

A network is considered to have small-world characteristics if it has short average path length and diameter and a clustering coefficient that is orders of magnitude higher than that of a random network with the same number of links and nodes (Watts & Strogatz, 1998). The clustering coefficient CC1 is defined as the proportion of a node's neighbors that are also neighbors of each other. CC2 is the proportion that links between neighbors of a word form out of all links the word's neighbors have. Table 1 shows that the lexicon is characterized by very high clustering but also by relatively long average path length and diameter. That is, like a small-world network, the lexicon contains neighborhoods in which all words are densely interconnected and between which the connectivity is lower. However, the between-neighborhood links that would allow access from a node in one neighborhood to a node in another neighborhood are harder to find than in a small-world network. Importantly, high clustering does not depend on the inclusion of morphologically complex words, although the exclusion of morphologically complex words does decrease the lexicon's clustering coefficient relative to a random network. Thus, morphology increases clustering but is not exclusively responsible for it. Interestingly, mean path length for the entire lexicon does not decrease noticeably from its value with the old metric (6.08 in Vitevitch 2004, 6.06 here), despite a large increase in the network's size.

Table 1. Small-world properties (actual data compared to random nets with the same number of nodes and links/node created using the Erdos-Renyi method in Pajek)

	Entire lexicon		**2-4 phoneme words**		**5-7 phoneme words**		**8-10 phoneme words**		**Monomorphemic words**	
	Real	**Rand.**	**Real**	**Rand.**	**Real**	**Rand.**	**Real**	**Rand.**	**Real**	**Rand.**
Average distance	6.06	4.30	4.86	4.08	5.46	4.74	7.79	6.24	5.27	3.88
Diameter	20	7	15	7	16	8	26	13	21	7
CC1	.235	.0007	.262	.002	.208	.0006	.173	.0006	.253	.0016
CC2	.040	.0005	.039	.001	.030	.0005	.047	.0005	.040	.0009

Figure 2 shows the entire lexicon. As we can see, the lexicon is a disconnected graph, which has a giant component comprising the vast majority of words in the lexicon. The visualization is derived using the Fruchterman-Reingold algorithm, which separates the giant component from the rest of the network and arranges the nodes in such a way that distance in terms of number of links corresponds to physical distance on the graph.

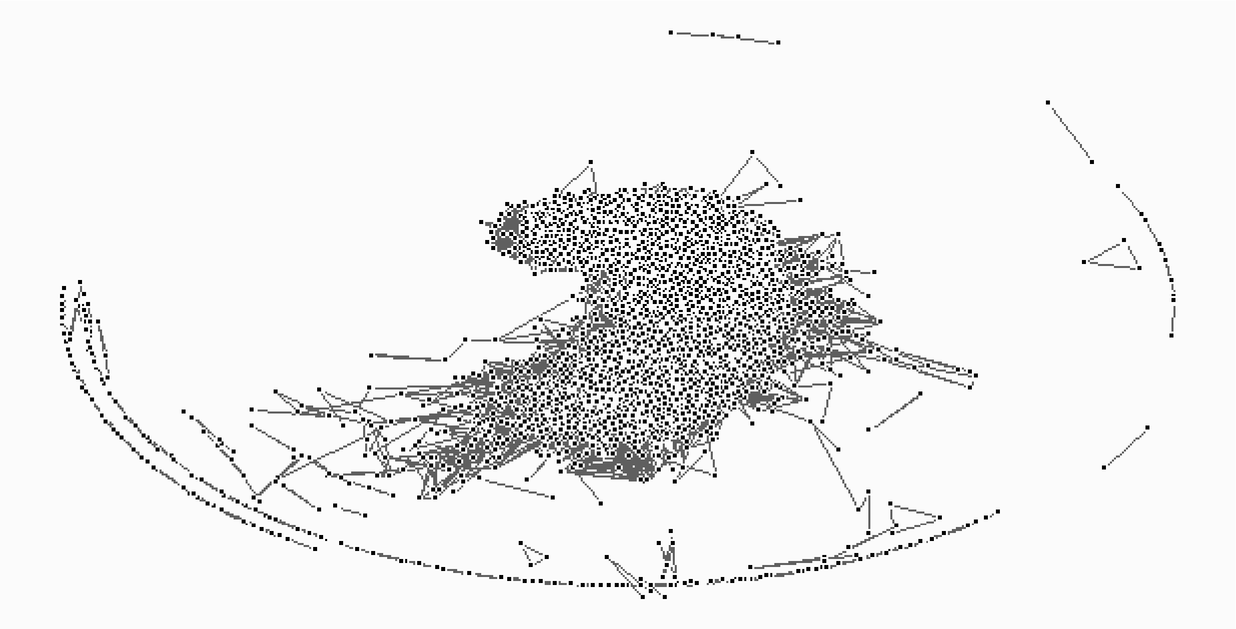

Figure 2. The complete lexicon. Dots are words, lines are connections. Two dots are connected if they share at least 2/3 of their segments.

The giant component is elongated, showing that some nodes can only reach each other through a long chain of intermediaries, thus leading to a relatively long mean path length. Interestingly, this elongation is obtained only if all words are included. It is not obtained with only monomorphemic words, or with words that are 2-4, 5-7, or 8-10 phonemes long. Thus, the elongation seems to result from the fact that the longest words in the complete lexicon cannot be connected to the shortest words: words of particular lengths form distinct connected regions of the giant component, such that the length of the words increases as one proceeds from the top of the giant component to its bottom. This hypothesis is confirmed by the finding that mean path length is larger relative to the corresponding random net for the entire lexicon than for any of its word-length-limited subsets (1.41 vs. 1.2, 1.15, 1.25).

In fact, the lexicon tends to fall apart into separate clusters when nodes with low degrees are eliminated. When nodes with degrees below 15 are eliminated, there is a noticeable bottleneck separating long words (more than 7 phonemes long) from shorter words. This bottleneck, which connects long words to short words, is formed entirely of /ʃən/-final words. Of the 2228 nodes with degrees above 25, the elimination of 4 nodes - ‘coalition’ (degree=40) or ‘colon’ (degree=43), ‘passion’ (degree=41), ‘nation’ (degree=39), and ‘fixation’ (degree=40) or ‘fission’ (degree=45) - would render the network disconnected. Of the six, only 2 (‘passion’ and ‘nation’) have more than one link to each mega-neighborhood. Thus, in this network, it appears that if one wants to render the network disconnected, it is not the biggest hubs (like 'pastor' with degree=112) that need to be taken out. Figure 3 shows a visualization of the lexicon with all nodes with output degrees below 40, i.e., all words with fewer than 40 neighbors, eliminated. The figure shows that the lexicon 'falls apart', showing that high-degree nodes fall into two sparsely-connected giant neighborhoods. The lower neighborhood comprises 8-10 segment words while the upper component contains shorter words with 24 segment words confined to the left side of the upper neighborhood.

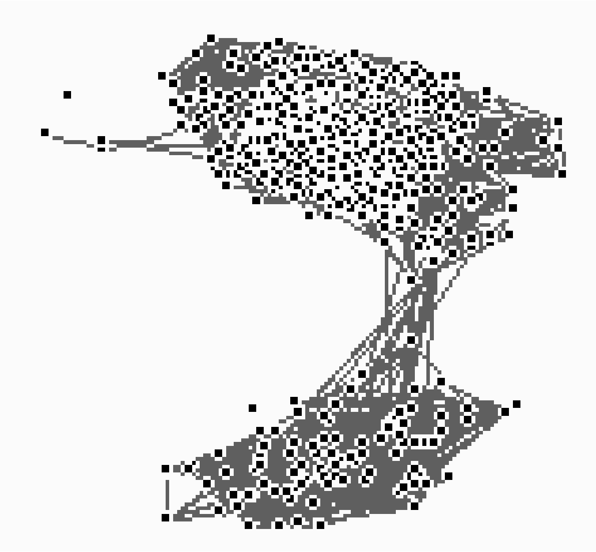

Figure 3. The network comprising all words with more than 40 neighbors.

Table 2 shows how the various reductions of the lexicon compare to the corresponding random networks. The table shows that as nodes with lower output degrees are introduced to the network, the network does not become less connected. That is, the low degree nodes do not just attach themselves to the outskirts of the network but also form shortcuts between high-degree nodes. Interestingly as nodes with lower degree are introduced, the average degree grows (until minimum degree reaches around 20), which indicates that the lower-degree nodes connect to many of the high-degree nodes, rather than forming long chains of low-degree nodes that must be traversed to reach a high-degree node from a randomly chosen low-degree node. In some cases, this even increases the connectivity of the network, e.g., when nodes with degrees between 35 and 39 are introduced, mean path length remains constant and diameter shrinks despite an increase in the size of the network. The same occurs when nodes with degrees between 15 and 19 are introduced.

In a small-world network, the growth of diameter and mean path length with the introduction of low-degree nodes would likely be steeper than in a random net because in such a network high-degree nodes are the ones that are more likely to have connections linking different neighborhoods. Traveling from a randomly chosen node down a randomly chosen link one is more likely to end up in a high-degree node than in a low-degree node. If the high-degree nodes are also more likely to provide a link to another neighborhood, the average path length between the neighborhoods is shortened. It makes functional sense for the lexicon to consist of poorly connected neighborhoods, since one would not want spreading activation to easily activate or inhibit neighbors of the stimulus's neighbors that are not neighbors of the stimulus and are therefore not at all similar to the stimulus.

Table 2. Elimination of low-degree nodes does not lead to decrease in average path length.

Minimum output degree	Mean path length relative to random net	Diameter relative to random net	Mean degree
40[3]	3.86/2.68=**1.44**	13/4=**3.3**	**16.21**
35	3.86/2.77=**1.39**	10/4=**2.5**	**16.68**
30	3.91/2.89=**1.35**	11/5=**2.2**	**16.98**
25	4.28/2.95=**1.45**	14/4=**3.5**	**17.73**
20	4.80/3.05=**1.57**	16/5=**3.2**	**18.04**
15	4.80/3.21=**1.50**	14/5=**2.8**	**18.02**
10	5.02/3.43=**1.46**	20/5=**4.0**	**17.38**
5	5.33/3.72=**1.43**	17/6=**2.8**	**15.22**
2	5.78/4.04=**1.43**	19/7=**2.7**	**12.77**
1	5.98/4.23=**1.41**	20/7=**2.9**	**11.73**
0	6.06/4.30=**1.41**	20/7=**2.9**	**11.33**

Low mean path lengths occur in networks which need to be traversed quickly. One purpose of such traversal is search. In the lexicon, on the other hand, search is unlikely to occur by traversing links between distantly located nodes. Rather, search in the lexicon involves activation of structured neighborhoods of words that all share a single sublexical chunk, the chunk consistent with the acoustic evidence at that point in word recognition (Marslen-Wilson, 1990). Below we will see that structured neighborhoods are a major feature of the lexicon.

Degree Distribution

Figures 4-8 show the degree distributions for the entire lexicon and its subsets modeled as a network in which two words are connected by only one link regardless of the strength of the connection. The trendlines, fitted in Microsoft Excel, show that the exponential distribution provides a much better fit to the data than the power-law-based one.

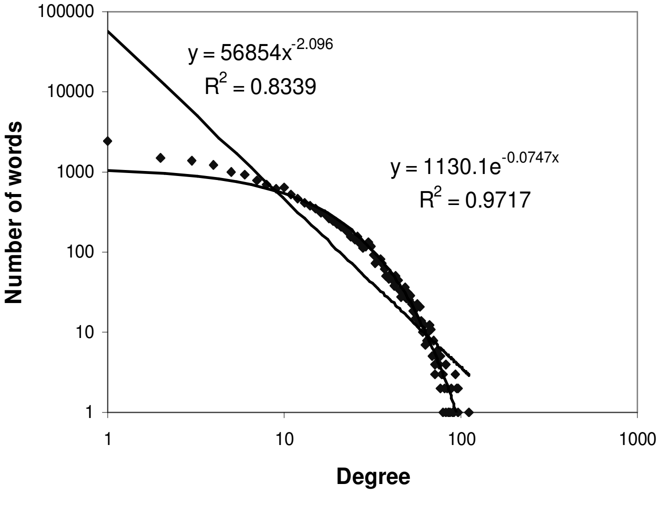

Figure 4. Degree distribution for the entire lexicon modeled with a power law and an exponential distribution.

---

[3] This was the highest degree for which mean path length and diameter measures were collected because the network no longer has a giant component at higher degrees.

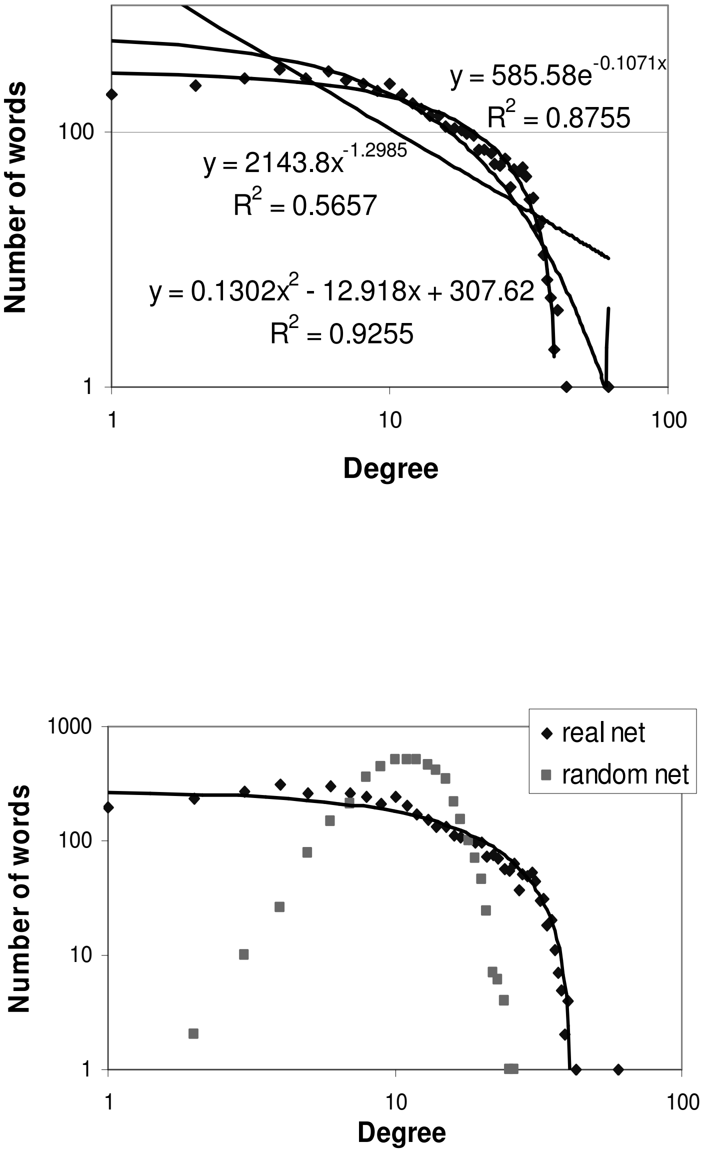

Figure 5. Degree distribution for words 2, 3, and 4 segments long with power law, exponential and parabolic models and compared to the Poisson degree distribution of a random network.

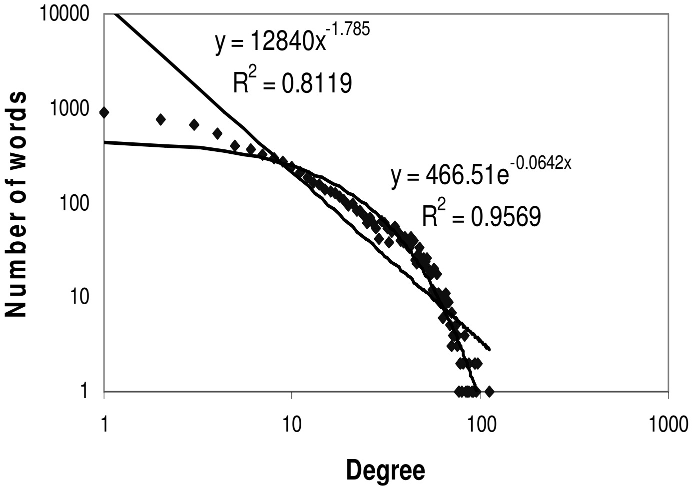

Figure 6. Degree distribution for words 5, 6, and 7 segments long with power law and exponential models.[4]

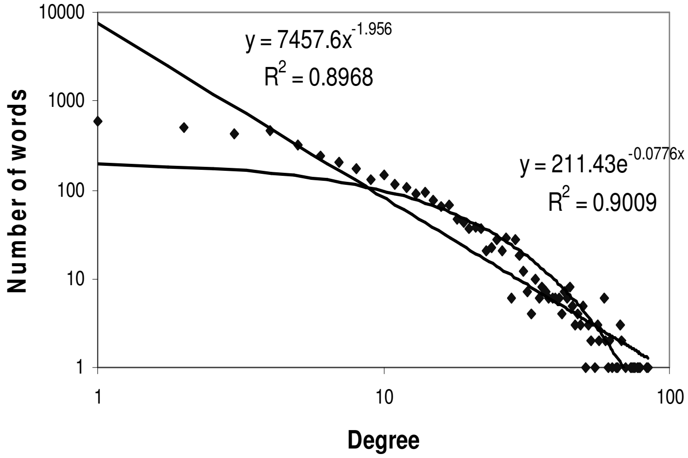

Figure 7. Degree distribution for words 8, 9, and 10 segments long with power law and exponential models.

---

[4] The reason the graphs do not show a random net distribution is that there is no lowering in the left-hand tail of the observed distribution, making a Poisson fit highly inappropriate.

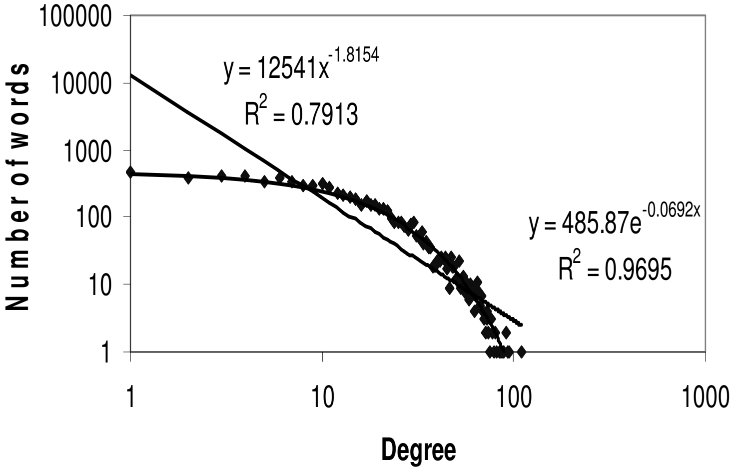

Figure 8. Degree distribution for the monomorphemic lexicon modeled with a power law and an exponential distribution.

To summarize, the degree distributions for the entire lexicon, the monomorphemic lexicon, and the set of words that are 5, 6, or 7 segments long are best approximated by an exponential curve, rather than a power law. The degree distribution of words that are 2, 3, or 4 segments long is best approximated by a parabola and not a Poisson distribution. The degree distribution of 8-, 9-, and 10-segment words is approximated equally well by an exponential curve and a power law. Where the curves diverge most is when predicting the number of words with very low degrees: the power law systematically overpredicts while the exponential distribution underpredicts. If the data for the entire lexicon are fit to a power law, the exponent is 2.1.

Perhaps, the most important feature of the new metric, however, is that the marked reduction achieved in the number of words with no neighbors (hermits), shown in Table 3.

Table 3. Number of hermits under old and new metrics.

Length of word	Percent hermits under old metric	Percent hermits under new metric
2-4 segments	2.4%	2.4%
5-7 segments	67.6%	8.6%
8-10 segments	92.7%	10.4%
Whole lexicon	58.1%	7.3%
Monomorphemic lexicon	33.2%	7.0%

Relations between Old Density, New Density, and Other Independent Variables[5]

Table 4 shows that new density correlates with other independent variables more highly than does old density. Thus to show that new density is a better predictor of behavior than old density it will not be sufficient to show that new density shows better correlation with behavioral dependent variables. Rather, we will need to show that it does better even when the other partially correlated variables are competing against density in a regression. Old density correlates with new density at r=.623. We are going to concentrate on modeling reactions to monomorphemic words to avoid confounding phonological and morphological links.

Table 4. Correlations between old density, new density, and other lexical variables for all monomorphemic words that are longer than four phonemes (n=4146).

		Number of syllables	**Number of letters**	**Number of phonemes**	**Mean phoneme frequency**	**Mean bigraph frequency**	**Log word frequency**
Old density	Pearson r	**-.373**	**-.339**	**-.387**	**-.029**	**.069**	**.103**
	Sig. (2-tailed)	.000	.000	.000	.061	.000	.000
New density	Pearson r	**-.490**	**-.429**	**-.499**	**.118**	**.115**	**.150**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000

Modeling Behavioral Data

Table 5 shows that new density demonstrates stronger correlations with behavioral variables than does old density.

Table 5. Old density vs. new density as correlates of behavioral variables for all monomorphemic words that are longer than four phonemes. The reason shorter words were excluded is because the new metric and the old metric make identical predictions for those words.

		Naming accuracy	**LDT accuracy**	**Naming RT**	**LDT RT**	**Familiarity**
Old density	Pearson r	**.142**	**.114**	**-.245**	**-.204**	**.136**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
New density	Pearson r	**.181**	**.151**	**-.310**	**-.273**	**.182**
	Sig. (2-tailed)	.000	.000	.000	.000	.000

Table 6 shows that when old density and the difference between old and new density are entered into a regression, the difference between old and new density makes a significant contribution

---

[5] All results reported in this section and the next one were derived with the network in which all relations are represented by a single link.

to predicting behavior. Thus subjects are sensitive to the word's distant neighbors brought in by the new metric but not by the old metric. Another interesting aspect of these data is that the clustering coefficient CC2 and hubness, which are measures that take into account characteristics of the neighborhoods of the stimulus's neighbors, are significant only when multimorphemic words are included. That suggests that words that are neighbors of neighbors of the stimulus play a role in the stimulus's processing only to the extent that they share morphemes with the stimulus or are neighbors of the stimulus's root.

Table 6. Old density vs. new density as predictors of naming and lexical decision accuracy and reaction time and familiarity judgments. The top row in each cell shows results for all word that are more than four phonemes long (n=10732). The bottom row shows results for monomorphemic words that are more than four phonemes long (n==4146). Significant effects are in bold.

Predictor	**Naming accuracy**		**Lexical decision accuracy**		**Naming time**		**Lexical decision time**		**Familiarity rating**	
	Coeff.	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**
Old density	.018 .034	.207 .063	.020 **.033**	.123 **.050**	-.021 **-.031**	.081 **.048**	-.005 -.026	.689 .103	.017 .023	.145 .121
New density minus old	**.030** **.082**	**.021** **.000**	**.051** **.072**	**.000** **.000**	**-.038** **-.093**	**.001** **.000**	**-.050** **-.076**	**.000** **.000**	**.046** **.073**	**.000** **.000**
Log word frequency	**.424** **.476**	**.000** **.000**	**.585** **.625**	**.000** **.000**	**-.472** **-.499**	**.000** **.000**	**-.541** **-.567**	**.000** **.000**	**.595** **.645**	**.000** **.000**
Mean phoneme frequency	**-.051** **-.061**	**.000** **.000**	**-.033** **-.034**	**.001** **.023**	**.114** **.127**	**.000** **.000**	**.042** **.043**	**.000** **.003**	**-.051** **-.057**	**.000** **.000**
Mean bigraph frequency	-.004 -.019	.693 .244	-.002 .009	.834 .551	-.006 -.023	.465 .091	**.018** .015	**.034** .287	.012 **.032**	.163 **.012**
Number of phonemes	.023 **.090**	.276 **.001**	.022 **.053**	.248 **.036**	**.073** .008	**.000** .731	**.057** .040	**.001** .098	-.012 .003	.497 .891
Number of letters	**.095** .029	**.000** .236	**.175** **.083**	**.000** **.000**	**.113** **.180**	**.000** **.000**	**.150** **.093**	**.000** **.000**	**.154** **.127**	**.000** **.000**
Number of syllables	**-.319** **-.162**	**.000** **.000**	**-.154** **-.089**	**.000** **.000**	**.257** **.133**	**.000** **.000**	**.169** **.123**	**.000** **.000**	**-.153** **-.114**	**.000** **.000**
Number of morphemes	**.137**	**.000**	**.139**	**.000**	**-.102**	**.000**	**-.050**	**.000**	**.169**	**.000**
CC2[6] (new)	**.045** .027	**.000** .113	**.027** .025	**.006** .107	-.015 -.015	.093 .296	-.017 -.003	.055 .817	.026 **.005**	.004 **.721**
Hubness (new)	-.004 .011	.717 .447	**-.029** .000	**.003** .997	**.025** -.007	**.006** .587	**.022** .011	**.010** .374	**-.040** .003	**.000** .824
r^2	.227 .272		.354 .405		.438 .457		.477 .465		.383 .444	

Table 7 shows that new density can make a significant contribution to successfully predicting subjects' reactions to the words that the old metric considers to be hermits. Table 8 shows that the correlations between new density and subjects' behavior in response to 'hermits' are significant. These results suggest that, contrary to the old metric, these words do not all have the same density. It is important to point out that the old metric cannot predict differences in behavior in response to different hermits even if neighbors are defined as words that are 1 or 2 links away from the stimulus (Gruenenfelder & Pisoni, 2005).

---

[6] The other clustering coefficient, CC1, is not a significant predictor for any dependent variable.

Table 7. Predicting reactions to former hermits. The top row in each cell shows results for all hermits that are more than four phonemes long (n=7795). The bottom row shows results for monomorphemic words that are more than four phonemes long (n==2054). Significant effects are in bold.

Predictor	**Naming accuracy**		**Lexical decision accuracy**		**Naming time**		**Lexical decision time**		**Familiarity rating**	
	Coeff.	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**
New density	.009 **.058**	.520 **.011**	**.036** **.070**	**.005** **.001**	**-.030** **-.076**	**.015** **.000**	**-.028** **-.052**	**.017** **.009**	**.033** **.080**	**.004** **.000**
Log word frequency	**.430** **.507**	**.000** **.000**	**.582** **.650**	**.000** **.000**	**-.495** **-.540**	**.000** **.000**	**-.548** **-.570**	**.000** **.000**	**.586** **.654**	**.000** **.000**
Mean phoneme frequency	**-.059** **-.060**	**.000** **.005**	**-.044** **-.042**	**.000** **.027**	**.116** **.105**	**.000** **.000**	**.050** **.048**	**.000** **.011**	**-.065** **-.065**	**.000** **.000**
Mean bigraph frequency	.008 -.006	.509 .757	.014 .032	.202 .080	-.007 -.019	.492 .303	.015 .011	.150 .559	**.031** **.058**	**.002** **.000**
Number of phonemes	.026 **.099**	.265 **.007**	.013 **.065**	.528 **.046**	**.060** -.022	**.003** .487	**.040** .009	**.038** .782	-.010 .022	.592 .465
Number of letters	**.085** .025	**.000** .435	**.162** **.066**	**.000** **.019**	**.094** **.177**	**.000** **.000**	**.170** **.130**	**.000** **.000**	**.138** **.111**	**.000** **.000**
Number of syllables	**-.277** **-.147**	**.000** **.000**	**-.136** **-.060**	**.000** **.013**	**.261** **.147**	**.000** **.000**	**.160** **.118**	**.000** **.000**	**-.145** **-.101**	**.000** **.000**
Number of morphemes	**.139**	**.000**	**.148**	**.000**	**-.106**	**.000**	**-.059**	**.000**	**.179**	**.000**
CC2 (new)	**.055** .019	**.000** .381	**.031** .029	**.007** .123	**-.024** -.032	**.028** .087	**-.029** -.011	**.006** .541	**.030** -.003	**.005** .880
Hubness (new)	-.001 .009	.911 .656	**-.037** -.005	**.002** .768	**.033** -.004	**.005** .826	**.026** .010	**.019** .547	**-.040** -.010	**.000** .532
r^2	.271 .277		.404 .426		.456 .453		.464 .443		.443 .444	

As seen from Table 8, new density is equally good at predicting behavior to former hermits as to former non-hermits in terms of lexical decision accuracy and judgments of familiarity but is slightly less successful on the hermits with the other dependent variables. Notably, the new density measure shows a better correlation with behavioral measures even when words considered hermits by the old metric are eliminated from comparison.

Table 8. Correlations between new density and behavioral variables for morphologically simple words considered hermits by the old density metric (top row, n=2054) and those words considered non-hermits by the old metric (bottom row, n=1628).

		Naming accuracy	**LDT accuracy**	**Naming RT**	**LDT RT**	**familiarity**
New density	Pearson Correlation	.107	.107	-.225	-.187	.136
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	Pearson Correlation	.162	.105	-.238	-.209	.135
	Sig. (2-tailed)	.000	.000	.000	.000	.000
Old density	Pearson Correlation	.090	.049	-.151	-.096	.075
	Sig. (2-tailed)	.001	.067	.000	.000	.002

Discussion

In this paper, we have proposed a definition of neighborhood that incorporates findings from confusability (Kidd & Watson, 1992) and sound similarity judgment data (Kapatsinski, 2005b, in press b) in that two words are defined to be neighbors if they share a certain proportion of their length, and not some absolute number of segments. We have seen that even under this metric longer words are more likely to have fewer neighbors but that the proportion of words that have no neighbors decreases dramatically. We have also seen that the lexicon has an exponential degree distribution but for very long words a scale-free distribution is indistinguishable from an exponential one in terms of goodness of fit to the data while the distribution for very short words is roughly parabolic due mainly to differences in numbers of words that have very few neighbors. Thus, lexicon structure is not consistent with growth via preferential attachment (Barabasi & Albert, 1999) under the existing neighborhood density metrics.

We have also seen that the lexicon is highly clustered and has a higher mean path length than a random network. Thus, the lexicon is not a small-world network. One factor contributing to relatively high average path length is that while in a random network a node can link to any other node, in the lexicon a word can only be linked to another word with which it shares at least 2/3 of its segments. Therefore, some nodes are guaranteed not to be directly connected.

We have also seen evidence that the more distant neighbors have an effect on how fast the subject can recognize a word. In what follows, I will touch upon the reasons for the high clustering of the lexicon and discuss some limitations of the present metric.

High clustering results naturally if words tend to be similar to many neighbors in the same way. That is, if a word shares some part with many of its neighbors, then the neighbors automatically share the part and are likely to be neighbors with each other. Thus the high clustering of the lexicon results from certain suprasegmental parts being reused more often than others. That is, the lexicon consists of a large number of mega-neighborhoods or “gangs” (Bybee & Moder, 1983), each of which is characterized by what Albright and Hayes (2003) called “structured similarity.” One such gang is the gang of words ending in –ter, which includes the highest-degree words ‘pastor,’ ‘canter,’ ‘caster,’ ‘master,’ etc. Another large gang consists of words starting with str-. The vast majority of long, 8-10-segment words belong to the gang of words ending in -tion, which is subdivided into words ending in -ation and those ending in -ition. If the neighborhood is highly structured, the clustering coefficient CC1 is high because neighbors of the stimulus are very likely to be neighbors of each other: they all share the same parts.

One could imagine a lexicon in which there would not be gangs, as all parts of a word would be equally likely to occur in another word. However, the process of chunking ensures the emergence of these larger units: segments that are used together fuse together through Hebbian learning (Bybee, 2002). Product-oriented generalizations of the type “words (that mean/are X) have Y” are formed (Burzio, 2002; Bybee, 1995) making new words that conform to the generalizations more learnable and leading to analogical change in old words that do not conform to the generalizations. Thus, a rich-get-richer phenomenon occurs at the sublexical level: frequently used units are likely to be used even more frequently in the future.

If no phonotactic constraints on the shape of possible words existed, then under the single-phoneme deletion/addition substitution metric, a word consisting of n segments in a language that has k phonemes could be neighbor to a maximum of n words that are shorter than it by 1 segment, n*k words that are of the same length, and (n+1)*k words that are longer than it by 1 segment. Therefore the old metric would make the prediction that short words should have more neighbors than long words since a short word can link to more long words than a long word can link to short words and the space of possible words is more sparsely populated at the longer word lengths.

As shown in Table 9, the number of neighbors two phonemes away depends on the length of the word more than does the number of neighbors one phoneme away. Table 9 shows that when neighborhood radius, in terms of number of phonemes, is kept constant, number of neighbors under the new metric is even more sensitive to word length than number of neighbors under the old metric because of including more remote neighbors.

Table 9. Correlations between neighborhood density and length for old and new metrics when neighborhood radius is kept constant.

	Old metric	**New metric**
Words 5-7 segments long (n=3545)	-.38	-.56
Words 8-10 segments long (n=598)	-.10	-.25

The fact that the lexicon falls apart along length boundaries is a testament to the strength of the phonotactic constraints of English, which, for an average word, rule out many more additions than substitutions. It is not predicted by the above formulas since the maximum number of additions is slightly greater than the maximum number of substitutions.

There are certain limitations of our metric of lexical/phonological similarity. A fundamental limitation is that the raw number of links is used to predict reaction times and familiarity judgments, rather than the sum of the strengths of the links.

Several ways in which links could be weighted are apparent. One is that the weight of a link should reflect the proportion of the word's segments that are mismatched.[7] Vitz and Winkler (1973) have found that in their data there is a correlation of .81-.95 (depending on the experiment) between similarity judgments and the proportion of mismatched segments. In addition, we need to take into account how similar substituted segments are to each other and how salient the inserted or deleted segments are, factors shown to affect similarity and confusability (cf. Kapatsinski, 2005b for review). Furthermore, a mismatch of x phonemes has the same effect on similarity under the present metric regardless of whether the mismatched phonemes are adjacent to each other, while sound similarity judgment data show that discontinuous mismatches are more salient (Kapatsinski, 2005b, in press b), a finding that may have to do with the fact that a discontinuous mismatch is likely to involve several suprasegmental units. The number of mismatched syllabic constituents has been shown to affect sound similarity judgments in addition to the number of mismatched segments (Kapatsinski, 2005b, in press b).

---

[7] This does not necessarily mean that everything is connected to everything, since no link would be postulated if the ratio of segments shared by two words to the number of segments in the base word is smaller than the allowed minimum ('neighborhood radius').

Finally, changes in all positions in the word are weighted equally at present. There are two reasons why this is problematic. One is that final and initial positions are more salient in sound similarity judgments than medial positions (Kapatsinski, 2005b, in press b) and the end of a stimulus is especially important in determining confusability (Fallon Coble & Robinson, 1992; Kidd & Watson, 1992). The second reason, which is even more serious, is that in phonological priming words that share beginnings inhibit each other while words sharing ends often show facilitation (Radeau et al., 1995). Furthermore, words sharing beginnings appear to inhibit each other in picture naming (Vitevitch et al., 2004) which makes the use of position-insensitive neighborhood definitions an inadequate predictor of reaction times.

In addition, morphologically complex words were included. The results indicate that the inclusion of morphologically complex words leads to finding effects of the clustering coefficient CC2 and hubness. Given the findings that words are often (Hay, 2003) or exclusively (Stockall, 2004) accessed through their roots, the inclusion of morphologically complex words is problematic. Furthermore, it is not clear whether neighbors of all forms of the word can influence the processing of any given wordform or whether neighbors of the root can. That is, whether neighborhood relations are relations between inflected forms, derived bases, or roots.

The data against which the metric has been tested at present are not ideal. Words were presented to subjects visually in the lexical decision task, allowing facilitatory sublexical effects during orthography-to-phonology conversion to be overlaid on the inhibitory phonological neighborhood density effects. Familiarity judgments are a metalinguistic task, which involves postlexical processing.

A promising avenue for modeling behavior using neighborhood density as a predictor has been provided by Pylkkanen et al. (2002) who found two ERP (event-related potential) components on the MEG (magnetoencephalogram), only one of which was sensitive to neighborhood density. An early component occurring at about 170 ms after word presentation was found to be sensitive only to phonotactic probability but not to neighborhood density while a component peaking at 350 ms was sensitive to neighborhood density. Generally, the M350 is thought to index lexical access. Examining neighborhood density effects using MEG and EEG may allow us to investigate the effects of neighborhood density at the first stage of processing that is sensitive to lexical-level variables (M350 is also the first ERP component sensitive to word frequency, Embick et al. 2000). Such an early component is less likely to show strategic effects. Comparing M170 effects to M350 effects also provides a way to deconfound neighborhood density and phonotactic probability/sublexical unit frequency.

Conceptualizing the lexicon as a complex network provides us with a number of variables that may influence the speaker/hearer's processing of the words. Graph theory has provided us with various measures of clustering. One other theoretically promising variable for modeling priming and word recognition is the average degree of the word's neighbors, its 'neighbor density'. A promising network-based variable for modeling confusability is the number of neighbors that two potentially confusable words have in common. Finally, some node pairs provide connections between neighborhoods, while others lie within giant, almost fully interconnected clusters. For instance, two CVC words sharing a rare VC unit and containing different frequent CV units will form one of the few links connecting together two large CV-based neighborhoods and may be especially important for analogical extension of linguistic patterns.

While graph theory provides us with powerful tools for describing networks, it is not, on its own, a theory of the mental lexicon. To provide such a theory, we need a psychologically real similarity measure. We also need to understand which of the many characteristics of a node humans are sensitive to in a particular task. That relative sensitivity will surely be constrained by as well as provide constraints for our theories of how the lexical network is used in a wide range of behavioral tasks.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/ experimental study. *Cognition*, *90*, 119-61.

Andrews, S. (1997). The role of orthographic similarity in lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin and Review*, *4*, 439-61

Balota, D.A., Cortese, M.J., Hutchison, K.A., Neely, J.H., Nelson, D., Simpson, G.B., & Treiman, R. (2002). The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords. http://elexicon.wustl.edu/, Washington University.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509.

Batagelj, V., & Mrvar, A. (2003). Pajek - Analysis and visualization of large networks. In M. Jünger & P. Mutzel (eds.), *Graph Drawing Software* (pp. 77-103). Berlin: Springer.

Burzio, L. (2002). Missing players: Phonology and the past tense debate. *Lingua*, *112*, 157-199.

Bybee, J.L. (2002). Sequentiality as the basis of constituent structure. In T. Givón and B. Malle (eds.), *The evolution of language out of pre-language* (pp. 107-132). Amsterdam: John Benjamins.

Bybee, J.L. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*, 425-455.

Bybee, J.L., & C.L. Moder. (1983). Morphological classes as natural categories. *Language*, *59*, 251-270.

Embick, D., Hackl, M., Schaeffer, J., Kelepir, M., & Marantz, A. (2001). A magnetoencephalographic component whose latency reflects lexical frequency. *Cognitive Brain Research, 10*, 345–348.

Fallon Coble, S., & Robinson, D.E. (1992). Discriminability of bursts of reproducible noise. *Journal of the Acoustical Society of America*, *92*, 2630-2635.

Fujimura, O., Macchi, M., & Streeter, L.A. (1978). Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and Speech*, *21*, 337-346.

Greenberg, J.H., & Jenkins, J.J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, *20*, 157-177.

Gruenenfelder, T., & Pisoni, D.B. (2005). Modeling the mental lexicon as a complex graph. Ms. Indiana University.

Hay, J. (2003). *Causes and consequences of word structure*. London: Routledge.

Kapatsinski, V.M. (2005a). Characteristics of a rule-based default are dissociable: Evidence against the Dual Mechanism Model. In S. Franks, F. Y. Gladney, & M. Tasseva-Kurktchieva (eds.), *Formal approaches to Slavic linguistics 13: The South Carolina meeting*, 136-146. Ann Arbor, MI: Michigan Slavic Publications.

Kapatsinski, V.M. (2005b). *Productivity of Russian stem extensions: Evidence for and a formalization of network theory*. M.A. Thesis: U of New Mexico.

Kapatsinski, V.M. (In press a). To scheme or to rule: Defining attributes of the dual mechanism default are dissociable. *BLS 31*.

Kapatsinski, V.M. (In press b). Phonological similarity relations: Network organization of the mental lexicon. *Proceedings of VIII Encuentro Internacional de Linguistica en el Noroeste*.

Kidd, G.R., & Watson, C.S. (1992). The “proportion-of-the-total-duration rule” for the discrimination of auditory patterns. *Journal of the Acoustical Society of America*, *92*, 3109-3118.

Ladefoged, P., & Maddieson, I. (1986). Some of the sounds of the world's languages*, UCLA Working Papers in Phonetics, 64*.

Luce, P.A., Goldinger, S.D., Auer, E.T., & Vitevitch, M.S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception and Psychophysics*, *62*, 615–625.

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

Marslen-Wilson, W. (1990). Activation, competition, and frequency in lexical access. In: G.T.M. Altmann (ed.), *Cognitive models of speech processing: psycholinguistic and computational perspectives* (pp.148-173). Cambridge, MA: MIT.

Nusbaum, H.G., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception Progress Report No. 10*. (pp. 357-376). Bloomington, IN: Speech Research Laboratory, Indiana University.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 195-247.

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

Pylkkänen, L., Stringfellow, A., & Marantz, A. (2002). Neuromagnetic evidence for the timing of lexical activation: An MEG component sensitive to phonotactic probability but not to neighborhood density. *Brain and Language*, *81*, 666–678.

Radeau, M., Morais, J., & Segui, J. (1995). Phonological priming between monosyllabic spoken words. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1297-1311.

Steriade, D. (2004). Knowledge of similarity and narrow lexical override. *BLS 29*, 583-598.

Steriade, D. (2001). Directional asymmetries in place assimilation. In E. Hume & K. Johnson (eds.), *The role of speech perception in phonology*. New York: Academic Press.

Stockall, L. (2004). *Magnetoencephalographic investigations of morphological irregularity and identity*. PhD Thesis: MIT.

Vitevitch, M.S. (2004). Phonological neighbors in a small world. Ms. University of Kansas.

Vitevitch, M.S., Armbruster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 514-529.

Vitz, P.C., & Winkler, B.S. (1973). Predicting the judged "similarity of sound" of English words. *Journal of Verbal Learning and Verbal Behavior*, *12*, 373-388.

Watts, D.J., & Strogatz, S. (1998). Collective dynamics of `small-world' networks. *Nature*, *393*, 440-442.