Network Modeling of the Mental Lexicon: Phonological Links Within and Between Communities

Jennifer Anne Gerometta
Graduate Center, City University of New York

How does access to this work benefit you? Let us know!
Follow this and additional works at: http://academicworks.cuny.edu/gc_etds
Part of the Communication Commons

Recommended Citation
Gerometta, Jennifer Anne, "Network Modeling of the Mental Lexicon: Phonological Links Within and Between Communities" (2015). CUNY Academic Works.
http://academicworks.cuny.edu/gc_etds/940

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Graduate Works by Year: Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
NETWORK MODELING OF THE DEVELOPING MENTAL LEXICON:

PHONOLOGICAL LINKS WITHIN AND BETWEEN LEXICAL COMMUNITIES

by

Jennifer Gerometta

A dissertation submitted to the Graduate Faculty in Speech-Language-Hearing Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2015
This manuscript has been read and accepted for the Graduate Faculty in Speech-Language-Hearing Sciences in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy

Valerie Shafer

______________________________

Date Chair of Examining Committee

______________________________

Klara Marton

Date Executive Officer

______________________________

Michelle MacRoy-Higgins

Klara Marton

Franzo Law

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK
Abstract

NETWORK MODELING OF THE MENTAL LEXICON:
PHONOLOGICAL LINKS WITHIN AND BETWEEN
LEXICAL COMMUNITIES

by
Jennifer Gerometta

Adviser: Valerie Shafer

Graph theory is a branch of mathematics that is used to study networks. Recently, graph theoretic techniques have been embraced by the cognitive sciences, and used to study the developing lexicon, semantic memory, and first and second language organization (Carlson, et al., 2011, Kennet et al., 2011, Wilks & Meara, 2002, Zareva, 2007) Graph theory can give valuable insight into the underlying phonological structure of language. Studying phonological networks contributes to our understanding of how the mental lexicon develops, and results of experimental studies on lexical processing can be used to test whether the proposed network structure is plausible. The goal of this dissertation was to examine the organization of a lexicon of words from a storybook corpus in terms of phonological properties. This goal was achieved by using graph theory techniques.

Two networks were defined for graph theoretic analysis. Different metrics were used to define the edges for these two networks to model different organization of neighbors of the developing mental lexicon. Word types from storybooks frequently read
to 2-4 year-old children were represented in both networks. Using graph theoretic techniques, degree centrality and betweenness centrality measures were calculated for both networks. Words that represented nodes of the network with high degree and high betweenness centrality were examined. Age of acquisition, word frequency, and measures of phonotactic probability were calculated for these prominent nodes in both networks. Comparisons of lexical and sublexical characteristics for words that represented high degree and high betweenness centrality nodes were made. Comparisons were also made between the general structures of both networks related to word categories (function and content) and morphological complexity.

Results of this dissertation indicate that the words (nodes) that hold prominent positions in these two differently-defined networks are not identical, nor are their connections. Differences were also evident in lexical and sublexical characteristics of words that represent prominent nodes within each network. The two networks also revealed different features in overall connections between words (nodes) and word types. Implications regarding how this reflects child language development is discussed.
Acknowledgments

With gratitude to my devoted family, colleagues and mentors, this great odyssey has come to an end.
# Table of Contents

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
</tbody>
</table>

## CHAPTER 1: INTRODUCTION

- Graph theory ................................................. 2
- Properties of graph theory networks .......... 3
- Parameters for linking nodes in the graph theoretic network models of the mental lexicon .............. 10
- Graph theory and the phonological dimension of the mental lexicon ........................................... 14
- Lexical organization in terms of phonology ................................................................. 18
  - Neighborhood density ................................. 18
  - Phonotactic probability ............................... 19
  - Age of acquisition ....................................... 23
  - Word frequency ........................................... 24
- Representation of function words in the mental lexicon .............. 25
- Representation of morphological properties in the mental lexicon 27
- Summary .......................................................... 28
- Research questions ........................................... 29

## CHAPTER 2: METHODS

- Participants .................................................. 33
- Procedures ..................................................... 34
  - Child literature corpus ................................. 34
  - Participant and group analysis .......................... 35
Rank analysis of lexical and sublexical characteristics ...... 35
Algorithms for network construction.......................... 37
Network Analysis.................................................. 40
Algorithms for network analysis.............................. 40

CHAPTER 3: RESULTS

Centrality measures in phonological networks............. 45
Lexical and sublexical characteristics in prominent nodes..... 47
Age of acquisition............................................. 48
Word frequency ................................................. 49
Positional segment average................................. 51
Initial biphone frequency............................ 52
Occurrence of function and content words ............... 53
Morphological complexity............................... 54

CHAPTER 4: DISCUSSION

Centrality measures and lexical and sublexical characteristics ...... 60
Age of acquisition............................................. 60
Word frequency ................................................. 63
Positional segment average................................. 65
Initial biphone frequency............................ 68
Frequency of function words and morphologically complex words........... 70
Comparison of two network models.......................... 73
Network connectivity........................................ 73
Summary.......................................................... 78
Clinical implications......................................... 82
Future directions ............................................. 82
Appendix.......................................................... 84
Example words for high and low prominence nodes........84
Example words for lexical and sublexical characteristics ..86
Comparison of links to cat........................................88
Comparison of links to straw...................................89

References........................................................................90
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Participant information ........................................33</td>
</tr>
<tr>
<td>2</td>
<td>Number of words with high and low betweenness ...............37</td>
</tr>
<tr>
<td>3</td>
<td>Hermits and islands in one-phoneme difference and 75% similarity networks ........................................47</td>
</tr>
<tr>
<td>4</td>
<td>Occurrence of function words in literature corpus ............54</td>
</tr>
<tr>
<td>5</td>
<td>Morphological complexity in one-phoneme difference and 75% similarity networks ........................................55</td>
</tr>
<tr>
<td>6</td>
<td>Summary of descriptive statistics for lexical and sublexical characteristics of prominent nodes ....................57</td>
</tr>
<tr>
<td>7</td>
<td>Comparison of significant network characteristics .............80</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Entire network and giant component</td>
</tr>
<tr>
<td>2</td>
<td>Example node <em>rain</em> with high degree centrality</td>
</tr>
<tr>
<td>3</td>
<td>Example node <em>water</em> representing in-degree and out-degree links</td>
</tr>
<tr>
<td>4</td>
<td>Visualization of the one-phoneme difference network</td>
</tr>
<tr>
<td>5</td>
<td>Visualization of the 75% similarity network</td>
</tr>
<tr>
<td>6</td>
<td>Examples of high degree centrality and betweenness centrality</td>
</tr>
<tr>
<td>7</td>
<td>Overview of present network analysis</td>
</tr>
<tr>
<td>8</td>
<td>Example node <em>milk</em> with average degree centrality</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Young childhood is a period of remarkable language growth. How children represent and process words is an area of great interest to those concerned with typical and atypical language development. Current research demonstrates that lexical factors such as age of acquisition, word frequency, and neighborhood density influence spoken word access in young children (Garlock, Walley, & Metsala, 2001, Walley, 1993). Sublexical characteristics, such as phonotactic probability, have also been shown to influence word recognition in young children (Woodley, 2010). One way to understand the vocabulary young children have access to is to examine storybook literature samples. Literature samples have long been considered to be representative of the receptive lexicon (although low frequency words may be overrepresented in literature as compared to spoken language; Dickinson et al., 2001, Dollaghan 1994, Coady & Aslin, 2003, Kucera & Francis, 1967). Studying lexical and sublexical characteristics of words found in children’s literature and examining the connections between those words in real world models of the developing mental lexicon represented by a storybook corpus will shed light on the organization of the developing mental lexicon and spoken word access.

The goal of this dissertation is to examine the organization of a lexicon of words from a storybook corpus in terms of lexical and phonological properties. This goal will be achieved by using graph theory techniques. Graph theory techniques give valuable insight into the structure and function of the mental lexicon and the underlying phonological structure of language. Studying lexical and phonological networks
contributes to our understanding of how the mental lexicon is organized, grows and changes. Network representation using graph theory has an advantage over traditional modeling because real-world connections in an existing corpus can more effectively be examined. In addition, network representation using graph theory demonstrates connections and allows for analyses that are not available through other approaches.

**Graph Theory**

Graph theory has been used in the fields of computer science, economics, and the biological sciences to study connections in the World Wide Web, social networks and labor markets, and properties of cells within an organism. Only in recent years have the cognitive sciences embraced graph theoretic approaches to model systems. Graph theoretic techniques have been used to model properties (e.g., phonological or semantic properties) that result in patterns of organization that account for results demonstrated in experiments examining processing and organization in the mental lexicon. For example, Carlson et al. (2011) examined phonological connectivity (global structure) of words in children’s developing lexicon in comparison to adult lexicon. First and second language lexical organization (Wilks & Meara, 2002, Zareva, 2007), structure of phonological networks across languages (Arbesman, Strogatz & Vitevitch, 2010), and the influence of the structure of the lexicon on word learning, production and recognition (Chan & Vitevitch, 2009, 2010, Vitevitch, 2008) have all been examined using graph theoretic approach.

Graph theory is a branch of mathematics that is used to study relationships between objects. As applied to the mental lexicon, graph theory is used to examine these
relationships using networks in which words are nodes and the connections between nodes are defined by a given relationship (e.g., phonological or semantic). Local connections are defined, but then the global structure of the system can be examined. The behavior of the system at the global level is not always predictable from the local relationships, even if these local relationships are quite simple (Vitevitch, 2008). The organization of the system structure at the global level can be examined to gain insight on the relationship between structure and function of the mental lexicon.

Properties of Graph Theory Networks

In graph theory, a number of statistical properties (graph theoretic properties) of the network are used to describe the network organization. These statistical properties can be used to generate hypotheses concerning lexical processing. In turn, results of experimental studies on lexical processing can be used to test whether the proposed network structure is plausible. Graph theory can also be used to directly compare the consequences of different phonological metrics on lexical organization.

Networks in graph theory are made up of nodes, or vertices (e.g., the words of the lexicon) connected by edges, or links (Barabasi, 2002). These edges are defined by the parameters set for each network, for example, the phonological similarity of the words in the lexicon. Edges in networks can be directed or undirected. Nodes linked by an edge in an undirected network are each neighbors to the other in a bidirectional relationship similar to a two-way street. For example, mad, add, pad, and had are all neighbors to dad, and vice versa. A node linked by an edge in a directed network represents a connection in one direction, such as a one-way street. For example, 75% of the phonemes in grain are in strange, but only 50% of the phonemes in strange are in grain.
A *community* is another graph theoretic property of networks that can provide insight to the researcher. A community is a region of nodes that are very densely linked, or highly connected to one another, and, thus, are defined as a group belonging to a community. Nodes are considered part of a community if they are more highly linked to other nodes within their community than to those nodes that exist outside their community (Barabasi, 2002).

The *giant component* is the sum of all highly interconnected *communities* within the network. Some nodes that are not part of the giant component, *islands* (small groups of nodes that are interconnected but do not connect to the network outside of the island) and *hermits*, (nodes that have no connections) are by definition excluded from the giant component. In previous studies examining the phonological structure of the mental lexicon, the giant component has been the focus (Arbesman, Strogatz & Vitevitch, 2010, Chan & Vitevitch, 2010, Siew, 2013, Vitevitch 2008). The types of words that are excluded from the giant component (i.e., those in islands or hermits) are typically multisyllabic words sometimes found in children’s language or children’s books such as *chimney*, and *treasure*. Figure 1 shows an example of how a network is organized with islands and hermits and without islands and hermits (giant component).
Two additional properties are *degree centrality* and *betweenness centrality*. These *centrality measures* are used to indicate the importance of a node based on its relation to other nodes in the network. Nodes that have the highest *degree centrality* have the highest number of connections (edges or vertices) incident upon them. For example, the word *rain* has a high degree centrality in a model based on phoneme similarity. Edges incident upon this node would include ties to words such *race, ran, rainy, wren*, and *vane*. See Figure 2.
**Figure 2.** Example node *rain* with high degree centrality in the one-phoneme difference network.

In an undirected network, degree is measured simply by the number of connections between words. In a directed network, degree is measured in two ways: *in-degree* and *out-degree*. *Out-degree* is measured as the number of words pointing *from* the head word to other words. *In-degree* is measured as the number of words pointing *to* the
head word. For example, in Figure 3 below, the head word is water. Water points to other words in which 75% (or more) of the word water is present, such as watermelon, watermelons, watering, waters and daughters. These words are connected through out-degree of the head word water. Words that represent in-degree in this graph are ought and aww. These words point to water.

Nodes with high betweenness centrality are strongly connected to different communities, acting as a bridge between communities. This bridge exists on the shortest path between two nodes in different communities within the network. The node dad could act as a bridge between the neighbors add, mad, bad, sad, had, and pad to neighbors within a community represented by did, dead and died (Figure 6). These prominent node types hold special positions within a network. Importantly, analysis of nodes that link words between communities is not possible to examine without engaging in network analysis.

Nodes with high degree centrality spread information quickly among nodes within a community. Nodes with high betweenness centrality share information among communities. Both of these types of prominent nodes create cohesion within the network (high degree nodes create cohesion within the community, high betweenness nodes create cohesion within the network). Network representation using graph theory is necessary to investigate both of these prominent nodes.
Figure 3. Example node *water* with edges representing in-degree and out-degree links in a 75% phonological similarity network.

Although it has been shown that there is a moderately strong correlation ($r=.7$) between these two measures (degree and betweenness centrality) these measures look at two distinctly different functions in a network (Valente et al., 2008). Both measures are based on an adjacency matrix, representing which nodes of a graph are adjacent to which other nodes, however, the underlying data is subjected to different computations (representing degree centrality and betweenness centrality, see below). Both measures of high degree and high betweenness represent prominence in the network, but in different
ways. Nodes with high degree or high betweenness centrality have the potential to influence neighboring nodes. Nodes with high degree centrality can quickly transmit information and influence other nodes within their community through direct or short paths to neighboring nodes. A node with high betweenness centrality can influence the spread of information by facilitating or hindering communication between nodes in different communities. (Freeman, 1979, Newman, 2003, Valente, 2008). These two types of prominent nodes represent different levels of connection in the network. Nodes with high degree centrality represent the micro/word-in-community measure of the network. Nodes with high betweenness centrality represent the macro/word-outside-community connection measure of the network. Both levels of connection are represented by words with different lexical and sublexical characteristics. It is these lexical and sublexical properties of words that hold these positions in the network (e.g., high degree, high betweenness centrality), and may have distinct effects on phonological and lexical processing.

Graph theoretic models suggest that the mental lexicon can be modeled as a structured network that can be defined by growth and preferential attachment (Arbesman et al., 2010, Barabasi, 2002, Kapatsinski, 2006). Growth refers to the addition of nodes to the existing network. For example, as a child learns new words, representations for those words are added to the lexicon. These additional words represent growth of the network. Preferential attachment refers to the increased likelihood that new nodes will form vertices with nodes that are the oldest and already have many connections – the “rich get richer” phenomenon. The network of webpages on the World Wide Web is a common example. The most popular pages
are most often linked to other pages in the network. These pages continue to grow in popularity, and gain more links and hits. These are interesting properties because they are consistent with what we know about lexical development. Specifically, we can continue to add words to the lexicon across the lifespan and that early-learned words appear to have a special status based on preferential attachment (Allen, 1992, Garlock et al., 2001, Metsala, 1999).

Parameters for Linking Nodes in the Graph Theoretic Network Models of the Mental Lexicon

Graph theory can be useful for examining how the mental lexicon might be organized (Arbesman et al., 2010, Chan & Vitevitch, 2010, Chan & Vitevitch, 2009, Goldstein & Vitevitch, 2014, Kapasinsky, 2006, Siew, 2013, Vitevitch & Goldstein, 2014, Vitevitch et al., 2011, Vitevitch, 2008). Graph Theoretic techniques can be used to analyze the network of the mental lexicon to represent the overall structure of the phonological system. In addition, these techniques can help examine the connections between words in terms of their lexical and sublexical properties. Modeling the mental lexicon in the present paper is framed upon existing metrics of lexical organization.

One such metric used to define the connections between words is phonological similarity. The one-phoneme metric outlined by Luce and Pisoni (1998) is the most commonly used phonological metric. In this model, phonological word forms are associated or considered neighbors if they are the same, except for one phoneme addition, deletion or substitution. Words that differ by more than one phoneme would not be
considered neighbors. This metric primarily links monomorphemic words as neighbors, as different words with more than one syllable tend to vary by more than one phoneme, with the exception of inflectional morphemes such as the plural /s/ and past tense /ed/, that consist of one phoneme additions. A neighborhood structure of the metal lexicon defined by this metric is supported in the literature. In recent studies, this one-phoneme metric is commonly used to define neighbors when examining attributes or processing of the mental lexicon. This metric has been demonstrated to be a valid way of assessing phonological similarity (Luce & Large, 2001). Among other findings, Luce and Large reported that words that were commonly produced following a primed word were rhymes, or words that differed by only one phoneme. In the speeded word production experiment with adult participants, 71% of the responses to nonwords stimuli were words that differed from those stimuli by just one phoneme, suggesting that the one-phoneme neighborhood metric captured an important property of lexical organization in this experiment.

In a series of behavioral experiments, Vitevitch and Goldstein (2014) found that keyplayers (words that hold an important role in linking the lexical network) identified in the structure of Vitevitch and Goldstein’s network model of the mental lexicon were responded to more quickly and accurately than comparable words (in terms of word length, word frequency, neighborhood density, and phonotactic probability) that were not keyplayers. These researchers constructed a network using the one-phoneme rule. Specifically, words identified as keyplayers were identified more quickly and accurately in white noise as compared to word foils. Those words were also identified more quickly and accurately during an auditory naming task and an auditory lexical decision task.
These findings corroborate the network model based on the one-phoneme rule because these key words held prominent positions in the network as keyplayers based on statistical properties derived from the graph model. In other words, their status in the model was based on the one-phoneme rule rather than having been provided with any special status (preferential treatment) in the input. These findings support the assumptions about phonological organization that underlie the model used by Vitevitch and Goldstein.

A more recent metric suggests that words may be organized in terms of percent of phonological similarity (Kapatsinski, 2006). Words composed of phonemes with 75% similarity may be grouped together as neighbors. This metric tends to link words with one and two syllables, as words with more than two syllables in length tend to vary in phonemes by more than 75 percent, with the exception of some word derivations for example, *blackberry* and *blackberries*. Confusability studies (the chance that one word will be misheard as another), naming tasks, lexical decision tasks, and familiarity judgment tasks support a model of neighborhood organization such as this one (Kapatsinski 2006, Coble & Robinson 1992, Kidd & Watson, 1992). For example, in confusability studies, Kidd and Watson (1992) found that the sequence or pattern is the element that establishes confusability of sounds within the sequence, and that the total duration of the sound production (tones) impacted the confusability of two sounds during sound discrimination experiments. This supports the proposal for the 75% similarity metric, in that associations between words could be drawn based on percentage of phonological similarity between words. Kapasinski (2006) authored another study that supports the use of this metric. The researcher examined reaction times (RTs) to lexical decision and naming tasks for 40,000 words from 1200 subjects, which were both
available through the English Lexicon Project repository (Balota et al., 2007). RTs were obtained from 13,458 of the nearly 20,000 words in Kapatsinski’s corpus. The RTs were successfully predicted for words in high versus low density neighborhoods based on Kapatsinski’s 75% similarity framework. These studies demonstrate that the 75% similarity network is also supported as an organizing principle of the mental lexicon.

Differences between the number of isolates (islands and hermits) and differences in word length are two important points of comparison between the two networks. In one previous study comparing network models of the mental lexicon, 58% of the lexicon consisted of hermits (words with no neighbors) when modeled using the one-phoneme metric, as compared to 7% of the lexicon consisting of hermits using the 75% similarity metric (Gruenenfender & Pisoni, 2005, Kapatsinski, 2006). Longer words are more likely to have fewer neighbors in both metrics, but to a much greater degree in the one-phoneme similarity metric (Kapatsinski, 2006). These comparisons highlight differences between the two metrics.

The metrics of lexical organization used in this study to model the mental lexicon are derived from the shared principles of established models of lexical retrieval such as the Neighborhood Activation Model (NAM), Cohort, and TRACE, so-called because processing units form a working memory structure called the trace (Luce & Pisoni 1998, Marslen-Wilson, 1987, McClelland & Ellman, 1986). Although these are different models, they are similar in their assumption of phonological input activating word forms. In all of these models, candidate words are activated and compete with one another until a match is found for the input word (lexical competition). The graph theoretic structural models of the organization of neighborhoods in the mental lexicon based on the above
metrics are consistent with these theoretical models of the lexical retrieval process. Graph theory modeling can be used to test these models.

**Graph Theory and the Phonological Dimension of the Mental Lexicon**

Graph theory is becoming a methodology more commonly used in the cognitive sciences. In speech and language science, graph theory has been used to study the acquisition of words in typically developing children (Carlson et al., 2011), and phonological relationships among words in different languages (Arbesman et al., 2010, Vitevitch, 2012). This dissertation examines the phonological dimension of the organization of the mental lexicon in English. A few studies have examined this using graph theoretic techniques.

Vitevitch (2008) used graph theory to examine factors that describe the network structure in a network comprised of approximately 20,000 word types based on the one-phoneme metric. He examined three factors; *average path length* is the mean number of links it takes to connect any two nodes in the network, *clustering coefficient* is the probability that two neighbors of a given node are also connected, and *assortative mixing by degree* is the probability that highly connected nodes are connected to other highly connected nodes. Focusing on the giant component of the network, Vitevitch found that the adult lexicon exhibits *small world* characteristics (low average path length and high clustering coefficient). These characteristics indicate that transfer of information from one node to the next is rapid and efficient across the network. Thus, this network model is consistent with what we know about lexical retrieval within the mental lexicon. Vitevitch’s network also showed positive assortative mixing by degree. This property suggests only a portion of the lexicon will be activated when phonetic information is
received, further suggesting that this network is efficient. Some criticism of these findings came from Gruenenfelder & Pisoni (2009). They found that these same results could be derived from three different pseudolexicons, with the same CV frequency of occurrence of words two to five segments in length as the corpus used in Vitevitch (2008). These researchers conclude that Vitevitch’s results occurred based on the phonetic overlap of words in the giant component of the network rather than evidence of an efficient network. This indicates that structural components of the network occur naturally when phonetic overlap exists, and can be found in pseudolexicons when the same parameters for edges of the network are set.

In a behavioral experiment based on results from Vitevitch’s 2008 network, Chan and Vitevitch (2009) examined RTs for words that were represented in the graph theoretic network as having high or low clustering coefficients. These words, controlled for phonotactic probability (the frequency with which speech sounds occur in a language, see definition below), were presented as stimuli. In one experiment, words were presented to participants in noise, and participants were asked to type the word they heard. In the second experiment, words with either high or low clustering coefficient were presented, as well as nonwords. Participants were asked to decide whether or not the stimulus word was a real word. RTs in both perceptual identification tasks were assessed. RTs were slower for words with high clustering coefficients in both experiments. Results of this study suggest that the structure of the graph theoretic network models the structure of the mental lexicon, and clustering coefficients of words in the lexicon represent a factor that affects lexical retrieval time.
In a study by Goldstein and Vitevitch (2014), researchers suggest that characteristics of individual words do not alone influence processing, and the relationships that exist among lexical neighbors and their connections in the network must be examined to fully understand lexical processing. They demonstrated clustering coefficient of nonwords was a factor in learning of nonwords. Using a word learning methodology, adult participants were trained once a week for three weeks to match nonwords to nonobject pictures. After three trainings and a one-week retention period, more nonwords-nonobject pairs were matched if the nonwords (CVC) exhibited high cluster coefficients.

Siew (2013) demonstrated that real phonological networks exhibit characteristics that are not found in randomly associated networks, or random associations between words in networks. She examined the community structure of real and random phonological networks. She compared the lexical characteristics of word length, word familiarity, word frequency, neighborhood density, positional probability, biphone probability and age of acquisition of word nodes in a real network based on the one-phoneme metric, and a random network. The random network was made up of words from the real community randomly assigned to nodes. In both the real and random networks, 17 communities were compared. When making comparisons between the two networks (real and random), communities in the real network differed significantly from one another in terms of all of the above lexical characteristics, however, the 17 random communities did not. This implies that communities identified by the community detection algorithm in graph theory are capturing important relationships among words in the real phonological network. A robust community structure exists in the phonological
network, and the organization of the model network of the mental lexicon may facilitate lexical retrieval mechanisms.

Most studies have used a one-phoneme metric, but this metric has limitations because of the number of isolates in the network, and lack of connections between words with more than three phonemes (one morpheme). Kapatsinski (2006) suggested the use of a different metric, as discussed above. In this metric, nodes are linked by 75% similarity. Using the same corpus as previous studies, he found that the network did not exhibit small world characteristics. The network did demonstrate high clustering coefficient (found in small worlds), but also high average path length. Kapatsinski argues that low average path length (necessary or small worlds) is not needed for efficient lexical retrieval. Lexical search involves activation of neighbors of words that share sublexical similarity (the phonological components that constitute a word). Lexical search does not involve activation of all words in the entire lexicon (Marslen-Wilson, 1990), therefore a high average path length is not necessary for efficient retrieval.

The above studies use various algorithms of graph theory to measure the structure of word-forms in the mental lexicon. No known studies have used graph theoretic techniques to examine the relationships of words that exhibit high degree and high betweenness in a network of the mental lexicon. No known studies have compared lexical and sublexical characteristics of prominent node types (high degree and high betweenness) in the developing lexicon based on a corpus of children’s literature. Also, no known studies currently exist comparing prominent node types and lexical properties of words in the mental lexicon using a 75% phoneme similarity metric, and a one-phoneme difference metric. Examining these prominent nodes based on a storybook
corpus in two different network models will inform the research community how information is efficiently transferred within and between network communities. It will also serve to inform how these networks (one-phoneme difference and 75% similarity) may differ in efficiency and searchability of the developing mental lexicon. These comparisons will be made in the present dissertation.

**Lexical Organization in Terms of Phonology**

Certain lexical properties (age of acquisition, word frequency, phonotactic probability) have been associated with word learning, lexical retrieval, memory tasks, and literacy development (MacRoy-Higgins et al., 2012, Metsala et al., 2007, Sosa, 2012, Storkel, 2001, 2006, Vitevich and Luce, 2005, Zamuner et al., 2004, Zamuner, 2009). These lexical properties, when examined in the developing mental lexicon, could help researchers understand the organization of the mental lexicon and lexical access. Understanding network connections in the mental lexicon could shed light on how these characteristics influence word learning, lexical retrieval, memory tasks, and literacy development.

**Neighborhood Density**

The neighborhood is a structure that is commonly examined in psycholinguistic studies, and neighborhood analysis is at the root of the present study. Two metrics are compared, each defining neighborhoods differently. In this dissertation, neighborhoods have been defined in two different ways; however, phonological neighborhoods, as defined using either metric, influences lexical access.
A phonological neighbor is often defined as a word (or words) differing from a given word by one phoneme substitution, deletion, or addition (Luce and Pisoni 1998). This metric has been used in numerous studies examining phonological processing. Results of these studies demonstrate that neighborhood density influences word recognition, word production and word learning (Garlock, Walley & Metsala, 2001, MacRoy-Higgins et al., 2012, Sosa, 2012, Storkel, 2001, 2006, Vitevich and Luce, 2005, Zamuner et al., 2004, Zamuner, 2009, Metsala, 1999, Hollich et al., 2002). For example, in a pseudo-word learning task, MacRoy-Higgins and colleagues (2014) found that typically developing children who learned novel pseudo-word labels for unfamiliar objects were able to pronounce and detect mispronunciation of pseudo-words with high neighborhood density better than those words with low neighborhood density. This demonstrates that typically developing children are sensitive to neighborhood density during word learning.

Phonotactic Probability

One lexical property examined in the analysis of neighborhoods (regardless of which metric is used) is phonotactic probability. Phonotactic probability is a statistical property found in the speech signal, and used during the development of language. Phonotactic probability, as defined by Trask (1996), refers to the frequency of occurrence of a particular arrangement of phonemes in the words of a given language. For example, the probability of the phoneme /s/ preceding /t/ in a word in English is higher than the probability of /l/ preceding /s/.
Knowledge of phonotactic probability in one’s native language has been linked to processing during both perception and production of speech and language. Making use of phonotactic probability during speech and language processing has been specifically associated with speech signal segmentation and processing speed (Jusczyk, 2000, Mattys & Jusczyk, 2000, Mattys et al., 1999, Saffran et al., 1996, Bonte et al. 2006, Luce & Pisoni, 1998), word repetition (Storkel, 2001, 2003, Zamuner, 2009), word learning, speech production (MacRoy-Higgins et al., 2012, 2014, Munson 2005), and metalinguistic processing (Gross et al., 2000).

In adults, Vitevitch and Luce (1998, 1999) have shown that phonotactic probability affects lexical access. They have investigated lexical access and phonotactic probability in studies using RTs as the dependent measure. In a single-word shadowing task, these researchers have compared RTs to words with high (dense lexical neighborhood) and low (sparse lexical neighborhood) probability phonotactic sequences, and nonwords with high and low probability phonotactic sequences. In these studies, RTs to words with high probability phonotactic sequences (and from densely populated lexical neighborhoods) were slower than RTs to words with low probability phonotactic sequences. When nonwords were used as stimuli, the opposite results occurred. RTs were faster to nonwords with highly probable sequences, and slower to nonwords with sequences of low probability. Vitevitch and Luce attribute this result to differences in lexical versus sublexical processing. Lexical processing demonstrated inhibitive effects, increasing RTs (i.e. slowing access) to words with highly probable phonotactic sequences due to lexical density (dense neighborhoods). Sublexical processing (accessed when nonwords were used as stimuli) revealed facilitative effects, reflected as shorter RTs to
nonwords with highly probable phonotactic sequences. In these studies, the level of processing (lexical or sublexical) dictated the response time (lexical processing induced by “real word” stimuli, and sublexical processing induced by nonwords stimuli). Vitevitch (2005) found facilitative effects in words with high neighborhood density (ND) during speech production tasks. In picture-naming tasks, fewer errors were produced for words with high ND as compared to words with low ND. Words with high and low ND were controlled for PP, so that PP was equivalent in both groups. At face value, findings of these studies seem to be at odds with one another. In Vitevitch (1998), high PP demonstrated facilitative effects. In Vitevitch (2005), while controlling for PP, ND demonstrated facilitative effects. Instead of conflicting, however, these studies demonstrate that these two processes, sublexical (represented by phonotactic probability) and lexical, (represented by neighborhood density) operate at different levels of representation. Although operating at different levels of representation, bidirectional connections are evident between phonological and lexical segments. It is these bidirectional relationships at these two levels of representation that can be closely examined using graph theory.

An interesting question is whether children rely on perception of sublexical information during lexical access as adults do (Woodley, 2010). Two studies suggest that children treat all incoming stimuli as lexical (Woodley, 2010, Pierrehumbert, 2003). For example, Woodley measured RT of preschoolers who judged “sameness” of words and nonwords. She examined how the factors of lexicality, phonotactic probability and neighborhood density influence access time. High phonotactic probability (PP) did not facilitate nonword access (as it did in adults for Vitevitch and Luce), and slower RTs
were observed for both words and nonwords with high phonotactic probability/high density neighborhoods. Woodley attributes these results to the likelihood that children process all incoming stimuli at the lexical level. This notion is supported by Pierrehumbert (2003). These studies suggest that strong phonological categories are developed as a child’s vocabulary grows. Sublexical contrasts between lexical entries are established and recognized, only after these phonological categories are established.

Storkel and Rogers (2000) also studied how phonotactic probability and neighborhood density affect lexical acquisition in children. In their interpretation of results of word learning tasks, they found that sublexical processing was essential in establishing an initial mental representation of new words. Participants (seven to 13-year-old children) participated in a delayed nonword recognition task. Following presentation of nonwords matched to pictures, the authors found that more high probability nonwords were learned than low probability nonwords. The authors concluded that the results support the hypothesis that sublexical processing is the dominant level of processing in lexical acquisition in children, and integral in establishing the initial mental representation. In a later study with younger children, Hoover, Storkel, & Hogan (2010) found that words with few neighbors and low phonotactic probability create an optimal condition for word learning in three- to five-year old children. During familiarization tasks (storybook reading) children were introduced to nonwords that varied in phonotactic probability and neighborhood density. After familiarization, children engaged in picture-naming tasks for CVC nonwords presented in the story. More nonwords with low phonotactic probability were learned as compared to nonwords with
high phonotactic probability. The authors concluded that these younger listeners identify unique sounds as novel, and this triggers word learning.

The results of these studies (Hoover, Storkel, & Hogan, 2010, Pierrehumbert, 2003, Storkel & Rogers, 2000, Woodley, 2010) demonstrate two levels of lexical processing, the lexical level and the sublexical (phonological) level. In a two-representation model, lexical and phonological levels interact and activate one another. Lexical activation typically dominates, as we typically perceive real words. When words are not perceived as real (or unknown) phonological activation dominates lexical processing (Storkel & Morrissette, 2001).

**Age of Acquisition (Word AoA) of Lexical Items**

The effects of word AoA on lexical processing has been well studied. Among other tasks, research has focused on word AoA in picture naming, word naming, object recognition and lexical decision tasks. Words that are acquired earlier have been found to be processed more quickly than words that are acquired later in life. For example, Brysbaert and colleagues studied word AoA in semantic processing tasks (2000) in college-age adults. In a word association task, 144 words were presented to participants. Words were separated into categories of either early or late acquired words (kindergarten teachers judged whether or not words would be known by their students) and controlled for word frequency. RTs were faster to early-acquired words, as compared to words acquired later in language development. In preschool children, Garlock, Metsala and Walley (2001) found that word AoA facilitated word recognition and production for preschool-age children. In both word repetition and gating tasks, preschoolers performed
more accurately when words with early AoA were presented as compared to later AoA. These children required less input to recognize early words in the gating task, and repeated words with early AoA more accurately. These finding supports the notion that word AoA influences lexical access.

**Word Frequency**

Frequency effects are a crucial aspect of language acquisition (Ellis, 2002). Adults and children identify, discriminate, name, and recall words faster if they frequently occur in the native language (spoken-word recognition). Even in the cohort model of speech recognition (Marslen-Wilson, 1990), words are activated not only by the sounds that occur in those words, but also the frequency with which words occur. Higher frequency words tend to get more activation. Both adults and children are sensitive to the frequency with which words occur, and this influences word processing. In Chambers and Forster’s classic reaction time study (1975), adult participants were instructed to judge pairs of words to be the same or different. Words were high frequency, low frequency, legal or illegal word forms (based on phonotactics). RTs in this study were faster for high frequency words than low frequency words or nonwords. In children, word frequency has been found to have a facilitative effect in a picture-naming task (Cirrin, 1983). Picture naming latencies were shorter for pictures representing frequent words in children in kindergarten, first and third grade. These findings have been supported in many studies (Allen et al., 1992, Bybee et al., 2001, Hogan et al., 2011).

According to Zipf’s Law (1955), the frequency of any word is inversely proportional to its rank in a given frequency table, and that relationships between words
in a corpus represent a power law distribution, with few words having many occurrences, and many words having just a few. For example, using the Brown Corpus, Zipf found that the most frequently occurring word in this corpus, *the*, accounts for 7% of all word occurrences, the next word, *of*, accounts for 3.5%. Zipf’s observations account for a power law distribution in a network. Although different network models have found the mental lexicon to have a truncated power law distribution (Vitevitch et al., 2008), or a parabolic distribution (Kapatsinski, 2006), a “rich get richer” point of view still remains in both accounts.

**Representation of Function Words in the Mental Lexicon**

Researchers often separate words into the categories *function* and *content* words. Content words (such as nouns, verbs, adjectives and adverbs) have meaning that is often denotational and can be determined even for the word in isolation. Function words are often difficult to define in and of themselves, but serve to express grammatical relationships and link content words in a phrase or sentence. Function words are defined as auxiliary verbs, conjunctions, determiners, prepositions, and pronouns. Function words occur much more frequently than content words in connected speech. Biphone frequency of words considered to be highly frequent function words (such as *the*, *you*, *of*) tend to be associated specifically with function words, resulting in traditionally-defined small neighborhoods for function words. The use of function words in connected speech has been speculated to enhance the ability to parse the speech stream, and the learning of new content words, as the use of articles such as “a” and “the” cue the child listener to prepare for a content word to follow (Selkirk, 1996).
Infant studies suggest that recognition of function word forms is highly important during language development. Recognition of function words may support lexical word recognition, development of the lexicon, phrase segmentation from the connected speech stream, and the development of syntactic classes (Shi et al., 2003).

Function words and content (lexical) words are often different acoustically and functionally. Content words are often more acoustically complex, carry stress in an utterance, and are often referential or, at least, more easily mapped onto the infant’s surroundings. Function words, on the other hand, are rarely stressed, acoustically more simplified, often cliticized, and reduced, or minimized during production. As they are not referential, their meanings are often more abstract than their lexical counterparts (Selkirk, 1996).

Recent studies have indicated that infants and toddlers are attuned to function words in the speech stream (Shafer et al., 1998, Shi et al., 2003). In an ERP study by Shafer et al. (1998), 11 month old infants discriminated differences in typical vs. atypical function words embedded in a story. In Shi et al.’s study (2003), 13-month-old infants demonstrated a looking preference to real function words vs. nonsense function words in phrases, suggesting that 13 month olds were able to recognize real function words in a phrase. These studies demonstrate infants’ ability to recognize differences in real versus nonce function words in the speech stream during experimental tasks. These findings support the claim that infants are sensitive to whether function words they hear are ungrammatical or missing. These findings suggest that it is also important to examine function words in the input to children, since they play a role in language development.
Representation of Morphological Properties in the Mental Lexicon

Representations of morphological properties of words could also help us understand relationships between words in the network of the mental lexicon. Morphophonological properties lie at the intersection between word formation and speech sound production. More specifically, morphophonology can be defined as the speech sound change that occurs when words’ morphemes combine. Morphophonological properties may also be a part of phonological bootstrapping. Under this hypothesis the phonological properties of the speech signal allow for identification of lexical and syntactic units acquisition. Phonological representations of complex and compound words in the developing mental lexicon may facilitate learning of morphological forms (Christophe et al., 1997). For example, in the words roll, knock, bee and back speech sound changes occur with the addition of the /d,t,z,s/ phonemes in the final positions to produce rolled, knocked, bees and backs respectively. A speech sound change occurred as inflectional morphemes were added in each of these examples. This phonological difference in the word representation may enable the child to make rudimentary syntactic deductions. This paves the way for more detailed representations as more exemplars are added to the lexicon. Thus, examination of complex morphophonological patterns of input is also important for understanding language acquisition. To date, studies of organization of the lexicon in terms of phonological properties have ignored morphophonology.
Summary

Constructing real-world models of the developing mental lexicon based on words found in children’s literature allows the researcher to generate predictions about organization of the lexicon in terms of phonology. Phonological similarity between words is an important factor in both networks. Which organization principle, one-phoneme or 75% metric, better reflects how the lexicon is organized in the brain? Both models can be examined to see which more closely matches experimental evidence for factors such as phonotactic probability, word frequency, word AoA, word category (function or content) or morphophonological structure.

Centrality measures of lexical networks appear to reflect lexical access. High centrality reflects faster access. Nodes that exist in these positions (high degree centrality and high betweenness centrality) may be special, or exceptional in term of the lexical characteristics of word AoA, word frequency and phonotactic probability. Network modeling can be used to generate predictions regarding the organization of the mental lexicon that can then be tested experimentally.

The Present Study

In this dissertation, the mental lexicon (based on a corpus of children’s literature) was modeled in two different ways based on phoneme similarity: The one-phoneme difference metric and a 75% phoneme similarity metric. Nodes with high degree and high betweenness centrality were compared in terms of lexical and sublexical characteristics. Additionally, words serving as prominent node types and the lexical and sublexical characteristics were compared within and between networks. Real world examples of
networks of the mental lexicon were examined to analyze micro- (within communities) and meso (between communities) level connections within the developing mental lexicon. Prominent nodes and the lexical and sublexical characteristics assigned to them were compared. These lexical and sublexical characteristics have been associated with word learning, lexical retrieval and literacy development. Previous research has suggested that these characteristics could be associated with prominent nodes. These analyses addressed the following research questions. 1) Which model, one-phoneme or 75% similarity rule better reflects child language development? Lexical properties, word AoA, word frequency, positional segment average (phonotactic probability) and biphone probability (phonotactic probability) were used to address this question. 2) How are function words and morphologically complex words organized in the one-phoneme versus the 75% similarity networks? Which model is better supported by the literature? 3) What differences exist between 75% similarity and one-phoneme models? Which is consistent with current literature related to the development and organization of the mental lexicon?

Predictions

1) In terms of lexical properties, both models could reflect child language development. I predict that in both models high degree nodes will have higher word frequency, higher positional segment averages, and higher biphone frequency compared to low degree nodes. Also, high degree nodes in both models will have lower age of acquisition as compared to nodes with high betweenness. The difference between models will show to what extent these properties are different. Greater differences between these lexical properties in high degree and high betweenness nodes will be evident in the 75% similarity model, because more nodes are interconnected in this model.
In both models, high degree centrality will be characterized by several unique features. Words with high degree centrality will exhibit lower word AoA when compared to words with high betweenness centrality. Following the organizing principles of network theory, structured networks exhibit preferential attachment. That is, new nodes will preferentially attach to nodes that are the oldest and already have many connections. Thus, words with lowest age of acquisition should be those that are highly connected within communities – those with high degree centrality, as compared to those nodes with a high number of edges connections) outside of a given community. High degree centrality nodes will also exhibit higher phonotactic probability and higher word frequency, and possibly greater representation of function words (as function words are more frequent than content words) as compared to nodes with high betweenness centrality.

In both models, nodes with high betweenness centrality will be characterized by lower word frequency, higher age of acquisition and lower phonotactic probability as compared to high degree centrality. This will reflect their role as bridges between communities, and optimize spreading activation during the process of lexical retrieval (specifically low phonotactic probability and word frequency makes functional sense for bridges, so as not to activate communities that are not neighbors to the stimulus).

2) As mentioned above, function words are highly frequent, and should be represented within communities, but not between the communities. Therefore, function words will not be represented as nodes with high betweenness centrality in either network. Although function words are highly frequent, their phonological makeup is unique. This property
may result in membership in communities that are not represented as high degree nodes in either network.

Morphologically complex words could represent a possible extension of phonological bootstrapping, serving as an entry to facilitate learning of morphology. A higher proportion of morphologically complex words may exhibit high betweenness centrality. This may constrain spreading activation of the lexicon for more efficient lexical retrieval, due to the complex nature of the words. These more complex characteristics (coupled with lower phonotactic probability) will inhibit the spread of activation in terms of their lexical and sublexical nature. Lexical complexity and low phonotactic probability, present in nodes that make links between communities, may be exploited during lexical access to 1) constrain spreading activation and 2) promote phonological bootstrapping. Nodes with high betweenness centrality will also be characterized by words with more morphophonemic markers, as compared to nodes with high degree. This will occur because complexity may be a prerequisite from a structural standpoint to express edges between communities. From a functional point of view it creates a phonological template for morphological mapping. Bootstrapping of this nature is important. It makes structural and functional sense for this to occur in prominent nodes with high betweenness because these words can serve as a morphophonemic bridge between communities.

3) Literature supports both models of the mental lexicon (Vitevitch, 2010, 2014, Kapatsinski, 2006). I predict the difference between the two models will be that there are significantly fewer island and hermit nodes in the 75% similarity model of the mental lexicon. In addition, the 75% model will show a greater representation of complex
morphemes in nodes with high betweenness centrality, as this network is more inclusive and provides opportunities for more lexical connections and bootstrapping. As stated above, words with affixes will occur in a higher proportion in nodes with high betweenness centrality. This finding should occur in the 75% similarity network, because the metric itself permits connections that are friendlier to words with more than one morpheme.
Chapter 2

Methods

Participants

Caregivers of six monolingual 24-48 month old children participated in this study. All caregivers and children were monolingual English speakers. Participants were recruited through informal communication at parent groups in Connecticut and New York. Caregivers completed a parental permission form for consent of their child’s participation in research. Caregivers verbally confirmed that the status of their child’s development (typical, delayed or disordered). All caregivers reported typical development. Each caregiver also completed a socioeconomic status questionnaire.

Table 1. Participant Information: Sex, Socioeconomic Status and Cognitive Development

<table>
<thead>
<tr>
<th>Participant number</th>
<th>Sex</th>
<th>Socioeconomic Status</th>
<th>Cognitive Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1M</td>
<td>Male</td>
<td>Upper middle</td>
<td>Typical</td>
</tr>
<tr>
<td>P2M</td>
<td>Male</td>
<td>Middle</td>
<td>Typical</td>
</tr>
<tr>
<td>P3M</td>
<td>Male</td>
<td>Upper Middle</td>
<td>Typical</td>
</tr>
<tr>
<td>P1F</td>
<td>Female</td>
<td>Upper Middle</td>
<td>Typical</td>
</tr>
<tr>
<td>P2F</td>
<td>Female</td>
<td>Middle</td>
<td>Typical</td>
</tr>
<tr>
<td>P3F</td>
<td>Female</td>
<td>Upper Middle</td>
<td>Typical</td>
</tr>
</tbody>
</table>
Procedures

Child Literature Corpus

Frequently read children’s books served as the corpus for the models of the mental lexicon. Caregivers selected between 10 and 20 books that were read at least five times (frequently read) to their 2-4 year old child. These books comprised the storybook corpora. Words in these books were orthographically transcribed, and then phonemically transcribed according to the computer-readable Carnegie Mellon pronouncing dictionary (Carnegie Mellon Speech Group, 1993). Words were phonemically transcribed in order to use the metrics outlined below to define the parameters for the network, and to measure lexical and sublexical characteristics of the nodes (words) in the network. Words and nonsense words in the storybooks that were judged by experts (speech scientists) to be pronounced in various ways were excluded from the study, such as Xmas, atishoo, bzz, and moosay.

Since Kucera and Frances’ literature-based language corpus was collected in 1967, using literature samples to reflect known words is common practice (Kuperman, 2012, Vitevitch, 2004). It is useful to use frequently read literature samples in this study, because these samples reflect language frequently heard by a sample of children. Two-to-four years of age is a period of remarkable language learning for children, therefore studying exposure to literature during this time period is useful. Beyond age four, as phonological awareness skills develop, children may begin to attempt reading on their own, shifting their attention from the story to the text. For these reasons, use of a literature corpus during this age range is of particularly useful. In addition, there are many different themes in frequently read children’s books; these books offer language
that is not context-specific, which may occur in a language sample. This also better reflects language experienced by children in this two-to four-year-old age group. Literature samples demonstrate the range of words that a child may hear, instead of context specific words.

Participant and Group Analysis

Data from the storybook corpus was grouped based on books contributed by caregiver (frequently read books by child), books frequently read to female children, books frequently read to male children, and a group that includes all books read to all children. The current analysis was focused on the entire lexicon.

Rank Analysis of Lexical and Sublexical Characteristics

The values given for each node after calculating degree and betweenness centrality provided a ranking of the most important nodes. High degree and betweenness centrality are defined by centrality values in the 95th percentile for each measure. Low degree and betweenness centrality is defined by centrality values in the 5th percentile. These extremes were selected (top 5% and bottom 5%) to allow testing of the hypothesis using a well-separated set of words. The number of nodes that represent the 95th and 5th percentile in both networks for degree centrality and betweenness centrality are represented in Table 2. Significantly more words with low degree (747 words for one-phoneme difference network, 772 words for 75% similarity network) are represented in both networks as compared to words of high degree, high betweenness or low betweenness. In both networks, many words exist that have just one neighbor, and all words with just one neighbor were included in the bottom 5th percentile.
The lexical characteristics of *age of acquisition* and *word frequency* were identified for each node (word) in the network with high and low degree centrality and high and low betweenness centrality. Word age of acquisition was calculated using Kuperman et al. (2012) Age of Acquisition Rating for 30,000 English Words. Word frequency was calculated based on Kuperman’s scale using SUBLTEXus (Brysbaert & New, 2009), a database of word frequency of 51 million words found in English movie subtitles. In addition, frequency of occurrence was calculated from the children’s literature samples examined in the present study. The sublexical characteristic *average phonotactic probability* was also identified for each of these words. Positional and Biphone phonotactic probability averages were calculated using Storkel and Hoover’s (2010) Child Mental Lexicon (CML) online calculator. Following the identification of these lexical and sublexical characteristics of each relevant node, these nodes were further identified as content or function words, base words or words with morphological markers. When calculating scores for word AoA, word frequency and phonotactic probability for high degree centrality, one-phoneme words were excluded from the analysis. In the 75% similarity network, one-phoneme words establish links to an extremely large number of words. In order to make an equivalent comparison in both networks, these words (e.g., *eye, I ow, oh, aww, sh*) were excluded from the analysis. In addition, a random sample of words with low degree centrality (127 words) were analyzed for word AoA, word frequency and phonotactic probability for both networks rather than the over 700 words found in each network with low degree.
Table 2. Number of words with high and low degree and betweenness in both networks.

<table>
<thead>
<tr>
<th>Network</th>
<th>Total Number of Words (highest and lowest percentile)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One Phoneme Difference Metric</strong></td>
<td></td>
</tr>
<tr>
<td>High Degree</td>
<td>161</td>
</tr>
<tr>
<td>High Betweenness</td>
<td>96</td>
</tr>
<tr>
<td>Low Degree</td>
<td>747 (127)</td>
</tr>
<tr>
<td>Low Betweenness</td>
<td>105</td>
</tr>
<tr>
<td><strong>75% Similarity Metric</strong></td>
<td></td>
</tr>
<tr>
<td>High Degree</td>
<td>146</td>
</tr>
<tr>
<td>High Betweenness</td>
<td>75</td>
</tr>
<tr>
<td>Low Degree</td>
<td>772 (127)</td>
</tr>
<tr>
<td>Low Betweenness</td>
<td>154</td>
</tr>
</tbody>
</table>

**Algorithms for Network Construction**

Two different algorithms were written to define the edges of the network. One algorithm defined edges by words that differ by one phoneme addition, deletion or substitution in any position of the word (Luce & Pisoni, 1998). Another algorithm defined edges by words that were similar in 75% of phonemes within the word (Kapatsinski, 2006). This was calculated by starting from the beginning of each word and finding the first matching phoneme among all other words. If this first phoneme matched, the next phoneme was checked for a match. This process continued until the end of the
word. If 75% of phonemes of a given word matched another, then an edge was created between those two words. See Figures 4 and 5.

**Figure 4.** Visualization of the one-phoneme difference network. Different colors represent different communities within the network.
Figure 5. Visualization of the 75% similarity network. Different colors represent different communities within the network.
Network Analysis

Algorithms for Network Analysis

Degree centrality and betweenness centrality were statistical properties calculated from the network. Mathematical expressions of these two metrics are defined here. The algorithms to calculate these measures are implemented in the program Gephi (Bastian et al., 2009). The degree of a node is the number of edges (connections) incident to that node in relation to the highest possible number in the community of that node. The degree of a node is an important index of communication activity within a network (Freeman, 1979). It is a ratio of the degree of a given node to the maximum possible degree centrality \((n-1)\). It is not just an integer count but calculates the connections of a node related to the highest possible number in their community. Degree centrality for a point \((p_k)\) is calculated in the following manner:

\[
C'_D(p_k) = \frac{\sum_{i=1}^{n} a(p_i, p_k)}{n - 1}
\]

Where \(p_i\) is a point on the graph, and \(p_k\) is the point for which degree is being measured and \(a(p_i, p_k) = 1\) if \(p_i\) and \(p_k\) are connected by an edge. If \(p_i\) and \(p_k\) are not connected, then \(p_i\) and \(p_k\) equal zero, and \(n\) is the total number of nodes in the network.

Degree centrality for a chosen node is calculated by beginning with 1 and adding the values up to \(n\). That number is divided by \(n-1\), as the given node \((p_k)\) can have an edge with up to \(n-1\) other nodes on the graph.
Betweenness centrality is equal to the number of shortest paths from all nodes to all others that pass through that node. To calculate betweenness, compute the length and number of shortest paths between all pairs, then sum all pair dependencies (Brandes, 2001, Freeman, 1979). Betweenness centrality measure was developed by Freeman (1977), and is useful as an index of the potential point of control of communication within a network.

Betweenness centrality is mathematically defined as:

\[ g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \]

\( \sigma \) represents the total number of shortest paths from node s to node t and \( \sigma_{st}(v) \) is the number of those paths that pass through v (vertex). In Figure 6 below, \( cat \) represents an example of a node with high degree centrality, and \( dad \) illustrates the measure of betweenness.
**Figure 6.** On the left, *cat* represents a node with edges (connections) to many other nodes (words). *Cat* represents high degree centrality. On the right, *dad* illustrates a measure of betweenness. *Dad* is a bridge between two different communities.

---

**Data Analysis:**

Descriptive statistics (mean, standard deviation, median, skewness, and kurtosis) were calculated for the degree centrality and betweenness measures. Words in the upper and lower 5\(^{th}\) percentiles were compared for the factors of word AoA, word frequency, positional segment average, and initial biphone frequency. The Kruskal-Wallis H test was used to determine if there were statistically significant differences between all groups (high and low betweenness, high and low degree) for word AoA, word frequency, positional segment average and initial biphone probability. The Kruskal-Wallis H was used because the data is not normally distributed, but does have the same variability. When appropriate, a post-hoc test of pairwise comparisons was performed using Dunn’s procedure with a Bonferroni correction for multiple comparisons.
A Mann-Whitney U test was conducted to determine differences in word frequency in the literature corpus between high degree and high betweenness centrality measures for both networks. This is in addition to the above analysis based on word frequency using Kuperman’s (2012) word frequency rating scale. A Spearman rank-order correlation was calculated to measure the strength of association between Kuperman’s (2012) word frequency ratings and word frequency from the literature corpus. This correlation was calculated for words that occurred in the 95th percentile for degree and betweenness in both networks.
Figure 7. The input to the two network metrics was phonemic identity of the words from children’s literature. Different parameters were set to identify neighbors in both metrics. Nodes (words) with highest and lowest degree centrality (DC) and betweenness centrality (BC) were identified for each network. Age of acquisition (word AoA), word frequency (WF), phonotactic probability (PP), lexical class (LC), and morphophonemic complexity (MPC) were identified for each word with high or low DC or BC.
Chapter 3

Results

Centrality Measures in Phonological Networks

Centrality Measures of degree centrality and betweenness centrality were obtained for both networks (75% similarity metric and one-phoneme difference metric). For both metrics, the 4,163 nodes (words) were represented, each a different word type from the entire corpus (40,106 words/tokens). Children between the ages of two and four years of age are known to have a receptive vocabulary of between 2,000 and 5,000 words (Goulden et al., 1990). This corpus represents a total number of words consistent with the average receptive vocabulary of the age group observed.

The average degree for the one-phoneme metric is 3.82 (SD=5.3), representing the average number of links each word (node) has within the network. This contrasts with the average degree of the 75% similarity network. Within this network, the average out-degree was 4.79 (SD=24.89). An example of a word that shows the average number of links is represented in Figure 8.
Figure 8. Example node from one-phoneme similarity network, *milk*, with average degree centrality.

Hermits and Islands

The number of hermits and islands differed for both metrics. In the one phoneme difference metric, there were 1285 hermits (nodes with no connections to any other nodes in the network) and 1176 islands (nodes connected to just one other word in the network) totaling 59% of the entire network. In the one-phoneme difference model more than half of the network does not connect to other nodes in the network (e.g., *unlucky*, *suffocate*, and *rattling*). In the 75% similarity metric, there were no hermits, and 174 islands (e.g., *ninja*, *ancient*, and *ahoy*), totaling 4% of the entire network, as measured for both in- and out-degree. There is a striking difference in the number of isolates when comparing these two types of networks. See Table 2.
<table>
<thead>
<tr>
<th>Network</th>
<th>Number of Nodes</th>
<th>Number of Hermits (e.g. reptile, desperately, embarrassing)</th>
<th>Number of Islands (e.g. prowl, final, drivers)</th>
<th>Percentage of Islands and Hermits</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-phoneme Difference</td>
<td>4163</td>
<td>1285</td>
<td>1176</td>
<td>59%</td>
</tr>
<tr>
<td>75% Similarity (In- and out-degree)</td>
<td>4163</td>
<td>0</td>
<td>174 (e.g. hedge, twelve, volcano)</td>
<td>4%</td>
</tr>
</tbody>
</table>

**Table 3.** Hermits and islands in one-phoneme difference and 75% similarity networks.

**Lexical and Sublexical Characteristics of Prominent Nodes**

Lexical and sublexical characteristics of prominent nodes reflecting high degree centrality and high betweenness centrality (95th percentile) were analyzed. From the one-phoneme difference network, nodes with high degree ranged from 34 edges for *who*, to 17 edges for *hot* ($M=10.29, SD=3.57$). Those with high betweenness ranged from 53,117 for *rain*, to 19,267 for *like* ($M=25,138, SD=6,480.76$). For the 75% similarity network (out-degree only) edges ranged from 113 for the word *or*, to 22 for the word *in* ($M=32.84$, $SD=15.22$).
For the measure of betweenness, high betweenness ranged from 6076 for the word *heart*, to 1209 for the word *grain* (M=2600, SD=1408.79).

As a comparison measure, these same characteristics were examined in nodes with low degree and low betweenness centrality (5th percentile). All nodes with low betweenness centrality were the islands for both networks. All nodes with low betweenness centrality had only one edge.

The characteristics examined include age of acquisition, word frequency, positional segment average and initial biphone frequency (phonotactic probability). Results of the Kruskal-Wallis H, Mann-Whitney U, and Spearman correlation tests are summarized below. Descriptive statistics for these characteristics are found in Table 5.

**Lexical and Sublexical Characteristics**

**Age of Acquisition**

Age of Acquisition (word AoA) was measured using Kuperman’s Word Frequency Rating Scale (2012). A Kruskal-Wallis H test was conducted to determine if there were differences in word AoA scores between groups that differed in levels and type of centrality. This was completed for both networks. In the one phoneme network, high degree (n = 157), low degree (n = 127), high betweenness (n = 96) and low betweenness (n = 105) were compared. The same comparison was made in the 75% similarity network: high degree (n = 132), low degree (n = 127), high betweenness (n = 75) and low betweenness (n = 154). There was a significant difference in scores among all eight conditions (high and low degree, high and low betweenness for both networks) $X^2(7) = 32.081, p<.01$; however, the differences between high centrality measures (degree
and betweenness) were not significant for either network \( X^2(7) = -63.628, p=1.000 \) (one-phoneme network), \( X^2(7) = -42.015, p=1.000 \) (75% similarity network) indicating that word AoA is not a factor in distinguishing these two types of nodes.

When comparing high to low degree nodes in the one-phoneme difference network, significant results were obtained \( X^2(7) = -136.063, p<.01 \). Nodes with high degree centrality (e.g., wet, shoe) exhibited significantly lower age of acquisition than low degree nodes (e.g., count, mother). In contrast, a significant difference was not found between high degree nodes and low degree nodes in the 75% similarity network \( X^2(7) = -76.105, p=.582 \). Significant differences were also not found between nodes with high or low betweenness in either network \( X^2(7) = -54.766, p=1.000 \) (one phoneme network) \( X^2(7) = -48.490, p=1.000 \) (75% similarity network).

**Word Frequency**

Word frequency was assigned using Kuperman’s Word Frequency Rating Scale (2012). A Kruskal-Wallis H test was conducted to determine if there were differences in word frequency scores between groups that differed in levels and type of centrality. This was completed for both networks. There was a significant difference in frequency among all eight groups (high and low degree, high and low betweenness for both groups) \( X^2(7) = 72.545, p<.01 \); however, the differences between high centrality measures (degree and betweenness) were not significant for either network \( X^2(7) = 30.167, p=1.0 \) (one-phoneme network), \( X^2(7) = 79.073, p=.686 \) (75% similarity network) indicating that word frequency is not a factor in distinguishing these two types of nodes using Kuperman’s Word Frequency Rating Scale.
When comparing high to low degree nodes in the one-phoneme difference network, significant results were attained $X^2(7) = 172.701$, $p<.01$, indicating that word frequency is another characteristic (in addition to word AoA) that is significantly different between nodes that have different number of links to other nodes within a community. Nodes with high degree centrality exhibited significantly higher word frequency than low degree nodes. A significant difference was also found when comparing nodes with high and low betweenness in the one-phoneme difference network $X^2(7) = -119.751$ $p=.014$. Words with high betweenness were more frequent. A significant difference was not found between high degree nodes and low degree nodes in the 75% similarity network $X^2(7) = 91.426$ $p=.214$. In addition, significant differences were not found between nodes with high or low betweenness in the 75% similarity network $X^2(7) = 54.023$, $p=1.000$.

An additional analysis was made comparing only nodes with high degree and high betweenness using the word frequency measurements from the storybook corpus. The Mann-Whitney U test was used to determine if there were differences in word frequency between nodes with high degree and high betweenness in both networks. This test was selected because of the non-normal distribution of the word frequency data. Median, rather than mean scores are reported for this reason also, as median scores are more representative of a sample with non-normal distribution.

In the one-phoneme network, median word frequency was statistically significant between high degree centrality (e.g., to, we) and high betweenness centrality measures (e.g., like, come) $U = 6,327.5$, $z = -2.379$, $p = .017$. For the 75% similarity network, differences between high degree (e.g., it, and) and high betweenness (e.g., still, start)
were also statistically significant $U = 4,295.5, z = -2.665, p = .008$. These results indicate that words that represent high degree nodes in both the one-phoneme similarity and the 75% similarity network occur more frequently in the storybook corpus, as compared to those found in nodes with high betweenness.

A Spearman rank-order correlation was used to examine the association between Kuperman’s Word Frequency Rating Scale (2012) and the word frequency of the storybook corpus. There was a moderately strong positive correlation between the two measures of word frequency $r_s = .784$. The differences in the analyses may be attributed to raw frequency measures in the literature corpus, and frequency per million words measure used in the Kuperman (2012) scale.

Positional Segment Average (Phonotactic Probability)

Positional Segment Average was measured using the Child Mental Lexicon Calculator (Storkel & Hoover, 2010). Median scores for Positional Segment Average were statistically significantly different between the levels of centrality measures for both networks $X^2(7) = 53.359 \ p < .01$. Following this analysis, pairwise comparisons were performed using Dunn’s (1964) procedure with a Bonferroni correction for multiple comparisons. Adjusted $p$-values are reported here. This post hoc analysis revealed statistically significant differences in Positional Segment Average between high degree ($Mdn = .054, SD=.020$) and high betweenness ($Mdn = .060, SD=.017$) ($p = .002$) for the one-phoneme network, but not for high degree ($Mdn = .060, SD=.023$) and high betweenness ($Mdn = .056, SD=.051$) ($p = 1.00$) in the 75% similarity network. In the one-phoneme difference network, words represented by high betweenness nodes had
higher Positional Segment Average than high degree nodes e.g., *sit* and *so* for high degree; *city* and *sack* for high betweenness). A significant difference in positional segment average between words that are represented by these nodes suggests that the sublexical characteristic of words in the lexicon determines their position in the network and the role they play in lexical access.

In addition, comparisons between high (Mdn = .054, SD = .020) and low degree nodes (Mdn = .050, SD = .013) (p = .05) were significant in the one-phoneme network, but not the 75% similarity network (Mdn = .060, SD = .023 vs .051, SD = .013) (p = 1.00). Nodes with high (Mdn = .06, SD = .017) and low betweenness (Mdn = .052, SD = .012) (p < .01) exhibited significant differences in the one-phoneme network, however, these results were not significant in the 75% similarity network: (Mdn = .056, SD = .051) high betweenness (Mdn = .051, SD = .013) low betweenness (p = .085). These results give us some information about the how the general structure of the network influences its function.

**Initial Biphone Frequency**

Comparisons of initial biphone frequency was also measured using Storkel & Hoover’s (2010) Child Mental Lexicon Calculator. As with Average Positional Probability, median scores for Initial Biphone Frequency were significantly different among all eight conditions (high and low degree, high and low betweenness for both networks) $X^2(7) = 55.512$ $p < .01$. As with other measures, pairwise comparisons were performed using Dunn’s (1964) procedure with a Bonferroni correction for multiple comparisons. Adjusted $p$-values are presented. Statistically significant differences were
determined for Initial Biphone Frequency between high degree \((Mdn = .003, SD = .002\) vs. \(Mdn = .004, SD = .004\)) and high betweenness \((Mdn = .005, SD = .002\) vs. \(Mdn = .006, SD = .006\)) \((p = .002, p < .01)\) for both the one-phoneme and 75% similarity networks, respectively, in post hoc analysis (examples can be found in Appendix B). When comparing differences within-network, differences between high (e.g., zoo, show) and low degree (e.g., hurray, visit) were significant for the one-phoneme difference network only \((Mdn = .003, SD = .002\) vs. \(Mdn = .005, SD = .003\)) \((p < .001)\). Differences between high and low betweenness for the one-phoneme difference network \((Mdn = .005, SD = .002\) vs. \(Mdn = .0041, SD = .003\)) \((p = 1.00)\), and high and low degree \((Mdn = .004, SD = .004\) vs. \(Mdn = .004, SD = .004\)) \((p = 1.00)\) and betweenness \((Mdn = .006, SD = .006\) vs. \(Mdn = .005, SD = .004\)) \((p = .303)\) for the 75% similarity network were not significant.

**Occurrence of Function and Content Words**

As addressed earlier in this paper, function and content words serve different purposes in the language in terms of language acquisition. Function words, such as I, and, the, and of exhibit phonological qualities different from content words. Lexical and sublexical characteristics of function words are examined in this dissertation. Percentages of function words in the 5\(^{th}\) and 95\(^{th}\) percentile for all words in the corpus are reported in Table 3. For both the one phoneme difference metric and the 75% similarity metric, function words are more highly represented in high degree nodes. Function words occur at a slightly higher percentage (2%) in high degree nodes in the 75% similarity network. Although many function words differ from content words in terms of lower frequency phonemes and decreased vowel stress (e.g., the, of, a), some are found as high degree
nodes in both networks (e.g., *of, a you*). That is, some function words are highly connected.

**Morphological Complexity**

Morphological complexity was measured for words found in the 5th and 95th percentile for degree centrality and betweenness centrality in both networks. Descriptive statistics for these nodes are found in Table 4. Low degree and low betweenness nodes from both networks exhibited the highest percentage of morphologically complex words (61%). Nodes from the 75% similarity network with high betweenness centrality followed in morphological complexity. Importantly, the prominent, highly connected nodes in a network that include many morphologically complex words (42%) are those of high betweenness in the 75% similarity network only.

**Table 4.** Occurrence of Function Words in Literature Corpus

<table>
<thead>
<tr>
<th>Network</th>
<th>Total Number of Words (highest and lowest percentile)</th>
<th>Number of Function Words</th>
<th>Percentage of Total Function Words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One Phoneme Difference Metric</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Degree</td>
<td>161</td>
<td>19</td>
<td>12%</td>
</tr>
<tr>
<td>High Betweenness</td>
<td>96</td>
<td>2</td>
<td>2%</td>
</tr>
<tr>
<td>Low Degree</td>
<td>747</td>
<td>16</td>
<td>2%</td>
</tr>
<tr>
<td>Low Betweenness</td>
<td>105</td>
<td>2</td>
<td>2%</td>
</tr>
<tr>
<td>75% Similarity Metric</td>
<td>Number of nodes</td>
<td>Number of morphologically complex words</td>
<td>Percent morphologically complex</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------</td>
<td>----------------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>High Degree</td>
<td>146</td>
<td>20</td>
<td>14%</td>
</tr>
<tr>
<td>High Betweenness</td>
<td>75</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>Low Degree</td>
<td>772</td>
<td>2</td>
<td>2%</td>
</tr>
<tr>
<td>Low Betweenness</td>
<td>154</td>
<td>2</td>
<td>1%</td>
</tr>
</tbody>
</table>

**Table 5.** Morphological Complexity in One-Phoneme Difference and 75% Similarity Networks

<table>
<thead>
<tr>
<th>Lexical/sublexical characteristic</th>
<th>Number of nodes</th>
<th>Number of morphologically complex words</th>
<th>Percent morphologically complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-phoneme Difference Network</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Degree</td>
<td>161</td>
<td>7 (eg. bees, bought, caught, sighed, they’re, we’d, we’ll)</td>
<td>4%</td>
</tr>
<tr>
<td>High Betweenness</td>
<td>96</td>
<td>4 (eg. caught, raced, saw, signed)</td>
<td>4%</td>
</tr>
<tr>
<td>Low Degree</td>
<td>747</td>
<td>456</td>
<td>61%</td>
</tr>
<tr>
<td>Low Betweenness</td>
<td>105</td>
<td>62</td>
<td>59%</td>
</tr>
</tbody>
</table>
Words like *clears* and *picked* were found to have high betweenness in the 75% similarity network, and words like *stomped, streaks,* and *trying* were found to have low degree in the one phoneme difference network. Both inflectional and derivational morphemes were represented. The high degree nodes in the one-phoneme difference network had no words with multiple morphemes. Most (61%) words with morphological complexity were found in low degree and betweenness nodes in both networks, and nodes with high betweenness in the 75% similarity network (42%). Very few words had more than two morphemes. A few instances of three morphemes existed, such as *bathtubs* and *firemen.* Words with morphological complexity found in prominent positions in the network, such as nodes with high betweenness centrality could support phonological bootstrapping of morphological/syntactic forms.
Table 6. Summary of Descriptive Statistics for Lexical and Sublexical Characteristics of Prominent Nodes

<table>
<thead>
<tr>
<th>Node Type</th>
<th>Median Word AoA</th>
<th>SD Word AoA</th>
<th>Median Frequency</th>
<th>SD Frequency</th>
<th>Median PSA</th>
<th>SD PSA</th>
<th>Median BP</th>
<th>SD BP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One-phoneme Difference</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Degree</td>
<td>5.0</td>
<td>1.562</td>
<td>105.2</td>
<td>2960.8</td>
<td>0.054</td>
<td>0.02</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td>High Betweenness</td>
<td>5.4</td>
<td>1.32</td>
<td>62.0</td>
<td>780.6</td>
<td>0.060</td>
<td>0.017</td>
<td>0.005</td>
<td>0.002</td>
</tr>
<tr>
<td>Low Degree</td>
<td>5.94</td>
<td>2.19</td>
<td>14.14</td>
<td>273.0</td>
<td>0.050</td>
<td>0.013</td>
<td>0.005</td>
<td>0.003</td>
</tr>
<tr>
<td>Low Betweenness</td>
<td>5.70</td>
<td>1.60</td>
<td>20.90</td>
<td>152.7</td>
<td>0.052</td>
<td>0.012</td>
<td>0.0041</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>75% Similarity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.10 1.53 74.92 3073 .06 0.023 .004 0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High degree</td>
<td>5.26 1.68 27 124 0.056 0.051 .006 .006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Betweenness</td>
<td>5.80 1.94 22.6 1471 0.051 0.013 0.004 0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Degree</td>
<td>5.76 2.01 16.72 223 0.051 0.013 .005 .004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Betweenness</td>
<td>5.76 2.01 16.72 223 0.051 0.013 .005 .004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4
Discussion

The goal of this study was to examine the phonological organization of the developing mental lexicon by investigating the prominent nodes in two different networks, based on lexical and sublexical characteristics assigned to those prominent nodes. More specifically, centrality measures, degree (measurement of number of nodes directly connected to a given node) and betweenness (measurement of nodes on shortest paths that bridge communities) were examined in two different models, the one-phoneme difference model, and 75% similarity model. Lexical and sublexical characteristics (word AoA, frequency, positional segment average, and biphone frequency) were examined for nodes that occurred in the top and bottom 5\textsuperscript{th} percentile. Results of the Kruskall-Wallis H test revealed significant differences between nodes of high degree versus high betweenness in initial biphone average in both networks. The one-phoneme difference network demonstrated significant differences when comparing these two centrality measures for positional segment average. Results of the Mann-Whitney U test comparing word frequency of the literature corpus in high degree to high betweenness nodes revealed significant differences for both networks.

Function words were found to occur more frequently in high degree nodes than nodes with high betweenness, or low degree or low betweenness. Morphologically complex words occurred most frequently in nodes of low degree and betweenness (nodes
of low prominence) for both models, and nodes of high betweenness (high prominence) in the 75% similarity network.

Statistical analyses were completed independently for each variable (word AoA, word frequency, positional segment average, initial biphone probability). It is clear that these variables are not at all independent of one another. In addition, these factors represent both lexical and sublexical (phonological) levels of processing. However, any factor that demonstrates that a significant difference exists between two types of prominent nodes (degree and betweenness) or between the two model networks (one-phoneme difference and 75% similarity) reveals a structural difference in the network(s), which may be revealed in a difference in network function. These results are discussed further, below.

Centrality Measures and Lexical and Sublexical Node Characteristics

The first research question asked: Which model, the one-phoneme or 75% similarity rule, better reflects child language development? Lexical properties, including word AoA, word frequency, positional segment average, and biphone frequency were used to address this question.

Word Age of Acquisition

It was predicted that word AoA would be higher in high degree nodes, because of the network property of preferential attachment. Graph theory suggests that early-acquired nodes tend to attract more nodes to their edges as the network grows (Barabasi, 2002). It was predicted that this would have been evident in both network models, as seen as lower word AoA found in high degree nodes. Age of acquisition of words was not significantly
different when comparing words of high degree centrality to words of high betweenness in either network. One factor that may have impacted the results is the use of the Kuperman rating scale (2012) as an instrument to gauge word AoA for the words with high degree and betweenness centrality. Only uninflected forms are found on the Kuperman (2012) Rating Scale. Many of the words of high betweenness centrality in both networks were morphologically complex. On the Kuperman scale, the word AoA scores assigned for morphologically complex words are those of the root word (e.g. beaten, harder, pours). It is possible that these morphologically complex words have different word AoA than their uninflected forms (e.g., beat, hard and pour). Although most singular and present tense forms are learned earlier than their inflected counterparts, not all words are learned this way (e.g., foot, vs. feet, walk vs. walking).

Another factor that may have played a role in not finding the predicted pattern may be related to the corpus itself. Words ranged in word AoA from 2.4 years to 13 years, with the mean for words in the top and bottom 5% of both networks around 5 years of age, and SD around 1.5 years of age (see Table 6). Children’s books tend to have richer vocabulary than spoken language. Context clues and illustrations support the understanding of these words. Children’s books tend to include richer vocabulary to promote interest and word learning. The median age for all centrality measures for both groups was 4.72 years of age. In a fully developed lexicon of an adult, AoA of words are likely to vary to a greater extent. The range of word AoA may have been too narrow to show an effect. Perhaps most words in this corpus could be considered “early” words. In addition, book reading may influence an idiosyncratic lexicon. Words such as excavators and mink may become a part of the child’s lexicon at an early age if frequently read
books include such words. This would also influence the age of acquisition, as compared to words found in the Kuperman scale (2012).

In contrast, word AoA was significant for both the one-phoneme difference model when comparing high and low degree nodes. Words with high degree showed a significantly lower word AoA than words with low degree centrality. This finding supports the network property of preferential attachment, as early-acquired nodes in both networks did attract more nodes to their edges than later acquired words, rather than the stems found in Kuperman (2012). In future studies, it will be important to determine if the difference in word AoA will be significant when comparing high degree and high betweenness when word AoA is calculated using a rating scale that includes morphologically complex word forms. It will be interesting to know whether words that link communities (high degree) or bridge communities (high betweenness) have higher or lower word AoA. This would inform us about the relationship word AoA has with the structure of the networks within and between their communities. The current results indicate that word AoA is only a significant factor within communities in the one-phoneme network, with high and low degree nodes demonstrating different word AoA.

This research demonstrates that lower word AoA facilitates retrieval times, when isolated as a factor in lexical processing. However, in these model networks, words with low AoA have a higher number of links to other nodes, potentially increasing lexical search, and search time. However, these low AoA words appear to lead to a higher level of resting activation, which may allow them to be retrieved faster even though they have more connections, or neighbors than higher AoA words. Further investigation is necessary to tease apart these factors.
Word Frequency

Differences in word frequency should also follow the principle of preferential attachment, a principle of graph theory that reveals nodes that have the most attachments are likely to accumulate more.

It was predicted that words with high degree centrality in both networks would have higher word frequency than words with high betweenness centrality (words with high degree were predicted to have higher frequency than words with high betweenness). In previous studies (Carlson et al., 2011, Siew, 2013, Steyvers & Tennenbaum, 2005) nodes (words) with high frequency were also found high degree neighborhoods. It was predicted that this pattern would be observed in the present dissertation. This prediction was confirmed for both networks between words with high degree versus high betweenness when the literature corpus was used as the dependent variable, but not when Kuperman’s (2012) scale, based upon SUBTLEXus (2008), was used. The SUBTLEXus (2008) corpus is based upon subtitles of movies in English. It is not a sample based upon words children are likely to hear. Although this sample is a much larger sample (51 million words) it may not reflect the frequency with which children hear words as well as a sample from children’s literature.

Some of the most frequent words, such as the, a highly frequent function word, were not found in high degree nodes of either network. It has been demonstrated that function words are processed differently in the brain than content words (Pulvermuller et al., 1995, Shafer et al., 1998, Shi et al., 2003). Further investigation of this finding is warranted. It could be the case that, in a two-representation model of the mental lexicon
(lexical and sublexical), words could have different levels of prominence. In a study by Concho and Sole (2005), semantic relations of adjacent words (at the sentence level), function words were found to be among the most highly connected nodes in the network. It could be the case that function words are more highly connected at the semantic level of lexical access. Although phonological and semantic levels of the mental lexicon interact, words can have different levels of prominence in a two-representation model. In addition, Jackson & Bolger (2015) found that co-occurrence is the acquisition and storage mechanism that accounts for the relationship between words.

Significant differences in word frequency were observed between high and low degree nodes in both networks using Kuperman’s (2012) scale. High degree nodes exhibited higher word frequency than low degree nodes. This is in keeping with Zipf’s law, but, as stated above, this informs us about nodes within communities, not between communities. A comparison of prominent nodes within and between communities informs us about lexical access throughout the mental lexicon, rather than access only within communities.

Although these representations of frequent words in the lexicon supports graph theoretic principles, and previous work examining child and adult mental lexicon network models, it contrasts with behavioral results of linguistic studies, if one only considers the influence of degree (neighborhood density) on reaction time, and not the influence of differences in resting activation. RTs to frequent words are faster than those to infrequent words. If frequent words have a high degree centrality (many neighbors), this should increase lexical search time, and cause RTs to be slower. However, according to Cohort Theory (Marslen-Wilson, 1990), higher frequency words in a cohort have a higher level
of resting activation, giving these words an advantage during lexical search. Given the results of this study, the plausible explanation for the presence of frequent words in high degree nodes in the developing mental lexicon is that these words have higher resting activation levels, in order to rise above the many other words in the neighborhood. It is important to investigate the relationship between resting activation levels of frequent words in these communities as a possible factor in these results. These results are not incompatible, but serve to more precisely describe lexical processing based on word frequency in the developing mental lexicon.

**Positional Segment Average (Phonotactic Probability)**

As observed above, only the one-phoneme difference network revealed a significant difference in positional segment average (PSA) when comparing words with high degree versus high betweenness centrality. Specifically, nodes with high betweenness were represented by words with higher PSA in this network as compared to high degree nodes. This finding opposes the prediction that PSA would be higher in high degree nodes. It was predicted that nodes with high degree would demonstrate higher positional segment average, based on previous research (Luce & Pisoni, 1998, Vitevitch and Luce, 1998, 1999). It was predicted that words with high betweenness would have a lower positional segment average in order to constrain the lexical search within communities. Instead, these results revealed nodes with high betweenness to have a higher positional segment average than nodes with high degree. Close inspection reveals that nodes with high degree do not have low phonotactic probability (PP), just lower PP than those nodes with high betweenness. Those nodes with high betweenness have the highest PP. This may play a role in facilitating spreading activation during lexical access as sounds with
highest PP should reach activation threshold for recognition more rapidly. In addition, if words representing nodes with high betweenness have highest PP, this may narrow the search space (during lexical access) from the entire network to words with high betweenness centrality and their connections. This would represent a more efficient search mechanism.

In the following paragraph, I will explore the sublexical influence of lexical activation represented by these results. Phonological (sublexical) factors affect speed of access between communities (via words with high betweenness) due to higher resting activation levels for phonemes with higher PP. Words with lower PP (high degree nodes) will not be activated as quickly as those nodes with high betweenness. However, the lexical competition effect also affects processing speed, as high betweenness centrality nodes are also highly connected.

If words with high betweenness have higher PSA (as compared to high degree nodes) as suggested by these results, this would affect lexical access in the following way: A prominent node with high betweenness centrality is activated. This node, with highest PSA and therefore a higher resting activation, quickly activates all other node edges that meet phonological requirements for activation. These nodes are a starting point to facilitate lexical search to communities in which the target word may reside. It is then that the target word is identified and activated. This process narrows the search space from the entire network to those words with high betweenness and their connections. Although the prediction differs from the results, the impact of high(est) degree nodes on the network in nodes of high betweenness still reflect an efficient system.
In addition, Steyvers & Tennenbaum (2005) propose that high centrality (degree or betweenness) reflects a node’s authority in a network, and there may be a bias toward accessing words with high connectivity. These authors liken lexical search to the search process Google uses, ordering websites based on their degree centrality to other websites. In this way, more central, highly connected nodes may be accessed first in the lexicon.

No significant difference in positional segment average was observed in the comparison of nodes with high degree and high betweenness in the 75% similarity network. This may be due to the parameters under which edges are defined in this network. When words are defined by 75% similarity, words, with one, two and three phonemes must be similar (point to) neighbors if they are identical (100% similar) to part of the neighboring word in order to create an edge. This occurs because words with fewer than four phonemes cannot have 75% similarity to any word if they differ by even one phoneme. For example, words with three phonemes, such as cat can only have an out-degree relationship with words that have the phonemes c-a-t in succession within the word, such as catfish, caterpillar, and cats. This may influence the results of positional segment average of words in this network. In this way, defining an edge when words differ by 75% similarity may not be sensitive enough to create a network that reflects significant differences between words that are prominent within communities versus words that are prominent between communities.

Another reason that significant differences in PSA were not detected in the 75% similarity network may be due to the corpus used to calculate PSA using Storkel and Hoover’s (2010) CMLC calculator. Only uninflected words were used to calculate PSA in this calculator. This may make calculating PSA for inflected words from the present
literature corpus difficult. Words with common affixes such as *cats*, *picky*, and *playing* are not found as part of her corpus (only *cat*, *pick*, and *play*). More inflected words are found in the 75% similarity network sample found in this dissertation. This may have affected PSA calculations for this network. Words with common phonemes in final positions representing common inflections were not calculated as highly frequent phonemes using this calculator.

**Initial Biphone Frequency (Phonotactic Probability)**

Initial biphones may play a crucial role in lexical retrieval. Research suggests that word onset is important in lexical search, as these first phonemes influence response times in similarity judgment and reaction time studies (Kapatsinski, 2005, Vitevitch et al., 2004, Kidd & Watson (1992); see Radeau et al., 1995 for importance of word endings). In this dissertation, both networks (75% similarity and one-phoneme difference) demonstrated significant differences in initial biphone frequency when comparing high degree centrality and high betweenness centrality. Specifically, nodes with high betweenness centrality demonstrated higher initial biphone frequency than nodes with high degree centrality in both networks. This suggests that biphone frequency is a factor that distinguishes these two types of nodes. This distinguishing factor may give us insight into lexical access in the developing mental lexicon.

As described above, there is evidence of interaction at the lexical and phonological levels of lexical processing. Lexical retrieval is facilitated by low phonotactic probability (and sparse neighborhoods), and inhibited by words with high phonotactic probability and dense neighborhoods). Lexical competition is also either
facilitated or inhibited by words and their position in the network. The description of spreading activation for nodes with high degree and high betweenness in the above section describing positional segment average holds true for initial biphone frequency, as these are both measures of PP.

Nodes with high betweenness centrality consist of words with higher initial biphone probability (as compared to high degree nodes). For example, when the initial biphone pair /st/ is activated, spreading activation from high betweenness /st/ nodes results in activating nodes within communities that share those initial biphones. These nodes are activated quickly, as higher PP increases speed of access based on higher levels of resting activation. This represents an efficient lexical search mechanism. The search space for words with initial biphone /st/ is narrowed from the entire network to high betweenness nodes and edges.

These results are in opposition to the author’s predictions. The author predicted that high degree words would favor high phonotactic probability, and low degree words would favor low phonotactic probability, as reported in Walley (2010) However, Siew (2013) examined lexical and sublexical characteristics of community structure of the mental lexicon compared to a random network. She did not find a correlation between community size and mean biphone probability. A possible explanation may be that behavioral studies revealed longer reaction times to words with both dense neighborhoods and high phonotactic probability, but graph theory is able to demonstrate that lexical processing based on these phonological and lexical characteristics may also be facilitated or inhibited by their position in the network.
The sublexical characteristic of initial biphone probability may be a stronger indicator of lexical retrieval than positional segment average. The Cohort Model is well supported in current research, and demonstrates the importance of initial phonemes in lexical retrieval. Therefore, higher initial biphone probability would be a strong measure of the role of this sublexical characteristic in lexical retrieval. The sublexical characteristic of positional segment average is also a measure of phonotactic probability, but may be less of an indicator of lexical access, as this measure *averages* the likelihood of each speech sound occurrence in each position for each word, rather than pinpointing the probability of occurrence of beginning speech sounds.

**Frequency of Function Words and Morphologically Complex Words**

This dissertation also asked: How are function words and morphologically complex words organized in the one-phoneme versus the 75% similarity network? Which model is better supported by the literature?

Current behavioral and ERP studies suggest that function words are processed differently from content words, even in infants (Shafer et al., 1998, Shi et al., 2006). Function words are known to have distinct signatures across multiple levels of representation, including fewer syllables, simple onsets and codas, a more restricted phoneme inventory, high token frequency, and morphological simplicity (Morgan et al., 1996).

In the present study, function words occurred more frequently in high degree nodes in both networks. Of the 20 most common function words in Chung & Pennebaker’s (2007) text archive, *(I, the, and, to, a of, that, in, it, my, is, you, was, for, ...*
have, with, he, me, on, but), seven (35%) are represented in high degree nodes of the one-phoneme network, and ten (50%) are represented in high degree nodes of the 75% similarity network. None of the top 20 function words were represented in the high betweenness, low degree, or low betweenness nodes of either network. As function words are highly frequent in English, it is consistent with the present finding that more frequent words occur in high degree nodes (based on the notion of preferential attachment). In addition, this may occur because phoneme pairs found in function words such as I, an, or, and in are commonly found in other words, however, the difference would be more than one phoneme for many word pairs (e.g., I vs. bite). In fact, these words in their entirety are often found in other words; therefore function words “point to” many other words in a directed network (such as the 75% similarity network). For example, the function word or points to four, snore, order, escort, and horn.

In total for both networks, 70% of the most frequent function words are present in the high degree nodes (I, and, to, a, of, in, it, my, is, you, for, he, me, on) yet all of the most frequent function words were present in the corpus itself. Of the most frequent phonemes not represented in high degree of either network, those phonemes range in a low degree of one neighbor (for the in both the one-phoneme and 75% similarity networks) to an average eight neighbors (for but in the 75% similarity network) to a relatively high number of neighbors (10) for that in the one-phoneme similarity network. In both networks, the highly frequent function word the is a neighbor only to the word a, which is also a highly frequent function word. It is possible that these words hold special status, or there is a unique relationship between these words in the lexicon. Not all frequent function words were found in nodes of high degree, but the majority of the most
frequent function words were found in this position. This finding warrants further investigation into the structure of the lexicon and the status of function words within the developing lexicon. For example, an electrophysiology study examining the n400 response to function words found in high degree nodes as compared to function words not found in high degree nodes may shed light on the status of different function words in the developing mental lexicon.

Morphologically complex words in this corpus consist primarily of compound words, derivational morphemes, and inflectional morphemes, including irregular past tense, past tense -ed, plural-s, comparative -er, and present progressive -ing. Aside from the nodes represented by irregular forms, all of these nodes include markings that can be identified morphophonologically. As stated in the introduction, phonological bootstrapping can be used as a spotlight to draw the child’s attention to the syntactic rule, initially. Following the initial attention to syntactic rules, implicit exposure to phonological forms that mark morphological differences can be reinforced by their existence in prominent nodes. These phonological features serve as nontransparent markings to cue the child about morphosyntactic rules (Ellis, 2002). It makes sense that morphologically complex words are found in words with high betweenness centrality, such as filled, picky and spills. These nodes bridge communities and allow a great number of communities to be linked morphophonologically. This leads phonologically based communities to extend to linguistic patterns, specifically information related to morphosyntax. Listeners could extract phonological information from the speech stream to facilitate learning of syntactic information.
Comparison of Two Network Models

A third question addressed by the dissertation was: What differences exist between 75% similarity and one-phoneme difference models? Which one is consistent with current literature related to the development and organization of the mental lexicon?

In order to address this question, each network is compared according to lexical and sublexical characteristics in prominent nodes, (high degree and high betweenness) position of function words and words with high morphological complexity in the network, and network connectivity (small world characteristics, average shortest path length, average degree, and number of islands and hermits).

Network Connectivity

Two models were compared using an identical corpus, one that defines edges by a one-phoneme difference (addition, deletion or substitution), and one that defines edges by 75% similarity of phonemes within a word. Although the corpus was identical, the networks created highly different connections.

The average degree for words was greater for the 75% similarity model (4.79) as compared to the one phoneme difference model (3.82), meaning more neighbors within a given node’s community would be activated when a target word is presented. Coady & Aslin (2003) reported that children’s neighborhoods were denser than those of adults based on total number of word types in the child speech sample. This could impact lexical competition and lexical processing, however, Carlson et al. (2010) found that network models representing child speech (CS) and child-directed speech (CDS) had higher degree distribution than network models based on adult directed speech (ADS),
and reported that CS and CDS lexicons were more stable and searchable than ADS. The 75% similarity model of the developing mental lexicon demonstrates a better fit with Carlson and colleagues’ representation of the developing mental lexicon.

The average shortest path length was longer for the one-phoneme difference model (5.8) versus the 75% similarity model (4.47), making this network longer to traverse. In other words, on average, the path (number of nodes) between a given node and any other node in the network is shorter in the 75% similarity network than the one phoneme difference network. This may be a detriment for the mental lexicon if the goal is efficiency of lexical access, particularly because the one-phoneme difference model has many islands and hermits compared to the 75% similarity model. In addition, the number of connected components within the 75% similarity network is nearly double that of the one-phoneme network, yet this network demonstrates the shorter average shortest path length.

One question is whether further research will support a model with a larger or with a smaller number of hermits/isolates. Steyvers & Tennenbaum (2005) propose that concepts become semantically inaccessible when they are disconnected as islands and hermits in a network. Nearly half of the words from the literature corpus were disconnected in the one-phoneme difference model. The 75% similarity model demonstrates a more inclusive model of the developing mental lexicon.

Another way to compare these two networks based on connectivity is by examining small-world characteristics (Watts & Strogatz, 1998). Small world characteristics are those that speak to the efficiency and robustness of a given network.
These characteristics are measured by *average clustering coefficient* and *average shortest path length*. As described above, the clustering coefficient measures the degree to which nodes in a graph cluster together. It measures the number of neighbors of a given node that are also neighbors to one another. Average shortest path length is an indication of how easily traversable a network may be. Taken together, networks with a high clustering coefficient and low average shortest path length are considered to be efficient and robust. They are easily traversable, efficient to search, and robust to perturbations, so that links between communities tend to stay intact even if nodes are attacked or removed (Barabasi, 2002). In Carlson et al.’s work (2010), networks based on CS and CDS exhibited higher clustering coefficients and lower average shorter path lengths than ADS. In the current study, the 75% similarity network demonstrated higher clustering coefficient (.371) as compared to the one-phoneme difference network (.347), and a shorter average path length (4.47) as compared to the one-phoneme difference network (5.82). Based on small world characteristics, the 75% similarity model demonstrates a more efficient network for the developing mental lexicon.

**Age of Acquisition and Word Frequency**

The two models of the developing mental lexicon did not reveal significant differences in word AoA when comparing centrality measures within each network. Comparison within communities demonstrated differences in high and low degree in the one phoneme network only. This demonstrates some element of defined structure within the one-phoneme difference network. As stated above in detail, the instrument used to assess AoA did not include inflected words. It is possible that this may have influenced the results. In addition, the mean AoA for the storybook corpus was around 5 years of age,
with a SD of around 1.5 years of age. This small range of AoA for words in the corpus may have been too narrow to show an effect.

**Word Frequency**

Differences in word frequency between prominent nodes was significant in both networks, but only for word frequency in the storybook corpus. Word frequency differences were not significant when using the word frequency measurement tool SUBTLEXus (2008). As discussed in detail above, this may have occurred because of idiosyncracies in the child lexicon, or because the SUBTLEXus (2008) is based upon the adult lexicon.

**Positional Segment Average**

Comparison of high degree and high betweenness for positional segment average within each network revealed a significant difference in the one-phoneme difference network only. Higher PSA was found in words with high betweenness as compared to words with high degree. As stated above in detail, significant differences were not found for the 75% similarity network, possibly due to either the parameters set for linking nodes as neighbors, or the differences in words in the literature corpus versus the CMLC based on inflection (Storkel & Hoover, 2008).

**Initial Biphone Frequency**

Statistically significant differences were found between high degree and high betweenness nodes in both networks for initial biphone probability. Both networks demonstrated higher initial biphone probability in high betweenness nodes versus nodes
with high degree. This also supports the research above in a similar way. Nodes with high betweenness have higher initial biphone frequency, and modulate lexical access by spread of activation to different communities. An additional factor to consider is the relevance of initial biphone frequency as a measure of phonotactic probability, as compared to positional segment average. Marslen-Wilson’s (1987) Cohort Theory of spoken word recognition relies on the temporal component of lexical access. Candidates for target word selection are eliminated if successive phonetic information does not match the target word. In this case, initial phonemes would be integral to the selection of a target word. If spoken word recognition proceeds as the Cohort Model suggests, then initial biphone frequency is a more relevant sublexical characteristic to measure. Alternatively, Luce & Pisoni’s (1998) Neighborhood Activation Model (NAM) does not consider this temporal component, but other lexical and sublexical factors that contribute to the identification of the target word. These factors include frequency of the target word, number and word frequency of neighbors, and confusability of constituent phonemes of the target word. Positional segment average may be a better indicator of the sublexical contribution to word recognition in this case.

**Function Words**

In both networks, function words occurred most often in high degree nodes (12% of nodes in the one-phoneme difference network, and 14% of nodes in the 75% similarity network). Of the 20 most frequently occurring function words, seven occurred in the one-phoneme difference model, and 10 in the 75% similarity model. In both networks, function words (also frequently occurring words) were a significant presence in high degree nodes. RTs to frequent words have been shown to be faster than infrequent words,
so their presence in larger neighborhoods (words with high degree) suggests higher resting activation levels for frequent words, as research also demonstrates slower RTs to words with dense neighborhoods due to lexical competition. Both Cohort and NAM models of word recognition account for this phenomenon by assuming elevated resting activation levels or preference for frequent words.

**Morphological Complexity**

It was predicted that high morphological complexity in high betweenness nodes would indicate an environment that could be friendly to phonological bootstrapping. Phonological experience with morphological affixes could spotlight these affixes and facilitate learning of syntactic information. This only occurred in the 75% similarity network. Only the parameters for the 75% similarity network allowed for words with this level of complexity to be highly linked. The parameters for the one-phoneme difference network do not allow for such connections. In this way, the structure of each network dictates their possible function. The structure of the 75% similarity network allows this network the possibility for phonological bootstrapping, however, the one-phoneme difference network does not.

**Summary**

Examination of lexical and sublexical properties of prominent nodes (nodes with high degree centrality and high betweenness centrality) revealed qualities important to the structure and function of networks based on the developing mental lexicon, but did not reveal differences between the two networks, with the exception of morphological complexity. Examination of types and number of connections between nodes (network
connectivity) revealed important differences between these two networks. As previously stated, there are many factors involved in lexical access; however, examination of these isolated factors leads to the conclusion that the 75% similarity network is a better fit for the developing mental lexicon based on the following:

The lexical characteristics AoA was not a significant feature of the network that demonstrated differences in centrality measures. Comparison of centrality measures for PSA was significant for the one-phoneme difference network only. This measure of phonotactic probability can be influential in some models of lexical access (NAM, TRACE). Comparison of centrality measures for biphone frequency was significant for both networks, and can be influential models of lexical access such as the Cohort Model. High frequency function words commonly occur in high degree nodes of both networks, but not all frequent function words exist in high degree nodes in either network. This suggests a need for further studies in this area. Morphologically complex words occur in nodes of high betweenness only in the 75% similarity network, suggesting that the 75% similarity model supports phonological bootstrapping. An extremely high number of words are isolates (islands and hermits) in the one-phoneme network, but not the 75% similarity network. This suggests overall stronger connectivity in the 75% similarity network. All other aspects of network connectivity (including fewer isolates) are dominated by the 75% similarity network (average degree, average shortest path, average clustering coefficient) Distinct differences exist between models. The 75% similarity model more strongly represents current literature reflecting lexical access in the developing mental lexicon. See Table 7.
<table>
<thead>
<tr>
<th>Age of acquisition</th>
<th>X</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positional segment average</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Initial biphone probability</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Function word representation</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Morphological complexity</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Network Connectivity – Fewer isolates</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Network Connectivity – Neighborhood density that represents current literature</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Network Connectivity – Small World Characteristics</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

**Table 7.** Comparison of significant network characteristics.

Both network models are consistent with lexical competition as a mode of lexical access. Analysis of prominent nodes using centrality measures revealed that statistically significant differences in lexical properties and sublexical properties (e.g., word frequency, positional segment average, initial biphone frequency, function word presence
and morphological complexity) between degree centrality (connections within a community) and betweenness centrality (bridges between communities) existed in both networks. Observations in these centrality measures in both networks support current literature demonstrating speed of access based on phonotactic probability and word frequency (Balota & Chumbley, 1984, Vitevitch & Luce, 1998).

Based on principles of graph theory and network connectivity, the 75% similarity network demonstrated a more stable and efficient environment for lexical development. The average degree of the 75% network is greater, representing a neighborhood structure supported by Carlson and Tennenbaum (2010). Fewer isolates, and a higher average shortest path length and higher clustering coefficient make this network more efficient in terms of search and stability.

Using network modeling to examine consequences of different relational rules is a fruitful way to examine the organization of the mental lexicon. Comparing networks built upon different parameters allows the researcher to consider possibilities for relations between words, and to generate and test hypotheses regarding the nature of the organization of the developing mental lexicon. Results of this study demonstrate that, using graph theoretic techniques, the two models of the developing mental lexicon follow the principles of structured networks. In addition, this study demonstrates that although a one-phoneme difference metric is commonly used to define the network of the mental lexicon, the 75% similarity metric may more closely represent the developing mental lexicon. In the future, researchers may consider using the 75% similarity metric to define parameters for networks examining the developing mental lexicon, rather than the one-phoneme difference network. Lastly, examining lexical connections beyond the level of
individual characteristics (e.g., word AoA, word frequency) and relationships between individual words (neighbors), this work promotes better understanding of how words relate to many others in the developing mental lexicon. This occurs by examining how individual words (nodes) fit within the larger network.

Clinical Implications

Further examination of lexical and sublexical characteristics of words existing as prominent nodes may reveal clinical applications for these findings. If evidence is found for phonological bootstrapping (in the 75% similarity network model), strategies for capitalizing on this phenomenon may be implemented. In addition, words with high betweenness may be implicated in the facilitation of lexical access. If this is the case, clinical strategies could be developed to take advantage of this occurrence. Extending this research to the bilingual population may lead to a more tactical approach to language learning. The results of this study provide opportunities to explore the application of network science in the clinical setting.

Future Directions

An analysis using mixed modeling to understand how the factors of word AoA, word frequency, PSA and biphone probability influence one another within both models of the developing mental lexicon would be informative. Mixed modeling may reveal interactions between lexical and sublexical characteristics of words that represent prominent nodes in the developing mental lexicon. This may promote better understanding of the influence of these characteristics in the network. Future studies could also select prominent nodes as stimuli for behavioral studies to assess and compare
which calculation of phonological similarity better reflects how we organize this information in our brain. For example, studying response rates to words with high degree and betweenness in RT studies could support the models presented in this study. Further examination is needed to understand the distribution of function words in the network, and the relation between age of acquisition of words and their location in the network. For example, behavioral studies examining differences in participant responses to function words with low, average and high degree could shed light upon the relationship between neighborhood density and function words. Comparing the male and female corpus could reveal similarities and differences in the input children of both sexes receive, particularly from books. A specific area that could be examined using this methodology is that of differences in lexical and sublexical characteristics in the input boys and girls receive during exposure to literature. Books read to boys and girls differ in vocabulary and subject matter, but it is unknown how the lexical and sublexical characteristics of these books may differ. Finally, replicating this study using a corpus of child directed speech could expand the knowledge of the lexical organization input children receive.
## APPENDIX

### Appendix A.

Example Words for High and Low Prominence Nodes

<table>
<thead>
<tr>
<th>Centrality Measure</th>
<th>Example Words</th>
<th>Function Words</th>
<th>Words Demonstrating Morphological Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One-phoneme difference network</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Degree</td>
<td>Bee, moo, high, we</td>
<td>At, be, by, for, he</td>
<td>Bees bought sighed, we’ll</td>
</tr>
<tr>
<td>High Betweenness</td>
<td>Rain, dumb, sit, bar</td>
<td>Are, will</td>
<td>Caught, raced, saw, signed</td>
</tr>
<tr>
<td>Low Degree</td>
<td>Splashed, shovel, question, carefully</td>
<td>Across, beside, into, onto, shall</td>
<td>Animals, boosters, curvy, remembered, trembled</td>
</tr>
<tr>
<td>Low Betweenness</td>
<td>Kinds, animal, easy, snuggle</td>
<td>-</td>
<td>Arrived, Celeste’s, pounded, snuggles, witch’s</td>
</tr>
</tbody>
</table>
### 75% Similarity Network

<table>
<thead>
<tr>
<th>High Degree (out-degree)</th>
<th>Are, ray, low, eyed, straw, thump</th>
<th>A, at, be, for, I, in, it, of, we, you</th>
<th>Adding, bears, it’ll, poured, went</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Betweenness</td>
<td>Heart, picked, order, tilt</td>
<td>Apart, upon</td>
<td>Bored, clears, lucky, tipped, trees</td>
</tr>
<tr>
<td>Low Degree (out-degree)</td>
<td>Foundation, furiously, organize, monkey, altogether</td>
<td>Beside, could, did, either, this, through, your</td>
<td>Airplanes, blowers, fluffy, pleased, wriggled</td>
</tr>
<tr>
<td>Low Betweenness</td>
<td>Tiaras, pumpkins, loaders, chuckling</td>
<td>Herself, this’ll</td>
<td>Ambulances, capricorns, gentlemen’s, tiaras, wobbled</td>
</tr>
</tbody>
</table>
Appendix B.

Example Words for Lexical and Sublexical Characteristics

**Note:** Word AoA is based on Kuperman’s (2012) rating scale. Using Mechanical Turk, 1,962 participants responded to a questionnaire of 30,000 words (approximately 300 words each), indicating the age in years at which they understood the use of these words. Word Frequency is based upon SUBTLEX corpus of 51 million words from English movie subtitles. Scores are reported as word occurrences per million words. Phonotactic probability (both PSA and initial biphone frequency) was calculated using Storkel & Hoover’s (2010) Child Mental Lexicon Calculator (CMLC), a corpus of approximately 5,000 word tokens. PSA is calculated by 1) summing the log frequency for all words in the corpus containing the given sound (or sound pair, in the case of initial biphone frequency) in a given word position, then 2) dividing by the sum of the log frequency of all the words in the corpus containing any sound in the given word position.

<table>
<thead>
<tr>
<th>One-phoneme difference Network</th>
<th>High Word AoA</th>
<th>Low Word AoA</th>
<th>High Word Frequency</th>
<th>Low Word Frequency</th>
<th>High PSA</th>
<th>Low PSA</th>
<th>High Initial Biphone Probability</th>
<th>Low Initial Biphone Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High Degree</strong></td>
<td>lair (11.63), soar (9.58)</td>
<td>wet (2.47), shoe (2.60)</td>
<td>to (22,677), me (9,241), we (9,011)</td>
<td>neigh (.240), hoe (.920)</td>
<td>sit (.096), so (0.9)</td>
<td>wheel (.0025), laid (.0032)</td>
<td>hat (.0082), whole (.0084)</td>
<td>zoo (.0005), show (.0014)</td>
</tr>
<tr>
<td><strong>Low Degree</strong></td>
<td>thwack (12.25), surveyor (11.67)</td>
<td>count (2.61), mother (2.63)</td>
<td>backing (2,009), telling (1,724)</td>
<td>grapple (.24), wiry (.35)</td>
<td>danced (.0797), hunter (.079)</td>
<td>eagle (.011), ivy (.012)</td>
<td>stand (.0249), straw (.0249)</td>
<td>hurray (.004), visit (.001)</td>
</tr>
<tr>
<td><strong>High Betweenness</strong></td>
<td>wren (9.35), mass (9.17)</td>
<td>men (3.11), come (3.32)</td>
<td>like (3,999), come (3,149)</td>
<td>wren (.37), whine (1.63)</td>
<td>city, (.093), sack (.088)</td>
<td>ride (.0032), red (.0057)</td>
<td>cat (.0123), crab (.0113)</td>
<td>pour (.0011), tall (.0013)</td>
</tr>
<tr>
<td><strong>Low Betweenness</strong></td>
<td>flitted (10.60), cable (9.20)</td>
<td>hands (2.74), animal (2.89)</td>
<td>calling (861.39), kinds (590.69)</td>
<td>grumbled (0.18), scrunch (0.22)</td>
<td>carry (.0815), Sunday (.0748)</td>
<td>easy (.0102), open (.0213)</td>
<td>scrunch (.0131), crawling (.0113)</td>
<td>jobs (.0002), pointed (.0006)</td>
</tr>
</tbody>
</table>
## 75% Similarity Metric

<table>
<thead>
<tr>
<th></th>
<th>High Word AoA</th>
<th>Low Word AoA</th>
<th>High Word Frequency</th>
<th>Low Word Frequency</th>
<th>High PSA</th>
<th>Low PSA</th>
<th>High Initial Biphone Probability</th>
<th>Low Initial Biphone Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High Degree</strong></td>
<td>mink (10.32), mend (9.11)</td>
<td>hand (2.74), jump (2.84)</td>
<td>it (18,896), and (13,387)</td>
<td>thump (1.94), rink (2.14)</td>
<td>sent (.099), plate (.096)</td>
<td>law (.0023), aid (.0036)</td>
<td>straw (.0249), scent (.0119)</td>
<td>eek (.0002), ocean (.0003)</td>
</tr>
<tr>
<td><strong>Low Degree</strong></td>
<td>primrose (12.53), uncoupled (12.53)</td>
<td>spoons (2.50), sleeping (2.79)</td>
<td>what (9,842), your (6,445)</td>
<td>suspenders (.04), uncoupled (.06)</td>
<td>scene (.090), pants (.080)</td>
<td>job (.0225), loving (.0237)</td>
<td>stir (.0249), stitched (.0249)</td>
<td>mush (.0001), job (.0002)</td>
</tr>
<tr>
<td><strong>High Betweenness</strong></td>
<td>pluck (9.78), slick (9.30)</td>
<td>ducks (3.57), trees (3.50)</td>
<td>still (789), start (340)</td>
<td>spills (.22), plump (1.47)</td>
<td>tilt (.0824), picked (.0804)</td>
<td>apart (.0356), plum (.0043)</td>
<td>star (.0249), steer (.0249)</td>
<td>thumbs (.0010), order (.0012)</td>
</tr>
<tr>
<td><strong>Low Betweenness</strong></td>
<td>excavators (13.00), muster (12.22)</td>
<td>animals (2.89), asks (2.89)</td>
<td>telling (1724.49), even (875.92)</td>
<td>gentlemen’s (.06), excavators (.08)</td>
<td>don’t (.080), sixties (.077)</td>
<td>enough (.0180), even (.0182)</td>
<td>strawberries (.0249), stations (.0249)</td>
<td>Edward’s (.0001), morning (.0005)</td>
</tr>
</tbody>
</table>
Appendix C.

Comparison of edges incident upon the example word *cat* in the 75% similarity network (left), and the one-phoneme difference network (right).
Appendix D.

Comparison of edges incident upon the example word *straw* in the 75% similarity network (left), and the one-phoneme difference network (right).
REFERENCES


Bonte, M., Parviainen, T., Hytönen, K., & Salmelin, R. (2006). Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. *Cerebral Cortex, 16*(1), 115-123.


Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive neuropsychology, 14*(1), 177-208.


Woodley, M. (2010). For preschoolers, lexical access is purely lexical: Neighborhood density effects in child speech perception and the emergent phoneme hypothesis. *UC Berkeley Phonology Lab Report, (395-405).*


