Does the Net Work?
Computationally Modeling Differences in the Mental Lexicon

Xi Chen, Maryam Hedayati, Aman Panda, Camden Sikes, Duo Tao, and Tegan Wilson
Carleton College Computer Science Department

Abstract

This paper seeks to examine the robustness of the phonological network to individual differences in lexicon size and composition. The graph-theoretic properties of subgraphs of the phonological network are compared in order to model the fact that different people know different sets of words. In addition, an incremental approach is used to examine changes in graph-theoretic properties as individuals acquire words over time. Using both approaches, we determined that degree centrality is the most robust measure to changes in the graph, and that clustering coefficient and closeness centrality are both robust, but not to the extent of degree centrality.

Introduction

The process of recognizing and understanding spoken language is complex. Experts in the fields of computer science and psychology have applied various approaches to understand the factors contributing to the speed and accuracy of word recognition. Insight into the detailed workings of human language understanding gleaned from this research would be valuable for both disciplines. Prior research in the field has looked at the impact of factors such as frequency and number of neighbors in speech recognition (Luce & Pisoni, 1998).

During the process of speech perception, as a word unfolds, multiple lexically related words compete for recognition. For example, hearing the word "cat" may make internal representations of words such as "cot" and "cap" more salient. One common approach of modeling this process is by representing every word in the lexicon as a node in a graph. Any two words in the graph will be connected by an edge if they differ by one phoneme (by addition, substitution, or deletion). (Luce & Pisoni, 1998) The properties of this graph, including average path length and degree distribution have been previously investigated (Vitevitch, 2008).

However, the graph-theoretic approach outlined above has several limitations as it has been investigated thus far. Because of the complex nature of this graph, generalizations and simplifications have often been made that make it difficult to extrapolate wider implications from the findings. For example, some studies have used mini-graphs containing only the immediate and 2-hop neighbors of a word, and assumed these are representative of the entire lexicon (Vitevitch, Ercal, & Adagarla, 2011). Additionally, most papers draw conclusions based on the entire graph, despite the fact that we each know different words, meaning that each person's phonological network will look different.

There has been research investigating the extent to which modifications of a graph influence its properties (Wei, Joseph, Liu, & Carley, 2015; Tsugawa & Ohsaki, 2015; Borgatti, Carley, & Krackhardt, 2006). However, this research has not been on graphs of the human lexicon, which means the results may not be generalizable for our purposes. In addition, these papers usually modify graphs at random, whereas we are interested in representing the ways in which human lexicons may differ. For example, it is much more likely that someone will not know a low-frequency word in the language than a high-frequency word, so using a uniformly random approach to remove words from the graph may not be representative.

The goal of this project was to use a graph-theoretic approach to investigate the robustness of the human lexicon to individual differences in phonological network composition, aiming to model differences in people's vocabularies based on their age (number of words they know) or environment (words they have been exposed to).

Our approach to investigate the role of environmental differences is to create different lexicons, each representing the vocabulary of a different individual. We begin with a standard lexicon containing every word in the language, and use that to output multiple lexicons, representing the different mental lexicons that multiple people may have. After a graph is generated from each of these lexicons, its graph-theoretic properties are calculated. We then compare these graph-theoretic properties across multiple graphs. Using this approach, we found that degree was the most robust measure across probabilistic sublexicons. Clustering coefficient was found to be robust to error, whereas closeness centrality was robust to inversions.

Additionally, we are interested in investigating the way the phonological network evolves over time, and the extent to which people differ based on how many words have been acquired. To investigate this question, we used an incremental approach, by adding one node to the lexicon at a time and obseriving how degree, clustering coefficient, and closeness centrality change. Specifically, words were sorted by either their age of acquisition or frequency and compared to a uniformly random baseline in order to understand how the properties of the graph changed as words were added. We found that degree was a robust measure, that clustering coefficient followed a similar pattern to degree, and that closeness centrality was less robust to error. We also found that as words are added to the network as ordered by age of acquisition, the network densifies and the effective diameter shrinks

Graphs

Structure of the Phonological Network

The phonological network reveals the relation between words in terms of their pronunciations. The pronunciations are split into phonemes and represented by the International Pho-

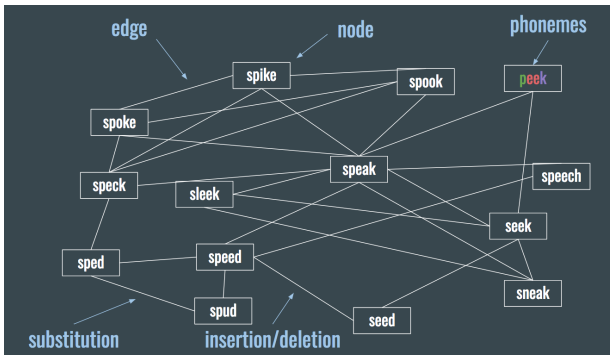

Figure 1: An example phonological graph representing the structures and concepts of nodes, edges, phonemes. Presented in the figure are spellings but what actually take effects is their pronunciations. Each word is a node, and there are edges connecting nodes differ by one phoneme. The "ee" in the peek is a phoneme, and the cases of connecting by one substitution, insertion and deletion are shown on the bottom of the graph.

netic Alphabet. Thus, each word is represented only by the phonemes, regardless of spelling.

Each node represents a word, with its pronunciation, spelling, and frequencies, or other relevant information according to our needs stored inside the node. An edge connects two nodes when the two nodes' pronunciations differ by one phoneme. By differing by one phoneme, it means that if we add one phoneme, delete one phoneme or replace one phoneme with another, we get another word's pronunciation. An example phonological network is shown in Fig.1.

It is worth noting that there are words in English that are pronounced the same but spelled differently and have different meanings. They are called homophones. Our strategy for dealing with this is to add the frequencies of the homophones together and combine the nodes into one. It makes sense because in the current stage of this study, we are mostly concerned with the pronunciations of the words themselves, not the meanings of the words. This method ensures that we have a well defined graph, with each node representing a distinct pronunciation.

Centrality Measures

1. *Degree centrality.* One of the measures we used to learn about the structure of the phonological network is degree centrality. Degree centrality is the measure of the number of neighbors in a node. In each node, its degree centrality is the number of neighbors it has. Thus, if we average the degree centrality of all nodes, we get a global measure, reflecting the average number of neighbors a node has in this graph. The higher this number is, the more dense the graph is. We can normalize this metric by dividing this by the total number of possible edges that could exists in the graph, which is $N(N-1)/2$, where N is the number of nodes in the graph. The expression for degree centrality is defined as
$$c_{deg} \equiv \frac{2E}{N(N-1)}. \quad (1)$$
From Eqn.1, we can see that if the graph is complete. The normalized degree centrality will be 1 and 0 if the graph is there is no edge in it.

2. *Clustering coefficient.* Another centrality measure used is the clustering coefficient. Clustering coefficient is a centrality measure that has a similar effect as degree centrality. The fact that the higher the metric is, the more centralized the graph is, also applies. However, cluster coefficient is slightly different in the way it works. It works by examining triplets of the nodes in the graph. If the triplet is fully connected, or forms a triangle, we count up by one. Then the clustering coefficient will be the number of triplets that are connected divided by the total number of connected triplet in the graph. The number of possible triplets is $N(N-1)(N-2)/6$. Denote the number of triangles in the graph as T. Then, the clustering coefficient would be
$$c_{clu} \equiv \frac{6T}{N(N-1)(N-2)}. \quad (2)$$
It is also clear that the clustering coefficient for a complete graph would be one, and zero for a graph without edge. We can think of the clustering coefficient in a way that the clustering coefficient a conceptual extension from the centrality measure when the number of nodes examined is increased from two to three. If we want the centrality measure, we examine two nodes at a time and see if they are connected; for clustering coefficient, we examine three nodes and, in the same way, if they are connected, we count up by one.

3. *Closeness centrality.* We also consider the centrality measure of closeness centrality. Closeness centrality works in a different way than the two above, degree centrality and clustering coefficient. Closeness centrality measures the closeness of a node to the rest of the graph. Thus, by distance, it means the length of the shortest path. If we calculate the shortest path of all distances from a node to all other nodes, we will have a list of numbers with one number corresponding to a destination node. Once the average of these numbers are taken, we will have the closeness of this node. We have one number for the whole graph, we have a number for each node in closeness. Thus, once we calculate all the distances between all pairs of nodes in the graph, we will be able to calculate a list of numbers based on that, indicating the closeness of each node. Mathematically, we define the centrality measure as
$$c_{clo}(n) = \frac{\Sigma_s d(n,s)}{N-1} \quad (3)$$
where N is the number of nodes, $d(n,s)$ is the distance from node n to node s. Thus, we sum up all the shortest distances from n to other nodes and take the average. By doing that, we get the closeness of node n.

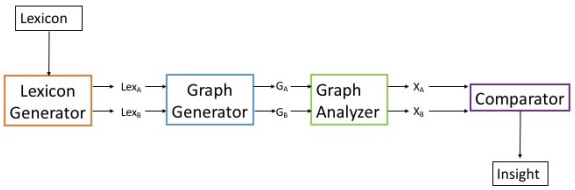

Figure 2: general workflow of the project

Probabilistic Approach

Pipeline

1. *Lexicon Generator* The first step of our pipeline is the lexicon generator. The purpose of the lexicon generation step is to create variations on the main lexicon. Ultimately, we wish to emulate the lexicon of individuals, which of course vary.

   One way to generate sub-lexicons is through word removal from the main graph on the basis of a frequency threshold. Specifically, this method involves using relative frequency data and defining a cut-off value. We then remove all words with frequencies lower than that value. Defining the cutoff roughly simulates different vocabulary sizes. In more concrete terms, given a lexicon L and a cutoff value v, where each word $w \in L$ has an associated frequency f_w, then we could produce a new lexicon $L' = \{w \in L | f_w \geq v\}$.

   In the real world, a person's vocabulary is more nuanced: people know some uncommon words and don't know some common words. This leads us to a probabilistic approach, which is our primary way to emulate an individual's lexicon. This method will be further elaborated upon in the data generation section. Additionally, to generate random sub-lexicons under this paradigm, we may simply assume a uniform distribution of word probabilities.

2. *Graph Generator* On generation of a lexicon, we construct a corresponding phonological network. A node in the network corresponds to a phonological representation of a word, where words with the same pronunciations are counted as one word. Two nodes are connected if the words they represent are phonologically similar - that is, the first word can be transformed into the second word by a substitution, addition, or deletion of one phoneme. For example, the word /kt/ (cat) would have phonological neighbors /bt/ (bat), /rt/ (rat), /t/ (at), and /kp/ (cap).

3. *Graph Analyzer* After a graph is constructed to represent a given lexicon, we wish to analyze the properties of this graph using centrality measures that have been outlined previously in this paper. We consider both global and local properties of the graph, where a global property assigns a score for the entire graph, and local property assigns a score to each vertex. Global properties include transitivity and average shortest path. Local properties include betweenness centrality, degree centrality, eigenvector centrality, clustering coefficient, and average path length (in a given component). Each of these measures indicate information about the structure of the graph and the lexicon it represents.

4. *Comparator* The graph analyzer outputs centrality measures about an inputted phonological network. The comparator piece of the pipeline can consequently generate an overlap n comparison when dealing with two graphs, and can compute average error when dealing with n graphs for $n \geq 2$. So for top n overlap comparison of two graphs, if we have V_a and V_b, the n vertices with highest centrality in graphs G_a and G_b respectively, then overlap n is the ratio of the size of $V_a \cap V_b$ to the size of $V_a \cup V_b$. Average error has been approached in two different ways: absolute error and relative error. This is delved into further in the data analysis section.

Generating Data

Our approach to altering the lexicon will begin by removing words in a probabilistic way based on frequency. People invariably do not know all the words in a given lexicon, and are more likely to know words that are used more often. The relative frequencies of words in our English lexicon data have been measured, so we can use this to decide which words to remove. If we consider a word w, we divide the frequency of w by the sum frequency over all words, and so can model the probability that given any word in our lexicon, that word is w. In more formal notation, let $F = \sum_{w \in L} f_w$, the sum frequency over all words. Then the probability of a word w is $P(w) = f_w/F$. If we make the simplifying assumption that after someone encounters any word at least once, it is in their mental lexicon, then we can model a person's mental lexicon: if a person has been exposed to n words, then we can probabilistically draw n words (with repetition allowed) from the total lexicon. This gives us a model to calculate the mental lexicon.

In our approach, we create 200 such lexicons, each created by taking one million word instances, and calculate their centrality measures using the techniques identified previously. We call these the frequency-based lexicons. We also generated lexicons by drawing word instances uniformly. We call these the uniform lexicons. For the uniform lexicons, we draw to ensure lexicons of a similar size to the frequency-based lexicons. Because of time constraints, we only generated 50 uniform lexicons.

Top N

The first way we began to analyze our results was by calculating the top n overlap between two lexicons. (Fig 3) The top n overlap value for some integer n is: if we take two sets, the n vertices with highest centrality in lexicon a and the n vertices with highest centrality in lexicon b, then the overlap n value is the percent of vertices which are in both sets. The purpose of doing this was to examine how similar any pair of lexicon

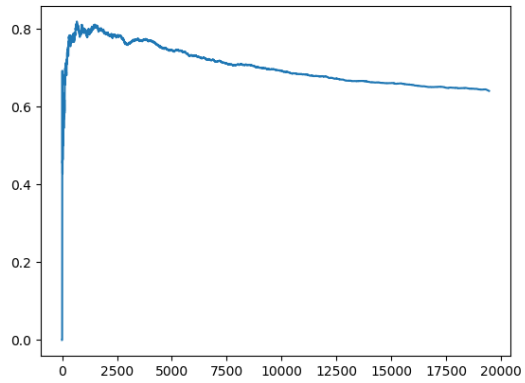

Figure 3: top n overlap graph on degree centrality.

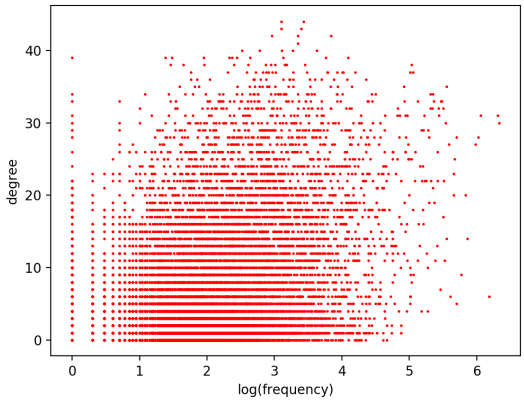

Figure 4: log frequency vs degree

models we generated would be. What we found was quite interesting: theoretically, the overlap n value should increase as n does, because if n is larger, then the sets we compare are larger, and so we have a higher likelihood of getting overlapping vertices. However, what we find in the figure presented is that for degree centrality, instead overlap n peaks early, at around $n = 1500$, and decreases afterward. This seems to indicate that the vertices that are more likely to overlap in both graphs in general are also more likely to have high degree, which points to a correlation between degree centrality and frequency.

On further inspection of degree centrality and frequency, a statistically significant relationship was observed. We found that when plotting degree as a function of log frequency, we observed an r^2 value of 0.174 with a p-value < 0.001, which gives us evidence that there is predictive power in frequency for the degree centrality measure. This could corroborate the observations in the previous paragraph. When thinking about why this may be the case, we came to the following conclusion: in language, words that are shorter tend to be simpler and used more frequently. Words that are shorter are also more likely to have neighbors in our lexicon graph; since neighboring words must be one phoneme apart from one another, smaller words have to match on less phonemes than longer words do to still be considered neighbors. In this sense, it is easier to have many neighbors as a shorter (and likely a simpler and more frequent) word.

Centrality Comparison

After generating both the frequency-based and the uniform lexicons, we compared their average centrality values in various ways, depending on the centrality measure. For centrality measures which return a single value for the entire graph (giant component size and transitivity), we calculated the average and standard deviation of the value over the generated lexicons of the same type. Our results are given in Table 1. Although the frequency-based and uniform lexicons have a similar number of nodes, the giant component of the frequency-based lexicons is much larger and more cohesive than of the uniform lexicons. We also see a small difference in their transitivity, again indicating that the frequency-based lexicons are denser than the uniform lexicons. Notably, the uniform lexicons have higher standard deviation in both the giant component and transitivity score, which shows that not only are they less dense than frequency-based, but also that their topology changes more significantly between generations, while the frequency-based lexicons maintain a more consistent topology.

Table 1: average and standard deviation of centrality measures on generated data

Centrality	Freq. Based Prob.		Uniform Prob.	
	Avg.	Std. Dev.	Avg.	Std. Dev.
Giant Component	8438	55.346	2860	66.92
Transitivity	0.295	0.001	0.282	0.006

For centrality measures which return a value for every node in the graph, we compared the centrality values of nodes in the generated lexicons to the value assigned by the full lexicon. Taking the difference between the two gave a measure of error, and we examined two types of error: absolute and relative. Given some actual value a and observed value o, the absolute error is the difference between a and o, that is, $(a-o)$. The relative error is the absolute error normalized by the actual value, that is, $(a-o)/a$. Note that this allows error to be either positive or negative, which differentiates between which of a and o is larger.

To examine the error of the generated lexicons (Fig 5), we first created a matrix which contains the centrality values for all 200 generated lexicons of a single type, ie. frequency-based or uniform probability, where index (i, j) of the matrix contains either the centrality value of vertex V_i in graph G_j, or null if $V_i \notin G_j$. We then applied error to each value in the

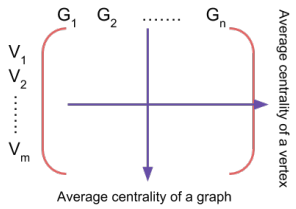

Figure 5: error averaging process

matrix. We investigated both absolute and relative error here. Finally, we averaged the error values across either rows, to get a vector containing the average error of each vertex, or averaged down columns, to get a vector containing the average error of each graph.

Figures 6 and 7 give histograms of the absolute and relative error of degree centrality, averaged over the vertices present within 200 frequency-based probabilistic lexicons. From both figures, a majority of the vertices have approximately 0 absolute error. In figure 7, a majority also have $+/-1$ relative error. The spike of around 1 relative error shows a subset of vertices for which $(a-o)/a \approx +/-1$. So, $(a-o) \approx +/-a$. If $(a-o) \approx a$, this indicated that $o \approx 0$. So, the spike can be attributed to vertices for which the average observed degree centrality over the 200 probabilistic lexicons is 0 or almost 0. This most likely occurs for words that only have low frequency neighbors in the full lexicon. Although such a word may be drawn into our frequency-based lexicons, it is unlikely that its neighbors will also be drawn, giving it an average degree centrality score of close to 0. But, if $(a-o) \approx -a$, then $o \approx 2a$. Since we are considering degree centrality rather than degree, this is technically possible. It occurs when some word keeps all or almost all of its neighbors, but the number of words in the generated lexicon is about half what it is in the full lexicon. This seems to indicate that even though the frequency-based lexicons are smaller in number of nodes that the full lexicon, many words are keeping all their neighbors through the transformation. This indicates that a decent portion of words have only neighbors of high frequency.

Figure 8 gives the absolute error histograms of degree centrality, averaged over the 200 frequency-based lexicon graphs. It models a normal distribution, focused around $6.75e^{-5}$. Since the frequency-based graphs have on average, around 19,400 words, this indicates an absolute error centered around 1.3 neighbors. So for any frequency-based lexicon, the degree (number of neighbors) of words is on average 1.3 neighbors off. The regularity of error values suggests that the lexicons hold some consistent patterns, which may not be an issue, or may be correctable. This leads us to our next way of examine error: counting inversions.

Say we take the main lexicon and order its vertices by their centrality score, and compare it to a sublexicon whose vertices are ordered by their sublexicon centrality. We can examine any pair of vertices in the sublexicon and see if their

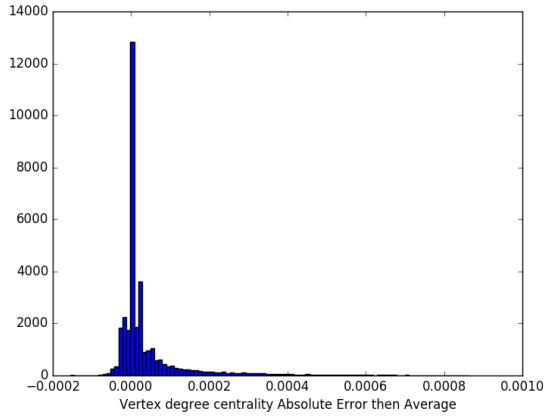

Figure 6: degree centrality absolute error calculated over average vertex centrality of frequency-based lexicons

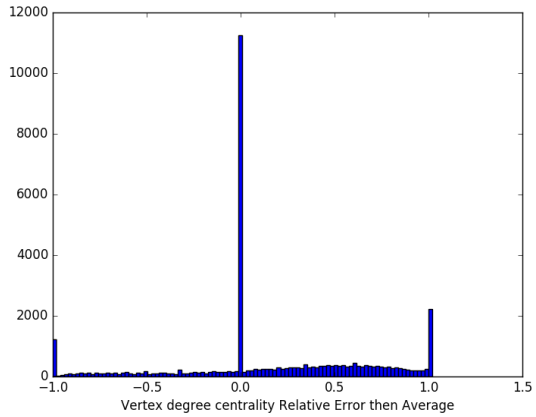

Figure 7: degree centrality relative error of frequency-based lexicons

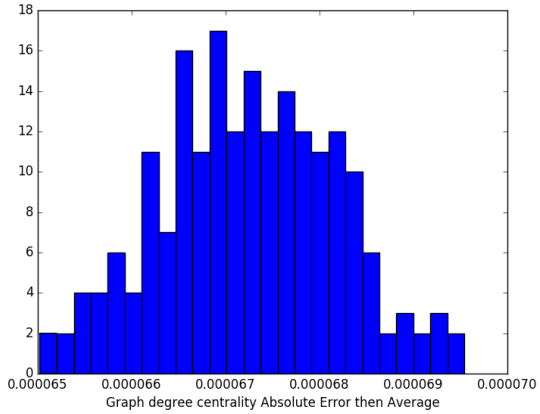

Figure 8: degree centrality absolute error calculated over average graph centrality of frequency-based lexicons

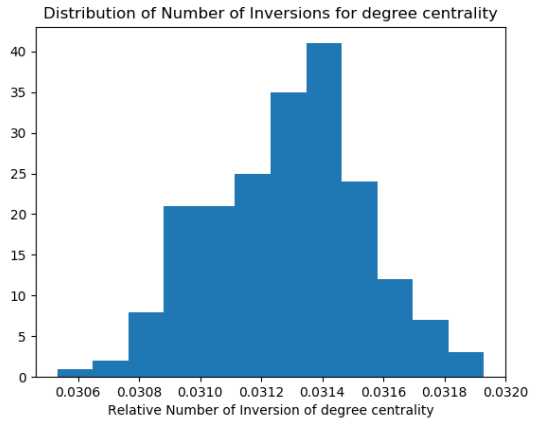

Figure 9: degree centrality percentage of pairs inverted in frequency-based lexicons

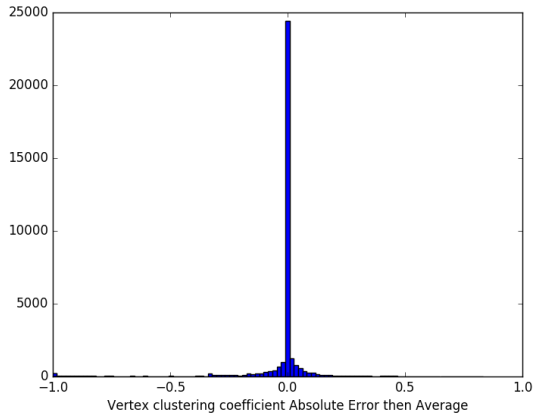

Figure 10: clustering coefficient absolute error calculated over average vertex centrality of frequency-based lexicons

relative order in the same in both the sublexicon list and the main lexicon list. If it is, then we say the pair is not inverted. But if the order is different, then we say the pair is inverted. We examined all possible pairs, and took a percentage value of how many were inverted, to get a single score for each sublexicon. We then plotted the value as a histogram, in Figure 9. When we calculated this for degree, we saw that on average, around 3.13% of vertex pairs were inverted. So for those vertices, it would be incorrect to use the main lexicon to compare their degree values. Although it is important to keep track of, we determined it to be acceptable.

Because degree had low error and few inversions on average, the main lexicon does a decent job approximating the sublexicons. Therefore, over the frequency-based sublexicons, degree is robust.

We examined various other centrality measures, as well. For a full set of histograms, please refer to our website.

We next examined clustering coefficient. We found that clustering coefficient tended to have extremely low error over vertices in frequency-based sublexicons, (Figure 10) both relative and absolute error. So from our first perspective, clustering coefficient seemed even more robust than degree. However, over 6% of vertex pairs were inverted by clustering coefficient values in the sublexicons. (Figure 11) This suggests that it may be correct to use the clustering coefficient of a single vertex from the main lexicon, it would likely be incorrect to compare two vertices' clustering coefficient score from the main lexicon in order to determine SWR findings. The result was somewhat inconclusive. The main lexicon is still useful at approximating the clustering coefficient of the sublexicons, but only in certain case. Therefore, clustering coefficient is somewhat robust, though is limited when comparing two vertices and their values.

Closeness centrality, was considered next. Closeness centrality also demonstrated a low absolute average error in 1.8 $\times 10^3$; however, as can be seen in the distribution of closeness

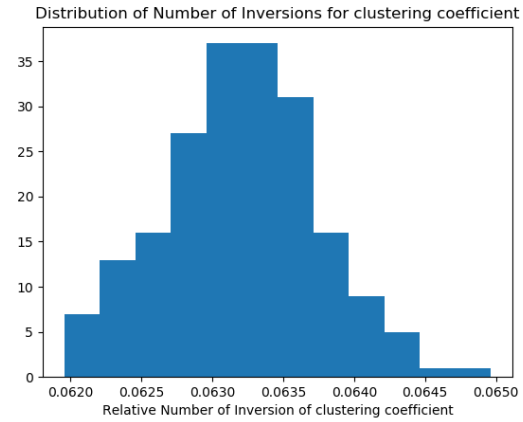

Figure 11: clustering coefficient percentage of pairs inverted in frequency-based lexicons

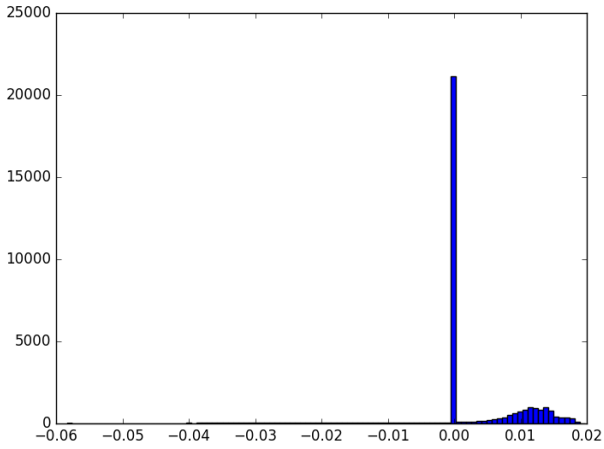

Figure 12: closeness absolute error calculated over average vertex centrality

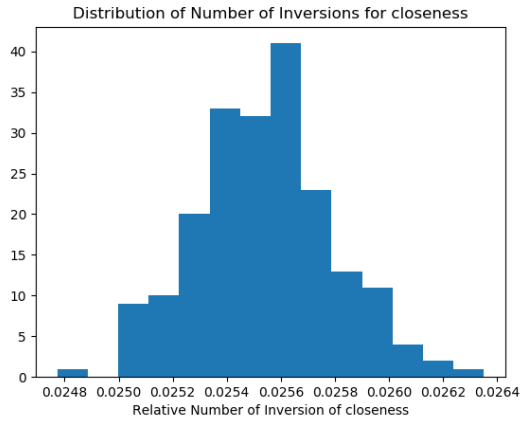

Figure 13: closeness percentage of pairs inverted in frequency-based lexicons

absolute error over vertices (Figure 12), there is a marked region in our distribution of error in which the frequency-based sublexicons are not accurately capturing closeness centrality values. This is represented by the density in the distribution centered around 0.015. We find here that for a non-negligible subset of nodes, the frequency-based sublexicons are systematically scaling values at a high error rate from the main lexicon. When relative error is concerned through counting inversions, we get a better result. Specifically, we find that only 2.56% of vertex pairs were inverted by closeness centrality values in the sublexicons. The distribution of closeness inversions is provided in figure 13. So our frequency based approach in constructing sublexicons gives us gives us positive results when we consider relative error, though our absolute error calculations are hinged with uncertainty. Hence, as we found with clustering coefficient, our findings are inconclusive. The main lexicon is not completely accurate in estimating closeness centrality in the sublexicons and we can say this centrality measure is only somewhat robust.

In total, we found the following: Degree is a robust measure. Clustering coefficient and closeness are somewhat robust, in different ways: clustering coefficient is robust to error, while closeness is robust to inversions.

Incremental Approach

Given our goal of understanding the robustness of speech perception findings based on individual differences in human lexicons, we were interested in another way that humans differ in their lexicons, their age. We learn words as we grow older, and an average 5-year old presumably has a very different vocabulary than a 22-year old. We were interested in using information about the order in which we acquire words to investigate how mental lexicons and their robustness evolve over time.

The approach we used to investigate the role of age of acquisition in the robustness of the lexicon was incremental. In this approach, the words were ordered by some parameter (representing the order in which the words were learned), and added to the graph in this order. For each node that was added to the graph, the centrality measures (degree and clustering coefficient) were updated accordingly. Thus, for every i from 1 to n (number of words in the lexicon), we have the centrality measures calculated for the graph containing the top i nodes, ordered by the specified parameter.

This method makes the assumption that we are able to order the nodes based on the order in which they are learned. However, obtaining accurate estimates of which words an individual knows is a great challenge, particularly with groups such as pre-verbal infants. Instead of recruiting participants at different ages and collecting data on which words they know at that point, a common approach to get estimates for the average age at which a given word is acquired (age of acquisition, or AoA) is to ask adults at which age they learned a given word. This has been done both in lab settings, as well as in an extensive online study which presented AoA ratings for 30,121 words (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). The results from both settings are highly correlated, and we therefore are able to obtain AoA ratings for the words in our graph.

However, although asking adults at which age they learned a word allows us to collect data more efficiently and extensively, it is unclear whether it accurately reflects the age at which the words were actually acquired. Kuperman et al (2012) point out that responders put the primary weight of learning between ages 6-12, which does not reflect other vocabulary size estimates. In addition, an average adult will not be able to retrieve an episodic memory of the moment at which they learned every single word, and they must therefore be using some other quality of the word to make these judgments. Kuperman et al (2012) argue that although the AoA estimates may not accurately reflect the age at which words were acquired, they still likely reflect the order in which the words were acquired, but this claim is speculative.

Despite the AoA dataset potentially not being an accurate reflection of the age at which or order in which words were acquired, the dataset is still interesting given that participants are not making these ratings arbitrarily. There is some consistency to the way that the words are being ranked, which may be some combination of factors such as word-length, familiarity, word retrieval fluency, and perceived simplicity. In word recognition experiments, AoA highly correlated with dependent variables such as lexical decision time, and accounted for some of the variance in addition to factors including word length and frequency (Kuperman et al., 2012). Thus, AoA differences may still be used to represent the fact that different people have different vocabularies.

We applied the incremental approach to the graph, with the nodes ranked by frequency. Due to the potential limitations of the AoA dataset, we also chose to use the same approach, but to order the nodes by frequency. It is likely that high frequency words would be acquired at a younger age, and frequency may therefore also provide insight into the order in

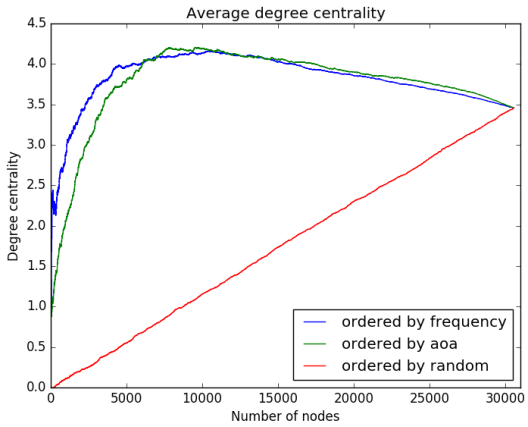

Figure 14: Average degree centrality across all words in the graph with the first n words.

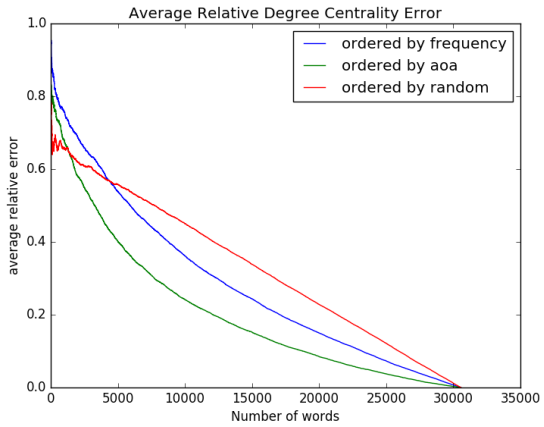

Figure 15: Average relative error in degree centrality across all words in the graph with the first n words.

which words are acquired. Finally, we also ordered the nodes randomly, as a point of comparison. The goal of these analyses was to compare the robustness of the graph and its properties when ordered by AoA, by frequency, and randomly.

In the following sections, we present our results for degree centrality, clustering coefficient, and closeness centrality using the incremental approach.

Degree Centrality

In figure 14, we plot the average degree centrality for the graph with the first n words, ordered three ways. For the random ordering, we see a linear increase in average degree centrality as nodes are added to the graph until it reaches the average in the graph including all the words. This is what we expect, since when half the words have been added we expect each word to have about half of it's neighbors, so the average degree centrality should be about half of what it is in the entire graph.

Looking at AoA and frequency, we see a very different pattern. The average degree centrality increases very quickly, peaks at about 10,000 words, and then decreases slowly until reaching the average in the whole graph. This indicates that words with high degree have lower ages of acquisition and higher frequencies, and that the neighbors of these nodes also have relatively low ages of acquisition and high frequency. The words in the graphs with the first 5,000 words already have on average a higher degree than the words in the whole graph, so they are high degree words and a significant number of their neighbors are already included in the graph as well.

Word length contributes to this trend. Short words tend to have lower ages of acquisition and higher frequencies than long words. Since there are fewer possible short words, and short words are more efficient and so are more common, short words tend to have more neighbors than long words. And these neighbors also tend to show up early on in the ordering.

This trend can also be observed in figure 15, where we plot the average relative error in degree centrality for the graph with the first n words, ordered three ways. Once again with the random ordering, we see a linear trend. This makes sense because after half the words are added to the graph each node should have on average half of its eventual neighbors. The error starts at about 0.7 instead of 1 because about 30% of the words have no neighbors, and thus have no error as soon as they are added into the graph.

For both AoA and frequency, the error is initially somewhat higher than it is for random. This is because words with degree zero tend to be infrequent with high ages of acquisition, so with these orderings there is not the initial boost of words with no neighbors having no error. However, the error drops much more quickly than it does for error, and is significantly lower for most choices of numbers of words. This is due to the fact that the neighbors of words with low age of acquisition and high frequency tend to also have these properties. Thus, as we saw in 14, the nodes included in the graph have more of their neighbors than you would expect with a random ordering, so their degrees are closer to what they will be in the full graph, so the average error is lower.

This plot also shows us that there is an appreciable difference between the orderings generated by age of acquisition and frequency. The average error for the words ordered by AoA is significantly lower than it is for frequency, indicating that the pattern of the neighbors of early occurring words also occurring early is even stronger for AoA. This could partially be because AoA is even more correlated with word length than frequency is. There are plenty of long words that adults use very frequently, but people are not very likely to think they learned a long, complex word early on, even if they use it regularly. This means that short words and their neighbors are added even earlier, further lowering the average error.

To get a better idea of exactly how the error in degree centrality is decreasing, we we plot the proportion of words with the same degree centrality in the full graph and the graph with only the first n words, ordered 3 ways, in figure 16. As ex-

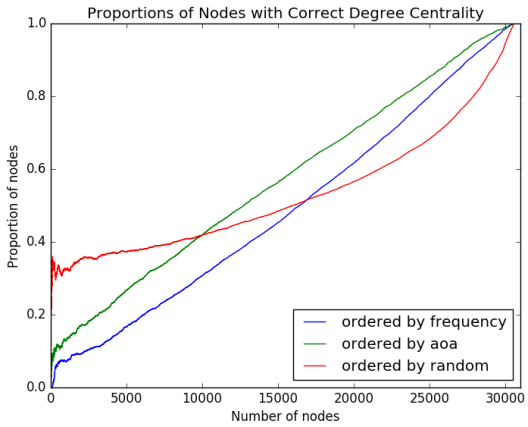

Figure 16: Proportion of words with the same degree centrality in the full graph and the graph with the first n words.

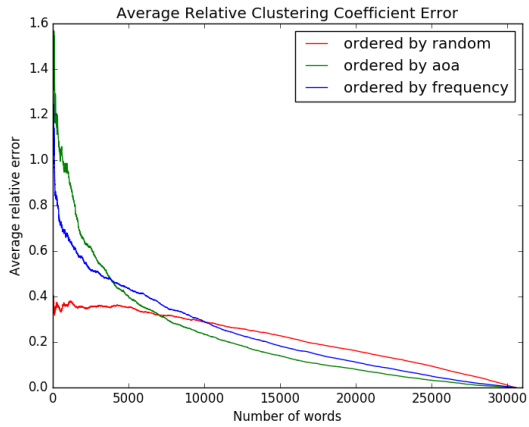

Figure 17: Average relative error in clustering coefficient across all words in the graph with the first n words.

plained above, about 30% of words have no neighbors, which is why the line for the random ordering starts at about 0.3. For AoA and frequency, we see that the proportion of nodes with correct degree centrality increases roughly linearly, which may be surprising. One may expect that most nodes would not have the correct degree centrality until very close to the end. In reality, a sizable number of nodes, which we know have higher than average degree, already have correct degree centrality when significantly fewer than half the nodes have been added to the graph. This cements the observation that nodes with low age of acquisition and high frequency tend to share these properties with their neighbors.

From these three plots we can clearly see that the order that words are added to the graph has a huge effect on both the average degree and the average error in the graph. Assuming AoA and frequency somewhat reflect the actual order that people learn words, then it seems that young people's mental lexicons have higher average degrees than adults. Additionally, using the full lexicon to model the mental lexicons of young people is more appropriate than it would be if they learned words in a different order.

Clustering Coefficient

For clustering coefficient, we see much the same results as we did with degree centrality. This is not surprising, as both measures are only concerned with the direct neighbors of a given node. Also, since in this context edges are always present if both terminal nodes are present, if a node has the correct degree centrality then it must also have the correct clustering coefficient. Thus, the patterns for error in clustering coefficient are quite similar than the patterns for degree centrality. This can be seen in Figure 17. On the other hand, it is possible for a node to have the correct clustering coefficient but not the correct degree. For instance, if none of a nodes neighbors are connected to each other, then the node will always have a clustering coefficient of 0. It turn out that quite a few nodes in the graph have this property – about 65% of them.

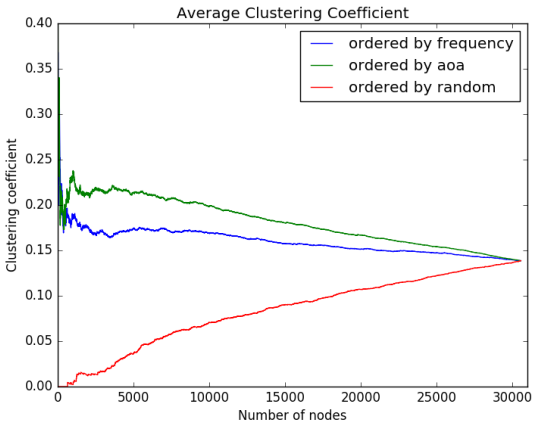

Figure 18: Average clustering coefficient across all words in the graph with the first n words.

The implication of this is that the error of the random ordering is capped at about 0.35, and the proportion of nodes with correct clustering coefficient starts very high.

In figure 18, we plot the average clustering coefficient for the graph with the first n words, ordered three ways. Similar to degree, the average clustering coefficient is significantly higher for small graphs ordered by age of acquisition and frequency than it is when ordered randomly. This once again supports the observation than words with low AoA or high frequency are clustering together.

Closeness Centrality

While degree centrality and clustering coefficient are very local measures, taking into account only immediate neighbors, closeness centrality is based on all the nodes in the graph. So, unsurprisingly, we see a very different result.

In figure 19, we plot the average relative error in closeness centrality for the graph with the first n words, ordered three

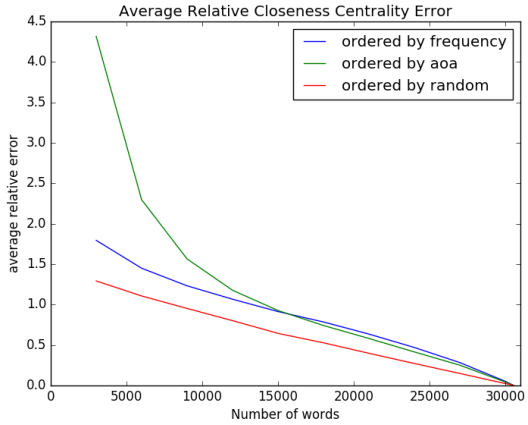

Figure 19: Average relative error in closeness centrality across all words in the graph with the first n words.

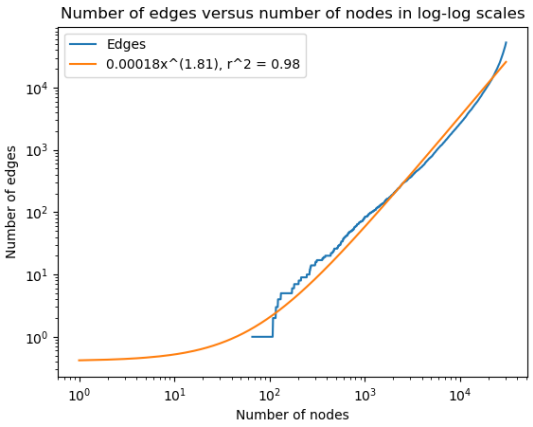

Figure 20: Number of Edges versus Number of Nodes in Log-Log Scales.

ways. Off the bat, we can see that the average error, regardless of the method of ordering, is much higher than it was for the other measures. Enough that the value in the full graph would be a bad approximation even for someone who has learned two-thirds of the eventual total number of words.

Also, The trends we see in this graph are the opposite of what we saw for the other two centrality measures. The graphs where nodes are added in a random order have by far the lowest error. Also, initially the error for words ordered by age of acquisition is significantly higher than frequency as well. This indicates that the order that humans learn words makes closeness centrality in the full graph a worse approximation for young people with small lexicons than it would be if we learned words in a different order.

In short, closeness centrality is significantly less robust to this type of error, meaning that it is less appropriate to use the full lexicon to estimate the closeness of words for a young person with a small vocabulary. This is not too concerning, since closeness centrality is not commonly used by speech perception researchers, while the other two centrality measures we investigated are.

Phonological Network Evolution

Motivation

Dense neighborhood might facilitate lexicon acquisition because the integration of a newly formed lexical representation with numerous existing lexical representations may serve to strengthen the new representation (Storkel, 2004). Empirical studies provide evidence that neighborhood density influences lexicon acquisition. 17-month-old infants briefly exposed to high-density neighborhoods learned the target word from this neighborhood better than the target word from the low-density neighborhood (Hollich, Jusczyk, & Luce, 2002). This implies that newly learned words might have many neighbors in the existing phonological network and the network densifies over time. In addition, we are interested in investigating how does the effective diameter of the phonological network changes over time because shorter average distance between nodes might imply faster lexicon retrieval speed. More specifically, we hope to answer two questions: Does the phonological network densify over time? Does the effective diameter shrink over time?

Densification and Power Law

First, we show that phonological network densifies over time with the number of edges growing superlinearly in the number of nodes. We demonstrate that the number of nodes and the number of edges follow the densification power law.

$$e(t) \propto n(t)^{\alpha}$$

where $e(t)$, n(t) is the number of edges and nodes at time e. We record the number of edges and number of nodes each time we add a node (word) in age of acquisition order. Then we plot the number of nodes vs number of edges in logarithmic scale and fit a linear line. Figure 20 shows that densification power law plot and the slope $\alpha = 1.81$ represents the exponent. Since $r^2 = 0.98$, we conclude that phonological network densification follows the power law. We also observe a high densification exponent $\alpha = 1.81$, which indicates a large deviation from linear growth ($\alpha = 1$). Notice that $\alpha = 1.81$ is a much higher exponent than the exponent reported in (Leskovec, Kleinberg, & Faloutsos, 2007), which implies phonological network densifies faster than most of the social networks.

Shrinking Diameter

We define the effective diameter of a graph G as follows: Let $g(d)$ be the fraction of connected pairs with shortest path length at most d for integer value d. We extend this definition for integer values to all real numbers by linearly interpolating the function value between $g(d)$ and $g(d+1)(d \leq x < d+1)$:

$$g(x) = g(d) + (g(d+1) - g(d))(x-d)$$

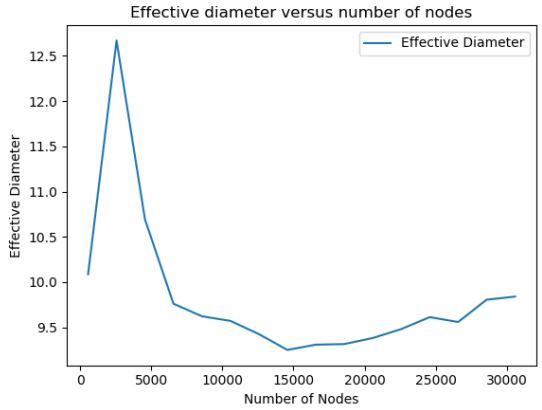

Figure 21: Effective Diameter Versus Number of Nodes.

(Leskovec et al., 2007). We define D as the effective diameter of graph G is $g(D) = 0.9$. The effective diameter d implies that roughly 90 percent of connected nodes are at distance at most d. It has been demonstrated that effective diameter shows qualitatively similar behavior as average diameter, but is more stable than average diameter because outliers such that small component and long chains have a smaller influence on effective diameter.

We remove 2000 nodes from the phonological network each time till we get the empty graph and calculate the effective diameter at each step. Figure 21 shows the result we get. The result does not entirely comply with the shrinking diameter phenomenon observed in social networks. We observe that the effective diameter starts at 10.08 for the first 568 words, then increases to 12.67 for the first 2568 words. This initial increase is expected since if we only have 568 words in the phonological network, it is likely that the shortest paths between nodes will be small because of the size of the graph. Starting from node 2568 to node 14568, we observe that the effective diameter is decreasing. Starting from node 14568 to 30568, we observe an increasing trend again. Because the unusual shape of the plot, we want to verify whether this trend is attributable to the way phonological network is constructed. We identify two factors that could contribute to this phenomenon.

- Disconnected Components: We suspect the decreasing trend is because of the large number of disconnected components in the phonological network. The largest component only contains 38.1 percent of the nodes. In addition, we have 10079 nodes each disconnected from other part of the graph and another 3439 components with size between 2 and 10. The definition of shrinking diameter does not account for the effect of disconnected components since it ignores shortest paths between disconnected nodes. The shortest paths between nodes in small components are expected to be small, which could potentially causes the effective diameter to decrease. In order to eliminate the effect of disconnected components, we plot the effective diameter

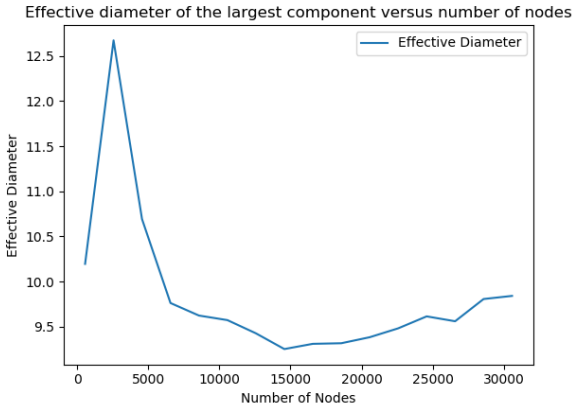

Figure 22: Effective Diameter of the Largest Component Versus Number of Nodes.

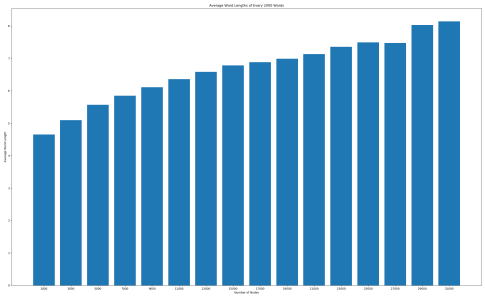

Figure 23: Average Word Length of Every 2000 Words

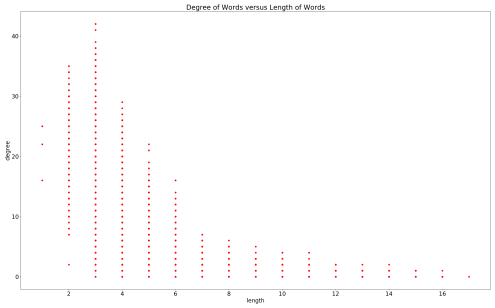

Figure 24: Degree vs Length of Words

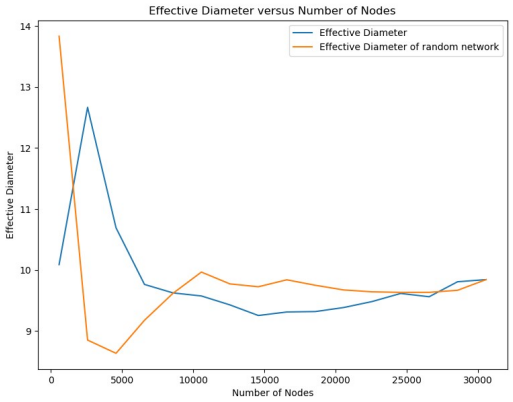

Figure 25: Effective Diameter versus Number of Nodes Ordered by Random

of the largest component and Figure 22 shows the plot. We can see that the plot looks roughly the same as the original plot and we conclude that disconnected components have little effect on the effective diameter.

- Word Length: Shrinking diameter could be a byproduct of the lengths of the newly added word. From Figure 23 we can see that the word length of earlier learned words are significantly shorter than later learned words. From Figure 24 we know that there is a strong correlation between degree and word length. It also make intuitive sense that the probability of seeing a word one-edit distance away from a short word is higher than a longer word. Adding high degree words could potentially create many shortest paths for many pairs of words and thereby decreasing the effective diameter. On the contrary, adding words with degree 1 or 2 could create long chains attached to a giant component and thereby increasing the effective diameter. In order to investigate the effect of word lengths, we randomly shuffle the words in a way that preserves the word length. For each word, we randomly choose a word with the same length without replacement. Figure 25 shows that result. From 5000 to 10000, we observe an increasing trend in the new plot, which indicates from 5000 to 10000, word length is not the factor that contributes to the shrinking parameter.

Conclusion

This paper sought to gain insight into the robustness of the phonological neighborhood graph to individual differences in lexicon composition. We began by looking at differences in lexicons generated via a frequency based probabilistic approach, simulating individual differences in vocabulary. We found that these lexicons tended to overlap more in the higher words as ranked by a given graph-theoretic measure, demonstrating a relationship between measures such as degree-centrality and frequency. In addition, we found that degree was the most robust measure, that clustering coefficient was robust to error, and that closeness centrality was robust to inversions.

Additionally, we were interested in changes in the mental lexicon as more words were acquired, which we looked at using data from adults' estimates of when they learned particular words, as well as frequency. We found that degree centrality and closeness centrality were similarly robust, and that clustering coefficient was less robust to error. We also found that the phonological network densifies as new words are learned, especially for the early learning stages. This suggests that the phonological network might be growing in a way such that it is optimally structured for search. This is logical because we know that lexical retrieval process is robust and fast from psychology research.

Our findings suggest that degree centrality, clustering coefficient, and closeness centrality are all relatively robust measures, and that degree centrality is the most robust of the three measures. This suggests that findings from prior work based on the phonological network are generalizable to individuals, especially in the most common case when only degree centrality was considered. In terms of future work, we are interested in comparing the English phonological network to other languages. We are also interested in looking at other types of error in the graph, such as changes in the edges based on individual differences in word pronunciation. We would also be interested in quantifying phoneme confusibility and analyzing the phonological network as a weighted graph. Finally, we are interested in looking at the robustness of other graph-theoretic properties of the phonological network.

Acknowledgments

We would like to thank our advisor, David Liben-Nowell for his guidance throughout this project. We would also like to thank Violet Brown and Julia Strand for their help and insight.

References

Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social networks*, *28*(2), 124–136.

Hollich, G., Jusczyk, P., & Luce, P. (2002). Lexical neighborhood effects in 17-month-old word learning. *Proceedings*

of the 26th Annual Boston University Conference on Language Development (pp. 314-323).

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007, March). Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, *1*(1). Retrieved from http://doi.acm.org/10.1145/1217299.1217301 doi: 10.1145/1217299.1217301

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

Storkel, H. L. (2004). Do children acquire dense neighborhoods? an investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(2), 201221. doi: 10.1017/S0142716404001109

Tsugawa, S., & Ohsaki, H. (2015). Analysis of the robustness of degree centrality against random errors in graphs. In *Complenet* (pp. 25–36).

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, *51*(2), 408–422.

Vitevitch, M. S., Ercal, G., & Adagarla, B. (2011). Simulating retrieval from a highly clustered network: Implications for spoken word recognition. *Frontiers in psychology*, *2*.

Wei, W., Joseph, K., Liu, H., & Carley, K. M. (2015). The fragility of twitter social networks against suspended users. In *Proceedings of the 2015 ieee/acm international conference on advances in social networks analysis and mining 2015* (pp. 9–16). Retrieved from http://dl.acm.org/citation.cfm?id=2809316