Comparing phonological networks

Philippa Shoemark

Master of Science by Research
School of Informatics
University of Edinburgh
2015
Abstract

The mental lexicon can be modelled as a network in which phonologically similar words are connected to one another. Examining the structural properties of phonological networks can reveal complex relationships among words, and studies combining network analysis with psycholinguistic experiments suggest that such relationships influence lexical processing. Comparing structural properties of phonological networks across languages and across different stages of language acquisition could potentially help to answer important questions about linguistic typology and the cognitive pressures that shape the acquisition and evolution of human language—but previous studies comparing phonological network structure across different lexicons have failed to account for confounding factors such as differences in the sizes of the lexicons and the inclusion or exclusion of inflectional variants. We show that these factors can substantially affect the statistics of a phonological network, and develop new methodologies for comparing phonological networks in more robust and informative ways.
Acknowledgements

I am immeasurably grateful to Sharon Goldwater, Rik Sarkar, and James Kirby for their unwavering support, patience, and encouragement, and for all of the invaluable suggestions, comments, and advice they have readily provided throughout the last year.

I would also like to extend my heartfelt gratitude to Andreas Kapouranis, Denitsa Saynova, George Papamakarios, and Harri Edwards for their diligence in maintaining the tranquility of the office, and to Charlie Nash, Clara Vania, Matt Pugh, Rafael Karampatsis, Sorcha Gilroy, and Theo Pavlakou for periodically reminding us about the existence of the outside world.

Last but not least, I am thankful to Samuel Abreu and to my family for their forbearance and moral support. I’m sure I could have done it without them, but I’m very glad I didn’t have to.

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.
Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Philippa Shoemark)
# Table of Contents

1 Introduction ................................. 1
  1.1 Motivation .................................. 1
  1.2 Overview ................................... 2
  1.3 Structure of the thesis ...................... 2

2 Background .................................. 4
  2.1 Technical background ...................... 4
    2.1.1 Definition of a phonological network ... 4
    2.1.2 Network statistics ...................... 5
    2.1.3 Network models ......................... 8
  2.2 Interpretations of Phonological networks .... 11
    2.2.1 Lexical processing ...................... 11
    2.2.2 Language acquisition and evolution .... 13
    2.2.3 Linguistic typology .................... 15
  2.3 Methodological issues .................... 15
    2.3.1 Size of the lexicon .................... 15
    2.3.2 Phoneme inventory size and word-length distribution .... 16
    2.3.3 Inflectional variants ................... 17

3 Comparing phonological networks across languages 19
  3.1 Background ................................. 19
  3.2 Brief overview of the new methodology .... 21
  3.3 Data collection and pre-processing .......... 21
    3.3.1 Construction of phonological networks ... 23
    3.3.2 Extraction of subgraphs ................. 23
  3.4 Empirical verification of the effects of word length distribution and phoneme inventory size .... 25
3.4.1 Comparison of pseudolexicons with fixed phoneme inventory size ........................................... 26
3.4.2 Comparison of pseudolexicons with fixed word length distribution ........................................... 28
3.4.3 Discussion .................................................................................................................................. 33
3.5 Robust comparison of phonological networks across eight languages ........................................... 34
  3.5.1 Comparison with Arbesman et al. (2010) ................................................................. 34
  3.5.2 Average word length ......................................................................................................... 35
  3.5.3 Number of edges ............................................................................................................... 39
  3.5.4 Fraction of nodes in the giant component ...................................................................... 39
  3.5.5 Fraction of edges in the giant component ...................................................................... 41
  3.5.6 Fraction of nodes which are isolates .............................................................................. 41
  3.5.7 Average shortest path length in the giant component .................................................. 44
  3.5.8 Average degree in the giant component ....................................................................... 47
  3.5.9 Assortative mixing by degree ....................................................................................... 47
  3.5.10 Transitivity ...................................................................................................................... 47
3.6 Conclusion ................................................................................................................................... 47

4 Comparing phonological networks across age groups ................................................................. 53
  4.1 Background ............................................................................................................................. 53
  4.2 Comparison of child, caregiver, and adult phonological networks ........................................... 55
    4.2.1 Data ...................................................................................................................................... 56
    4.2.2 Raw lexicons ..................................................................................................................... 57
    4.2.3 Sublexicons with fixed word-length distribution ................................................................. 58
    4.2.4 Discussion .......................................................................................................................... 61
    4.2.5 Conclusion .......................................................................................................................... 65

5 Conclusion ....................................................................................................................................... 67

A A
  A.1 Tables of frequency thresholds and resulting sublexicon sizes ............................................ 69

B B
  B.1 List of CHILDES corpora from which the Child Speech lexicon was extracted ..................... 72

Bibliography ....................................................................................................................................... 73
Chapter 1

Introduction

What makes some words easier to learn, recognise, produce, or recall than others? Studies suggest the mental lexicon is structured in such a way that phonologically similar words are connected to one another, and the number of other words to which a given word is connected (known as *neighbourhood density* in the psycholinguistics literature) affects the ease with which it can be accessed from or integrated into the lexicon (Luce and Pisoni, 1998), (Harley and Bown, 1998), (Vitevitch, 2002), (Ziegler et al., 2003). For example, Vitevitch (2002) found that words with many neighbours were produced more quickly and accurately than words with few neighbours in a word pair repetition task, a tongue-twister repetition task, and a picture naming task. In order to better understand how the structure of the mental lexicon affects the processing of words, psycholinguists have begun to model it as a complex network or graph, and quantitatively analyse its structure using measures from graph theory and network science. While most of these phonological network analyses have been focused on adult English lexicons, in this thesis we compare the structures of phonological networks across different languages, and we also compare the structure of children’s phonological networks with those of adults. We discuss methodological issues in previous studies that have compared phonological network structures across different languages or age groups, and develop new methodologies for comparing the structure of phonological networks in robust and meaningful ways.

1.1 Motivation

Arbesman et al. (2010) compared the phonological networks of five different languages and noted that they all had several properties in common, which are distinct from the
properties typically observed in other types of complex networks. They suggested that some of these properties may be advantageous for lexical processing, and suggested that phonological networks might be optimised for robustness to node removal and to facilitate efficient search. Carlson et al. (2011) compared phonological networks representative of children’s and adults’ productive lexicons, and suggested that the structure of children’s lexicons might be even more advantageous for stability and searchability than that of adults’. They hypothesised that children may preferentially acquire certain words in order to induce this structure in their developing lexicons, because of the apparent processing advantages it affords. These are compelling claims, but neither of these studies have satisfactorily addressed the methodological issues that arise when comparing phonological network structure across different lexicons, and so there are doubts about the validity of their analyses.

1.2 Overview

Our original goal was to develop a robust method for comparing lexicons of different sizes, but as the project progressed, it became apparent that there are additional confounding factors which make this objective more difficult. Stella and Brede (2015) argued from a theoretical standpoint that the distribution of word lengths in a lexicon and the size of its phoneme inventory constrain its phonological network topology. We demonstrate this issue empirically, and show that controlling for these confounds can dramatically change the interpretation of phonological network statistics. Another confounding factor which is rarely addressed in the literature is the choice between including or excluding inflectional variants in a phonological network. We explore the impact of this choice on the statistics of phonological networks, and show that analysing the differences in the phonological network statistics of lexicons which do and don’t include inflectional variants could be a promising avenue for linguistic typological research.

1.3 Structure of the thesis

In Chapter 2 we provide some technical background and define statistics and terminology from the network science literature. We discuss the various ways in which phonological networks have been interpreted, and explain the methodological issues that we aim to address in this thesis.
In Chapter 3 we develop a new methodology for comparing phonological networks across different languages, and demonstrate how it can provide novel insights in a comparative analysis of the phonological network structures of eight different languages.

In Chapter 4 we compare the phonological network structures of lexicons derived from corpora of child, child-directed, and adult-directed speech, and show that the apparently optimised structure of children’s phonological networks can be parsimoniously explained as an artefact of their shorter word-length distributions.

While Chapters 3 and 4 are in some ways distinct lines of inquiry, there are important methodological issues that underlie both, and the aim of this thesis is to demonstrate these issues empirically and to show how they can be overcome.
Chapter 2

Background

2.1 Technical background

2.1.1 Definition of a phonological network

A phonological network is a graph in which nodes/vertices correspond to the phonological word forms\(^1\) in a lexicon, and undirected, unweighted edges link nodes which are phonologically similar according to some metric. In the present work, as in most of the previous literature on phonological networks, we use a one-phoneme edit-distance metric, according to which edges are placed between any two nodes that differ only by the insertion, deletion, or substitution of a single phoneme. In other words, edges link any two nodes with a Levenshtein distance of one. Phonological networks could in principal be constructed using other similarity metrics, for example Bailey and Hahn (2001) proposed an edit-distance metric in which the cost of a substitution depends on the number of distinctive features the phonemes have in common – so that the word ‘duck’ would be more similar to the word ‘tuck’ than to the word ‘muck’, since the phoneme /d/ has more distinctive features in common with /t/ than with /m/. However, despite its crudity, there is a large body of evidence to suggest that the simple Levenshtein distance metric has psychological validity, as various properties of words derived from their position in phonological networks which use this metric have been shown to predict behaviour on a range of language processing tasks, as we will review in §2.2

\(^1\)By phonological word forms, we mean phonemic transcriptions of words as opposed to their orthographic forms.
2.1.2 Network statistics

Networks can be characterised by a variety of statistics designed to quantify their topological properties. These statistics can be used to compare different networks directly, as well as to classify them based on their similarity to canonical network models. In this section we introduce and define each of the network statistics that will be referred to throughout the rest of the thesis.

2.1.2.1 Average degree

The degree $k$ of a node $v$ denotes the number of neighbours it has – that is, the number of other nodes in the network to which $v$ is connected. In a phonological network, a word’s degree is equivalent to its neighbourhood density, i.e. the number of other words in the lexicon from which it differs by a single phoneme. While degree is a property of individual nodes, the network as a whole can be characterised by its average degree $\langle k \rangle$, which is the mean of the degrees of each of the nodes. The average degree can be computed with the equation

$$\langle k \rangle = \frac{2|E|}{|V|},$$

where $E$ is the set of edges in the network, and $V$ is the set of nodes in the network.

2.1.2.2 Degree distribution

A network can also be characterised by its degree distribution $p(k)$, which gives the fraction of nodes in the network that have degree $k$, or equivalently the probability that a node chosen uniformly at random from the network will have degree $k$, for any $k \geq 0$.

2.1.2.3 Fraction of nodes in the giant component

A connected component of a network is a maximal subgraph in which any node can be reached from any other by traversing the edges in the subgraph. Connected components are sometimes referred to as ‘islands’, and nodes which do not have any edges are referred to as ‘hermits’ or ‘isolates’. A common phenomenon in large, complex networks is the existence of a ‘giant component’ which contains a very large fraction of the nodes— much larger than the fraction of nodes in any other connected component. The larger the giant component, the more likely it is that a pair of nodes sampled uniformly at random from the network will be reachable from one another. Hence
the fraction of nodes in the giant component can be used to characterise the global connectivity of a network.

### 2.1.2.4 Average clustering coefficient

The clustering coefficient $c$ of a given node $v$ measures the ‘cliquishness’ of its neighbourhood, and is defined as the ratio of the number of edges that exist between the neighbours of $v$ to the number of edges that would exist between them if they were all connected to one another. Equivalently, the clustering coefficient of a node $v$ can be thought of as the ratio of the number of triangles of which $v$ is a member to the number of triples of which $v$ is the central node, where a triple is a set of three nodes $u$, $v$, and $w$ such that there exists an edge $\{u,v\}$ between $u$ and $v$ and an edge $\{v,w\}$ between $v$ and $w$; and a triangle is a set of three nodes $u$, $v$, and $w$ such that there exist three edges $\{u,v\}$, $\{v,w\}$, and $\{w,u\}$. The clustering coefficient can be calculated using the equation

$$c(v) = \frac{m(v)}{t(v)},$$

where

$$m(v) = |\{\{u,w\} \in E : \{u,v\} \in E \text{ and } \{v,w\} \in E\}|$$

and

$$t(v) = \frac{k(v)(k(v) - 1)}{2}.$$

We can see from this equation that the clustering coefficient is undefined for nodes with $k < 2$, since the denominator $t(v)$ reduces to zero for such nodes. The network as a whole can be characterised by the average clustering coefficient $\langle c \rangle$, which measures the cliquishness of a typical neighbourhood and is defined as the mean of the clustering coefficients of each of the nodes in the network. This leaves open the issue of how to deal with nodes for which the clustering coefficient is undefined: sometimes researchers set $c$ to zero for nodes with $k < 2$; sometimes they set $c$ to one for such nodes; and sometimes they simply exclude these nodes from the calculation of the mean. Hence the average clustering coefficient is variously defined in the network science literature as

$$\langle c \rangle = \frac{1}{|V|} \sum_{v \in V} c(v),$$

where $c(v)$ is set either to zero or to one for every node $v$ having $k(v) < 2$; or as

$$\langle c \rangle = \frac{1}{|V'|} \sum_{v \in V'} c(v),$$
where \( V' \) is the set of nodes with \( k \geq 2 \). As Schank and Wagner (2004) have shown, the choice of definition can make a considerable difference to the value of \( \langle c \rangle \)— but unfortunately, researchers do not always specify which definition they have used. In this thesis, we will report values of \( \langle c \rangle \) given by the first equation, with \( c \) set to zero for nodes with \( k < 2 \).

### 2.1.2.5 Transitivity

An alternative statistic for characterising the cliquishness of a network is the transitivity \( T \), which measures how many of the triples in the network are also triangles. It is given by the formula

\[
T = \frac{3 \delta}{\tau},
\]

where \( \delta \) is the number of triangles in the network and \( \tau \) is the number of triples in the network. Schank and Wagner (2004) have pointed out that transitivity can also be formulated as

\[
T = \frac{\sum_{v \in V'} m(v)}{\sum_{v \in V'} t(v)},
\]

and is equivalent to a weighted average clustering coefficient, where the clustering coefficient of each node \( v \) is weighted by the number of triples of which \( v \) is the central node. Hence the transitivity and the un-weighted average clustering coefficient have the same value in networks where all nodes have the same degree or the same clustering coefficient, but can differ considerably in networks with heterogeneous degrees or clustering coefficients.

### 2.1.2.6 Average shortest path length

The shortest path length \( d \) between two nodes (also known as the geodesic distance) is the minimum number of edges that must be traversed in order to get from one node to the other. The average shortest path length is then the mean of the geodesic distances between all pairs of nodes in the network, and is given by the equation

\[
\langle d \rangle = \frac{1}{|V||V| - 1} \sum_{v,w \in V} \frac{d(v,w)}{|V||V| - 1},
\]

where \( d(v,w) \) is the geodesic distance between nodes \( v \) and \( w \). In networks with more than one connected component, the calculation of the average shortest path length is complicated by the fact that the geodesic distance between a node in one component and a node in a separate component is infinite— and hence the mean geodesic distance
is also infinite. In order to meaningfully compare the average shortest path lengths of networks which are not fully connected, researchers sometimes just report the average shortest path length in the giant component, or sometimes they calculate the average shortest path length in each component and then report the mean of these values.

2.1.2.7 Assortative mixing by degree

A network in which nodes tend to connect to other nodes that have similar degrees to their own is said to exhibit assortative mixing by degree, whilst a network in which nodes tend to connect to other nodes that have dissimilar degrees to their own is said to exhibit disassortative mixing by degree. The extent to which a network exhibits assortative or disassortative mixing can be quantified using the degree assortativity coefficient $r$, which is defined as the Pearson correlation coefficient of the degrees of the nodes at either ends of an edge (Newman, 2002). The degree assortativity coefficient lies in the range $-1 \leq r \leq 1$, where $r = -1$ indicates that the higher a node’s degree is, the lower the degrees of its neighbours are; $r = 0$ indicates that there is no correlation between the degree of a node and the degrees of its neighbours; and $r = 1$ indicates that each node connects only to other nodes with the exact same degree as its own.

2.1.3 Network models

Network models are designed to generate networks with specific topological properties using simple growth or rewiring algorithms. Such models can be used to investigate how particular topological properties affect dynamic processes which operate on networks, and comparisons between empirically observed networks and networks generated by models can guide hypotheses about what sorts of mechanisms could plausibly have given rise to the observed network structure. We will now give a brief overview of some network models to which phonological networks have been compared in previous literature.

2.1.3.1 Erdős Rényi Random Graph model

The Erdős Rényi Random Graph model generates networks by starting with a disconnected set of $N$ nodes, and placing an edge between each distinct pair of nodes with probability $p$. Hence the presence or absence of an edge in an Erdős Rényi Random Graph is independent of the presence or absence of any other edge, and the total number of edges is a random variable with expected value $\binom{N}{2} p$. The degree distribution
of an Erdős Rényi Random Graph is given by

\[ p(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}, \]

and for large \( N \) can be approximated by a Poisson distribution with a peak at \( p(\langle k \rangle) \) (Albert and Barabási, 2002). Since edges are placed at random, the degree assortativity coefficient is 0 in the limit of large \( N \). The expectation of the clustering coefficient is given by

\[ \langle c \rangle = p = \frac{\langle c \rangle}{N}, \]

since the edges are independent and so the probability that two of a node’s neighbours are connected is just the probability that any randomly sampled pair of nodes are connected. The average shortest path length \( \langle d \rangle \) in an Erdős Rényi Random Graph is proportional to \( \ln(\langle k \rangle) \ln(N) \) (Albert and Barabási, 2002).

### 2.1.3.2 Small-world models

Like Erdős Rényi Random Graphs, many biological, social, and technological networks in the real world have short average shortest path lengths relative to their size. However, average clustering coefficients in these real-world networks are often much higher than would be expected in an Erdős Rényi Random Graph with the same number of nodes and degree distribution. Networks with these properties are said to exhibit the small-world phenomenon, since even though they contain a large number of nodes, any two nodes are likely to be connected through a short chain of intermediaries. Watts and Strogatz designed a model which could generate networks with both short average path lengths and high clustering, by starting with a regular ring lattice and randomly rewiring each edge with probability \( p \), creating short cuts between far-away nodes. Varying the value of \( p \) allows one to interpolate between a regular ring lattice, in which the average clustering coefficient \( \langle c \rangle \) is a constant and the average shortest path length \( \langle d \rangle \) scales linearly with \( N \); and an Erdős Rényi Random Graph, in which \( \langle c \rangle \) is inversely proportional to \( N \) and \( \langle d \rangle \) scales logarithmically with \( N \). Watts and Strogatz showed that as the rewiring probability \( p \) is increased from 0, the value of \( \langle d \rangle \) drops rapidly towards that of an Erdős Rényi Random Graph, but over a broad interval \( \langle c \rangle \) remains similar to that of a regular ring lattice.

Kleinberg (2000) generalised the Watts-Strogatz model by adding an additional tunable parameter \( \alpha \), which allows the probability of a short-cut being added between two nodes to be determined as a function of their lattice distance. In Kleinberg’s model,
networks are generated by starting with a regular lattice and adding an extra edge from each node \( u \) to another node \( v \) chosen with probability proportional to \( r^{-\alpha} \), where \( r \) is the lattice distance between \( u \) and \( v \). When \( \alpha \) is set to zero, \( v \) is chosen uniformly at random, and the model is equivalent to the Watts-Strogatz model. Kleinberg showed that if and only if \( \alpha \) is equal to the number of dimensions in the lattice, a decentralised search algorithm can operate on the network with an expected delivery time proportional to \( (\log N)^2 \).

2.1.3.3 Barabási-Albert model

Whereas Erdős Rényi Random Graphs are characterised by binomial degree distributions, the degree distributions of networks in the real world often follow power laws or exponential functions. While small-world models can reproduce the high clustering coefficients observed in real-world networks, their degree distributions are similarly shaped to those of Erdős Rényi Random Graphs. The Barabási-Albert (Barabási and Albert, 1999) model was designed to generate networks with power law degree distributions, using a growth mechanism with preferential attachment. Starting with a small number of nodes, at each time step a new node is added and connected to \( m \) nodes already present in the network, chosen with probability proportional to their degree. Since each new node is connected to existing nodes as soon as it is added, this model produces networks with a single connected component. The average shortest path length in a network generated by the Barabási-Albert model is shorter than that found in an Erdős Rényi Random Graph, but the clustering coefficient is considerably higher (Albert and Barabási, 2002).

2.1.3.4 Random growth model

In the random growth model introduced by Callaway et al. (2001), nodes are also added one-by-one to the network, but rather than using a preferential attachment rule, edges are placed with probability \( p \) between pairs of nodes chosen uniformly random. As a result, new nodes do not necessarily connect to any existing nodes when they are first added, and the randomly grown networks tend to contain isolates and connected components of various sizes, with a phase transition at which a giant component appears. These characteristics are also exhibited by Erdős Rényi Random Graphs, except that the giant component emerges at a much lower edge density in the randomly grown graphs. In further departures from the structure of Erdős Rényi Random Graphs, the
networks generated by Callaway et al.’s random growth model have exponential degree distributions, and positive assortative mixing by degree.

2.2 Interpretations of Phonological networks

We will now discuss the various ways in which a detailed understanding of the global structure of phonological networks could potentially enhance our understanding of human language.

2.2.1 Lexical processing

The original motivation for analysing the structure of phonological networks was born out of the psycholinguistics literature on neighbourhood density. Many studies in this field have indicated that the number of words in the lexicon from which a given word differs by a single phoneme affect the speed and accuracy with which it can be processed (Luce and Pisoni, 1998), (Harley and Bown, 1998), (Vitevitch, 2002), (Ziegler et al., 2003). For example, Vitevitch (2002) found that words with many phonological neighbours were produced more quickly and accurately than words with few phonological neighbours in a word pair repetition task, a tongue-twister repetition task, and a picture naming task. This prompted researchers to explore more nuanced measures of lexical similarity neighbourhoods, and lead to Vitevitch’s (2008) seminal paper in which he conceptualised the phonological similarity relationships among words in the lexicon as a complex network. In addition to degree or neighbourhood density, various other properties of words derived from their position in phonological networks have been shown to predict behaviour on language processing tasks. For example, a word’s clustering coefficient has been shown to predict the speed and accuracy of spoken word recognition (Chan and Vitevitch, 2009), (Altieri et al., 2010); spoken word production (Chan and Vitevitch, 2010); and retrieval of words from long and short term memory (Vitevitch et al., 2012). Vitevitch and Goldstein (2014) used Borgatti’s (2006) Negative Key Player algorithm to identify a set of 25 keywords in the English lexicon, which if removed would cause maximal partitioning of the phonological network. In three conventional psycholinguistic tasks, these keywords were responded to more quickly and accurately than a carefully controlled set of foil words.

These results are consistent with a theory of lexical processing proposed by Chan and Vitevitch (2009), in which activation spreads through the mental lexicon via the
phonological similarity links among words. If lexical processing does depend on the spread of activation through a phonological network, then the efficiency of lexical processing mechanisms will be influenced by the global structure of the network. Therefore, in order to understand precisely what sorts of mechanisms within this ‘spreading activation’ framework can account for the impressive speed and accuracy with which we process language, as well as the occasional failures we experience such as tip-of-the-tongue states and slips of the ear, we require a detailed understanding of how phonological networks are structured, and how that structure affects the dynamics of spreading processes. For example, it has been demonstrated both analytically (Newman, 2003) and with simulations (Kiss et al., 2008), (Payne et al., 2009), that the proportion of a network to which a commodity (such as activation, gossip, or disease) can spread, as well as the speed with which it spreads, are determined by several factors including the density of the network, the degree distribution and the coefficient of assortative mixing by degree. Arbesman et al. (2010) compared the global characteristics of phonological networks derived from lexicons representing five languages from different language families, and found that all five of them had unusually high levels of assortative mixing by degree, and degree distributions that were reasonably well fit to truncated power laws. In addition, all five phonological networks were less dense (in that they had a lower ratio of edges to nodes) than comparably sized random graphs. High assortative mixing by degree facilitates spreading processes on networks with low densities, but – at least in networks with power-law degree distributions (Kiss et al., 2008), cf. (Payne et al., 2009) – it also limits the proportion of the network that will ultimately be reached. Conversely, in networks with dissassortative mixing by degree, activation tends either to barely spread at all, or else to flood the entire network. As Vitevitch (2008) has suggested, the particular set of characteristics observed in phonological networks could be beneficial for lexical processing in that the spread of activation from acoustic-phonetic input to a limited set of candidates would mean that “not every word in the lexicon would have to be considered and rejected as a potential lexical candidate [...] ensuring rapid, seemingly automatic retrieval of the correct lexical candidate from the lexicon, even in less than ideal listening conditions.” (p. 9)

Arbesman et al. (2010) also found that some structural characteristics varied among the five phonological networks they analysed— for example, the average shortest path length in the Mandarin network was twice that of Hawaiian. Differences in phonological network structure across languages might mean that different processing strategies
are better suited to different languages. However, as Arbesman et al. have acknowledged, it is not yet clear how meaningful the differences in their results are, due to methodological issues which will be explained in §2.3.

2.2.2 Language acquisition and evolution

Several researchers have used the global structural characteristics of phonological networks to guide inferences about their growth mechanisms. Vitevitch (2008) observed that a phonological network representing the vocabulary of an average English-speaking adult had many separate connected components, a large number of isolated nodes, high assortative mixing by degree, and a degree distribution that was better fit by an exponential function than a power law. One growth mechanism which produces networks with qualitatively similar properties is the random growth model introduced by Callaway et al. (2001), in which at each time step, a new node is added to the network and edges are placed with probability $p$ between pairs of nodes chosen uniformly at random. Vitevitch suggested that similar processes might influence the acquisition of phonological word-forms, and could thus account for the occurrence of the aforementioned properties in mature phonological networks. He argued that the process in which new edges are added between random pairs of nodes was compatible with the Lexical Restructuring Hypothesis (Metsala, 1997) – the idea that early word representations are relatively course-grained, but become more detailed as the lexicon grows.

However, care must be taken when inferring growth models from the current state of a network, as multiple mechanisms can produce the same structures. Gruenenfelder and Pisoni (2009) criticised Vitevitch’s suggestion that the characteristics of the English phonological network reflect processes involved in its growth. They first pointed out that when we acquire a language, there is a pre-defined target lexicon to which we must converge, regardless of the learning strategy used. Thus, the phonological network structure of that target lexicon cannot really tell us anything about the order in which the words are acquired. Gruenenfelder and Pisoni conceded that the growth-plus-lexical-restructuring model could instead be applied to the case of language evolution, and used to explain how new words are added to the common lexicon, as opposed to the lexicon of an individual language learner. But they then went on to argue that many of the structural characteristics of the adult English phonological network can in fact be accounted for without appealing to any sort of growth mechanism—because they are natural consequences of the way the nodes and edges are defined.
Greunenfelder and Pisoni showed that qualitatively similar characteristics are in fact exhibited by phonological networks constructed from ‘pseudo-lexicons’, in which the words are generated completely independently of one another, and consist of sequences of phonemes chosen uniformly at random from the English phoneme inventory.

On the other hand, Stella and Brede (2015) recently pointed out that while some higher order characteristics of random phonological networks are qualitatively similar to those of the adult English phonological network, lower order statistics such as the degree distribution, the number of edges, and the distribution of component sizes are significantly different. Hence, they argued, there is still room for growth models which can account for properties of phonological networks that aren’t mere artefacts of the way they are defined. They constructed various growth models, and found that the model which best fit the structure of the adult English phonological network was one in which potential new words are preferentially added to the network if they are short and phonologically similar to at least one existing word—subject to constraints on the maximum degree and transitivity of any given node, which they suggested might arise from pressures to avoid too much confusability. This model is compelling, but ignores the fact that the English lexicon did not arrive at its present state just by incorporating more and more new words: words can also become obsolete, resulting in their removal from the lexicon; systematic sound changes occur, effectively resulting in a systematic re-wiring of the lexicon; and even the phoneme inventory can change over time, meaning the underlying space of possible words changes, too.

In an attempt to determine whether children preferentially acquire lexicons with structural characteristics that facilitate lexical processing and/or further lexical acquisition, Carlson et al. (2011) compared the global structural characteristics of three phonological networks derived from a corpus of normal adult speech, a corpus of child-directed speech, and a corpus of speech by children. Their results indicated that the phonological networks derived from child and child-directed speech were more densely connected than the network derived from adult speech, having higher assortative mixing by degree, higher average clustering coefficients, and a larger proportion of nodes in the largest connected component. Furthermore, the child speech network had a shorter average shortest path length between nodes than either the adult speech network or the child-directed speech network. Carlson et al. suggested that these properties may enhance the stability of children’s phonological networks and make them more efficiently searchable, supporting the notion that “children favor a lexicon whose global phonological structure allows for efficient functioning during speech production
and comprehension.” (p. 10) However, as we discuss in §2.3, the fact that the three networks are differently sized and have different word-length distributions confounds the interpretation of differences in their higher-order structural characteristics. In Chapter 4 we address these issues in detail, and present a more parsimonious explanation of the structural differences across the adult, child-directed, and child speech networks.

### 2.2.3 Linguistic typology

Regardless of whether phonological networks are psychologically real, and regardless of whether we can reasonably make inferences about their growth mechanisms based on static snapshots, analysing the structural characteristics of phonological networks could also be useful as a way of comparing different languages typologically, and better understanding the similarities and differences between their lexicons and phonotactic properties. Particular phonotactic or morphological traits might be reflected in phonological network structure, and so measures of network properties could potentially be used to quantify the degree to which different languages exhibit such traits, or the overall similarity of one language to another.

Whatever the ultimate application may be, when comparing the characteristics of phonological networks constructed from different lexicons, there are important methodological issues to consider. In §2.3 we discuss some of these issues, and in Chapters 3 and 4 we outline the approaches we have taken to address them.

### 2.3 Methodological issues

#### 2.3.1 Size of the lexicon

One issue is that many of the measures used to characterise topological properties of networks are influenced by the number of nodes in the network. Since we cannot know every word that exists in a language, or in a specific individual’s vocabulary, we approximate the average speaker’s mental lexicon with some subset of the (ever-changing) set of all words in the language. Usually, this subset is the set of unique words that occur in a particular corpus. Thus, the number of nodes in a phonological network depends on the number and diversity of word tokens in the corpus it was drawn from. For a handful of languages, very large phonologically transcribed corpora covering a wide range of domains are available; but for the vast majority of languages, and particularly for child and child-directed speech, resources are scarce and so the
size of the lexicon we can extract is limited. Therefore, it is important to consider the extent to which measures of network characteristics depend on the number of nodes in the network. If we compare differently-sized lexicons using a network statistic which is sensitive to the number of nodes, then how can we know whether differing values reflect meaningful differences in the structures of the lexicons, or are merely consequences of the fact that one lexicon was derived from a larger corpus than the other?

2.3.2 Phoneme inventory size and word-length distribution

Another important issue concerning the interpretation of network statistics arises from a peculiarity in the way that nodes and edges are defined in phonological networks. The issue was first raised by Gruenenfelder and Pisoni (2009), and was given a more formal exposition by Stella and Brede (2015). What is peculiar about phonological networks is that the nodes represent sequences of discrete symbols (phonemes) sampled from a fixed inventory, and the presence or absence of an edge between two nodes is determined by the overlap of these sequences. As Stella and Brede explain, “the set of all possible combinations of symbols (together with the phonetic similarity metric) defines a space, of which the actual word repertoire is a subset.” (2015)[pp.3-4]. Thus, the topology of a phonological network is constrained by the topology of the underlying space. To see how, consider the set of all possible ‘words’ that could hypothetically be formed using a given phoneme inventory $P$. As Stella and Brede have pointed out, the number of possible words of length $\ell$ is given by $|P|^\ell$, and the number of length-$\ell$ neighbours of a length-$\ell$ word is given by $(|P| - 1) \ell$. Note that the number of possible words of length $\ell$ scales exponentially with $\ell$, while the number of length-$\ell$ neighbours of a length-$\ell$ word scales linearly with $\ell$. We can extend this line of reasoning to the total number of possible neighbours of a length-$\ell$ word $w$, by recognising that the number of length-$(\ell + 1)$ neighbours of $w$ is given by $(\ell + 1)|P| - \ell$, and the number of possible length-$(\ell - 1)$ neighbours of $w$ is given by $\ell - g$, where $g$ is the number of consecutive phoneme repetitions in $w$. The number of possible words of length $l$ that have $g$ consecutive phoneme repetitions is given by $\left(\frac{l-1}{g}\right) \cdot |P| \cdot (|P| - 1)^{\ell - g - 1}$. Hence, if we randomly sample a pair of words $w$ and $v$ from the set of all possible words, such that $v$ is at most one phoneme longer than $w$, then the shorter their lengths are, the more likely it is that the sampled words will be neighbours. By this line of reasoning, it becomes apparent that lexicons whose word-length distributions are biased
towards shorter words “will naturally induce larger component sizes than word-length distributions that account for relatively more long words.” (Stella and Brede, 2015, p.7) Furthermore, as Gruenenfelder and Pisoni (2009) have pointed out, the probability that two length-$\ell$ neighbours of a length-$\ell$ word are also neighbours of each other is $\frac{1}{\ell}$ — so it follows that lexicons with word-length distributions biased towards shorter words would naturally have higher average clustering coefficients, too. Since the number of possible words also scales faster with $|P|$ than does the number of possible neighbours of a given word $w$, we would also expect the size of the phoneme inventory to affect the connectivity of a phonological network— albeit by a smaller factor than the word-length distribution. These theoretical arguments will be verified empirically in Section §3.4. While differences in word-length distributions and phoneme inventory size across lexicons might be interesting and informative in their own right, we do not need to construct and analyse phonological networks in order to discover them. If we wish to determine whether phonological network structure differs across lexicons in more revealing ways, which might provide novel insights into linguistic typology or suggest possible mechanisms that could be guiding their formation, then it will be necessary to control for differences in the word-length distributions and phoneme inventory sizes.

### 2.3.3 Inflectional variants

We are also forced to grapple with the fundamental issue of what constitutes a word. Should phonological similarity relations be defined over all inflectional variants, or only over lemmas? The phonological network in Vitevitch’s (2008) study was derived from a pocket dictionary, and thus the nodes corresponded to lemmas. Most of the phonological network analyses that have followed (e.g. Gruenenfelder and Pisoni (2009), Stella and Brede (2015)) appear also to have used lexicons consisting only of lemmas, although this is not always made explicitly clear: Arbesman et al. (2010) compared the phonological network structures of five languages, but while they used lexicons derived from dictionaries (and so presumably restricted to lemmas) for some of these languages, the databases from which their Spanish and Basque lexicons were derived contain entries for inflected forms as well as for lemmas. Judging by the reported sizes of the Spanish and Basque networks (Arbesman et al., 2010, p.681), it seems as though inflectional forms were indeed included in the Spanish and Basque lexicons — which calls into question the validity of directly comparing them with lex-

---

2LEXESP (Sebastián-Gallés, 2000) and E-Hitz (Perea et al., 2006), respectively
icons that contained only lemmas. Carlson et al. (2011) compared phonological networks derived from corpora of child speech, child-directed speech, and adult-directed speech, and stated explicitly that unlemmatised forms were included. Yet conversely, the authors of an earlier study which also compared the structures of child and adult lexicons (Coady and Aslin, 2003) stated explicitly that they had removed inflected forms. In summary, there is clearly a lack of consensus about whether inflectional variants should be included or excluded in representations of the mental lexicon, and in some cases the nature of the available data may preclude any choice in the matter. It is therefore important to investigate how the inclusion or exclusion of inflectional variants affects the topological properties of phonological networks. For example, one might expect that in languages with fusional morphology, the inclusion of inflectional variants would result in a higher average degree and a higher average clustering coefficient, since there would be many wordforms which are all variants of the same lemma and differ from one another only by a single morpheme (which could consist of just a single phoneme).
Chapter 3

Comparing phonological networks across languages

3.1 Background

Arbesman et al. (2010) compared the phonological networks of five different languages, and noted that they shared several distinctive structural characteristics, compared with other complex networks described in the network science literature. These include both higher assortative mixing by degree and a smaller percentage of nodes in the largest connected component. All five phonological networks were also found to exhibit small-world characteristics, and their degree distributions were well fit by truncated power-laws. The authors argued that “the properties of these networks suggest explanations for various aspects of linguistic processing and hint at deeper organisation within the human language”. (p. 679)

Insofar as models involving some sort of ‘activation’ spreading through a phonological network are a plausible account of lexical processing, the first claim is relatively uncontroversial. Regardless of how or why the five phonological networks came to have the observed structural characteristics, these characteristics have implications for the efficacy and efficiency of processes that might operate on such networks, as we discussed in §2.2.1. However, the claim that the observed structural characteristics hint at ‘deeper organisation’ within language implies an assumption that these characteristics deviate from what would be expected to occur in phonological networks constructed from lexicons which contain no such deeper organisation— i.e. lexicons consisting of random phoneme sequences. As Gruenenfelder and Pisoni (2009) and Stella and Brede (2015) have shown, some of the reported characteristics of the English phonological
network – i.e. small-world characteristics and the power-law like portion of the degree distribution – occur also in phonological networks constructed from randomly generated pseudolexicons with the same word length distribution. But other characteristics, including the exceptionally high assortative mixing by degree and the proportion of nodes in the largest connected component, were confirmed to be deviations from the expected characteristics of random lexicons. Interestingly, whereas Arbesman et al. (2010) noted that the proportion of nodes in the largest connected component of the English phonological network is small compared to networks constructed using the random growth process of (Callaway et al., 2001); Gruenenfelder and Pisoni (2009) and Stella and Brede (2015) have shown that it is actually much larger than the proportion of nodes in the largest connected component of a phonological network built from random phoneme sequences. Thus, what does warrant further explanation is not that the proportion of nodes in the largest connected component of the English phonological network is smaller than expected, but that it is actually larger than expected. This points to the importance of choosing appropriate random baselines.

While various qualitative characteristics are common to all five of the phonological networks analysed by Arbesman et al. (2010), there are considerable quantitative differences in the giant component sizes and average shortest path lengths. However, it is not clear whether these quantitative differences reflect meaningful differences in the structures of the lexicons of different languages, or whether they are artefacts of the methodological issues discussed in Section §2.3. Based on their observation that the number of hypothetically possible length-$\ell$ words scales much more quickly with $\ell$ than does the number of possible neighbours of a length-$\ell$ word, Stella and Brede (2015) argue that one would expect lexicons with a relatively high proportion of short words to have larger giant components than lexicons which contain relatively more long words. Since the number of possible words also scales faster with $P$ than does the number of possible neighbours of an arbitrary word, we would also expect the size of the phoneme inventory to affect the size of the giant component— albeit by a smaller factor than the word length distribution. In Section §3.4 we will verify empirically that both the word length distribution and the phoneme inventory size affect the connectivity of randomly sampled pseudolexicons. Two further confounding factors in the interpretation of Arbesman et al.’s results are the fact that the lexicons vary markedly in size, and the fact that some of the lexicons apparently include inflectional variants while others contain only lemmas. To address these issues, we will now present a more robust method for comparing phonological networks across different languages, and in
Section §3.5 we will apply this methodology to a comparative analysis of phonological network structure across eight different languages.

### 3.2 Brief overview of the new methodology

Where possible, we obtain for each language a lexicon that includes inflectional variants as well as a lexicon restricted to lemmas, and construct separate phonological networks for each lexicon type. This allows us to investigate the effect of including or excluding inflectional variants on the network statistics. After validating empirically the theoretical argument made by Stella and Brede (2015), we compare each of our lexicons with randomly generated pseudolexicons that are individually matched in word length distribution and phoneme inventory size, in order to determine the extent to which the phonological network structures of the different lexicons deviate from random expectations. As discussed in section §2.3.1, lexicons effectively approximate the average speaker’s vocabulary by sampling words according to their frequency of use, since the more frequently a word is used, the more likely it is to occur in whichever balanced corpus the lexicon is drawn from. We therefore reasoned that if we want to compare a large lexicon $A$ to a smaller lexicon $B$ which has $b$ nodes, it would make sense to sample the $b$ most frequent nodes from lexicon $A$ and compare lexicon $B$ to this sample, since $B$ is itself just a small sample of frequent words from a larger vocabulary. Hence we produce for each lexicon not just one phonological network, but an ensemble of phonological networks of various sizes, extracted by progressively filtering out batches of low-frequency words. Plotting the statistics of each of these phonological networks against their sizes reveals interesting differences across languages that can not be discerned from individual pointwise comparisons.

### 3.3 Data collection and pre-processing

We collected lexicons for English, Dutch, German, Spanish, French, Portuguese, Polish, and Basque. Where possible, we obtained for each language a lexicon consisting only of lemmas, and a separate wordform lexicon which includes inflectional variants. Each lexical entry consists of a phonologically transcribed word and a frequency count. The sources and sizes of the lexicons are listed in Table 3.1. For Portuguese and Polish, the frequencies came from different sources than the transcriptions, so for these languages we used the intersection of a phonemically transcribed lexicon and a lexicon
with frequencies. Similarly, the lexical database we used for Basque provided both frequency counts and pronunciations for wordforms, but only frequencies for lemmas—so the phonological network of Basque lemmas contains only those lemmas which also appear in the lexicon of Basque wordforms. We were unable to obtain phonemic transcriptions for Portuguese wordforms, or reliable frequencies for Spanish lemmas, so analyses could not be conducted for either of these lexicons. The nodes in phonological networks correspond to phonological wordforms, not orthographic forms; so if two words which are spelled differently have the same pronunciation, they constitute a single node in the network. Thus, when thresholding nodes by their associated frequencies in order to produce subgraphs of fixed sizes, we need to use frequency counts of phonological forms. The frequencies provided in the lexical databases were for orthographic forms, so in order to obtain frequencies for phonological forms it was necessary to sum the frequencies for all orthographic forms which had the same phonemic transcription.

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Language</th>
<th>Type</th>
<th>Source of pronunciations</th>
<th>Source of frequencies</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Wordforms</td>
<td>CELEX (Baayen et al., 1995)</td>
<td>CELEX</td>
<td>87,263</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lemmas</td>
<td>CELEX</td>
<td>CELEX</td>
<td>44,841</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>Wordforms</td>
<td>CELEX</td>
<td>CELEX</td>
<td>300,090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lemmas</td>
<td>CELEX</td>
<td>CELEX</td>
<td>117,048</td>
</tr>
<tr>
<td></td>
<td>German</td>
<td>Wordforms</td>
<td>CELEX</td>
<td>CELEX</td>
<td>353,679</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lemmas</td>
<td>CELEX</td>
<td>CELEX</td>
<td>50,481</td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>Wordforms</td>
<td>Lexique (New et al., 2001)</td>
<td>Lexique</td>
<td>71,334</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lemmas</td>
<td>Lexique</td>
<td>Lexique</td>
<td>43,361</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>Wordforms</td>
<td>CALLHOME (Garrett et al., 1996)</td>
<td>CALLHOME</td>
<td>42,461</td>
</tr>
<tr>
<td></td>
<td>Portuguese</td>
<td>Lemmas</td>
<td>Porlex (Gomes and Castro, 2003)</td>
<td>CORLEX (do Nascimento, 2003)</td>
<td>18,656</td>
</tr>
<tr>
<td></td>
<td>Polish</td>
<td>Wordforms</td>
<td>GlobalPhone (Schultz, 2002)</td>
<td>SUBTLEX-PL (Mandera et al., 2014)</td>
<td>25,623</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lemmas</td>
<td>GlobalPhone</td>
<td>SUBTLEX-P L</td>
<td>6024</td>
</tr>
<tr>
<td></td>
<td>Basque</td>
<td>Wordforms</td>
<td>E-hitz (Perea et al., 2006)</td>
<td>E-hitz</td>
<td>99,491</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lemmas</td>
<td>E-hitz</td>
<td>E-hitz</td>
<td>9102</td>
</tr>
</tbody>
</table>

Table 3.1: Sources and sizes of lexicons. Sizes refer to the number of distinct phonological forms.
3.3.1 Construction of phonological networks

Phonological networks were constructed using the one phoneme edit-distance metric, according to which undirected, unweighted edges are placed between pairs of words that have a Levenshtein edit distance of one. Networks were constructed using the NetworkX python package (Hagberg et al., 2008). The naïve way to construct a phonological network would be to compute the edit distance between each pair of phonological wordforms in the lexicon, adding edges between pairs for which the edit distance is one. Since words whose lengths differ by more than one phoneme cannot possibly have a Levenshtein edit distance of one, it is in fact only necessary to compare words of length \( \ell \) with words of length \( \ell \) or \( \ell + 1 \); but this still typically requires millions of edit-distance computations, even for lexicons containing only a few thousand words. Using this procedure, the time taken to construct a phonological network from the lexicon of 353,679 German wordforms was more than 30 hours. In Algorithm 1 we provide pseudocode for a much more efficient method, wherein edit distances are not explicitly computed but we instead create an inverted index of wordforms, keyed by templates into which they can be transformed by inserting or substituting a single ‘wildcard’ phoneme. Using Algorithm 1, the time taken to construct a phonological network from the German wordforms lexicon was dramatically reduced to less than 30 seconds.

3.3.2 Extraction of subgraphs

We extracted subgraphs of various sizes from each phonological network by removing all words with frequency counts below a given threshold, for various threshold values. We first extracted a subgraph from the phonological network of English lemmas by removing all words with a frequency of zero (that is, words which were listed in the CELEX lexicon of English lemmas but did not occur in the corpus from which the CELEX frequency counts were derived). We then extracted a second subgraph by removing all words with a frequency of one or below, and proceeded to extract eighteen more subgraphs, increasing the frequency threshold each time. The frequency thresholds used and resulting subgraph sizes are listed in Table 3.2. This procedure was repeated for each of the other lexicons, except that the frequency thresholds were adjusted for each lexicon such that they resulted in subgraphs of similar sizes to the English lemma subgraphs. The frequency thresholds used for each lexicon are listed in Appendix A.
Algorithm 1 Construct a phonological network

**Input:** a lexicon $L$

**Output:** a phonological network $G$

1: $G \leftarrow$ empty graph
2: $D \leftarrow$ empty dictionary \hspace{1em} $\triangleright$ keys will be templates; values will be lists of wordforms
3: $c \leftarrow$ some ‘wildcard’ character that does not occur in $L$

First, we populate $D$ with lists of wordforms that can be transformed into the same template by inserting or substituting a single instance of the wildcard character:

4: for all phonological wordforms $w \in L$ do
5: \hspace{1em} add $w$ as a node in $G$
6: \hspace{1em} $n \leftarrow$ number of phonemes in $w$
7: \hspace{1em} for $i = 0$ to $n$ do
8: \hspace{2em} $template \leftarrow w$ \hspace{1em} $\triangleright$ initialise a template as a copy of $w$
9: \hspace{2em} insert $c$ at index $i$ of $template$ \hspace{1em} $\triangleright$ insert $c$ at the specified index
10: \hspace{2em} append $w$ to $D[template]$ \hspace{1em} $\triangleright$ append $w$ to a list in $D$ keyed by the template
11: \hspace{2em} $template \leftarrow w$
12: \hspace{2em} replace the phoneme at index $i$ of $template$ with $c$
13: \hspace{2em} append $w$ to $D[template]$
14: \hspace{1em} end for
15: $template \leftarrow w$
16: append $c$ to the end of $template$
17: append $w$ to $D[template]$
18: end for

Now we simply connect up all wordforms that appear in the same list:

19: for all keys $k \in D$ do
20: \hspace{1em} for all unordered pairs of wordforms $(w, v)$ in $D[k]$ do
21: \hspace{2em} add the edge $(w, v)$ to $G$
22: \hspace{1em} end for
23: end for
24: return $G$
<table>
<thead>
<tr>
<th>Frequency threshold</th>
<th>Subgraph size</th>
<th>Frequency threshold</th>
<th>Subgraph size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>32,964</td>
<td>40</td>
<td>12,110</td>
</tr>
<tr>
<td>1</td>
<td>30,181</td>
<td>50</td>
<td>10,864</td>
</tr>
<tr>
<td>2</td>
<td>28,254</td>
<td>70</td>
<td>9190</td>
</tr>
<tr>
<td>3</td>
<td>26,788</td>
<td>100</td>
<td>7562</td>
</tr>
<tr>
<td>4</td>
<td>25,545</td>
<td>200</td>
<td>5128</td>
</tr>
<tr>
<td>5</td>
<td>24,445</td>
<td>300</td>
<td>4001</td>
</tr>
<tr>
<td>7</td>
<td>22,702</td>
<td>400</td>
<td>3315</td>
</tr>
<tr>
<td>10</td>
<td>20,671</td>
<td>500</td>
<td>2839</td>
</tr>
<tr>
<td>20</td>
<td>16,428</td>
<td>700</td>
<td>2228</td>
</tr>
<tr>
<td>30</td>
<td>13,886</td>
<td>1000</td>
<td>1694</td>
</tr>
</tbody>
</table>

Table 3.2: Frequency thresholds applied to the English lemmas lexicon, and resulting subgraph sizes.

### 3.4 Empirical verification of the effects of word length distribution and phoneme inventory size

As we discussed in Section §2.3.2, analytical analyses of the space of hypothetically possible words suggest that both the distribution of word lengths in a lexicon and the size of its phoneme inventory have an impact on the topological properties we can expect its phonological network to exhibit. We will now demonstrate this empirically by constructing random pseudolexicons with various word length distributions and phoneme inventory sizes, and comparing their statistics. In order to tease apart the effects of the word length distribution and the phoneme inventory size, we will first compare the statistics of a set of pseudolexicons which all have the same phoneme inventory size, but whose word length distributions are individually matched to each of the real lexicons (and frequency-thresholded sublexicons) in our dataset. We will then compare the statistics of a second set of pseudolexicons which all have the same word-length distribution, but whose phoneme inventory sizes are individually matched to those of the real lexicons. Based on the reasoning in Section §2.3.2, we would expect a random phonological network whose word-length-distribution is biased towards short words to have more edges, a larger giant component, and a higher average clustering coefficient than a random phonological network with the same number of nodes.
and the same phoneme inventory size, but a word-length distribution more biased towards long words. We would also expect a random phonological network with a small phoneme inventory to have more edges, a larger giant component, and a higher average clustering coefficient than a random phonological network with the same number of nodes and the same word-length distribution, but a larger phoneme inventory. Since the number of possible words of a given length scales exponentially with the length and linearly with the phoneme inventory size, we would expect the effect of the word-length distribution on the statistics of a random network to be greater than that of the phoneme inventory size.

### 3.4.1 Comparison of pseudolexicons with fixed phoneme inventory size

#### 3.4.1.1 Method

We generated a set of pseudolexicons which all had a common phoneme inventory size, and whose word-length distributions were each matched to a different lexicon (or frequency-thresholded sublexicon) from our cross-linguistic dataset. The common phoneme inventory size was set to 41, as this is the mean of the phoneme inventory sizes of the languages in our dataset. The phoneme inventory size of each language is given in Table 3.4.1.1. Pseudocode for the procedure used to generate the pseudolexicons is given in Algorithm 2. The procedure was repeated twenty times for each real (sub)lexicon, using a different random seed each time. Phonological networks were constructed from the resulting pseudolexicons, and the average clustering coefficient, number of edges, and fraction of nodes in the giant component of each resulting network were computed using the NetworkX python package. These statistics were averaged over each of the twenty random pseudolexicons matched to the same real (sub)lexicon, and plotted against the number of nodes in the network. Since they all have the same phoneme inventory size, any differences in the network statistics of pseudolexicons with the same number of nodes are purely due to their different word length distributions.
### Table 3.3: Phoneme inventory size of each language

<table>
<thead>
<tr>
<th>Language</th>
<th>Phoneme inventory size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basque</td>
<td>28</td>
</tr>
<tr>
<td>Spanish</td>
<td>29</td>
</tr>
<tr>
<td>Polish</td>
<td>36</td>
</tr>
<tr>
<td>French</td>
<td>37</td>
</tr>
<tr>
<td>Portuguese</td>
<td>39</td>
</tr>
<tr>
<td>Dutch</td>
<td>44</td>
</tr>
<tr>
<td>English</td>
<td>54</td>
</tr>
<tr>
<td>German</td>
<td>59</td>
</tr>
</tbody>
</table>

**Algorithm 2** Construct a pseudolexicon with a specified phoneme inventory size and the same word-length distribution as a specified lexicon

**Input:** a lexicon $L$, and a phoneme inventory size $I$

**Output:** a pseudolexicon $PL$

```plaintext
1: $PL \leftarrow$ empty list
2: for all phonological wordforms $w \in L$ do
3:   $pseudo_w \leftarrow$ empty string
4:   while $pseudo_w$ is the empty string or $pseudo_w \in PL$ do
5:     $partialpseudo_w \leftarrow$ empty string
6:     while length of $partialpseudo_w <$ length of $w$ do
7:       $p \leftarrow$ random integer in the range $(0, I)$
8:       append $p$ to $partialpseudo_w$
9:     end while
10:   $pseudo_w \leftarrow partialpseudo_w$
11: end while
12: append $pseudo_w$ to $PL$
13: end for
14: return $PL$
```
3.4.1.2 Results

The word length distributions for the ~6000 most frequent words\(^1\) from each of the real lexicons are shown in Figure 3.1. Compared to the other lemma lexicons, the French and English lemma lexicons contain considerably more words with 2 to 5 phonemes, and considerably fewer words with 7 phonemes or more. The Dutch lemma lexicon has the next most short words, and the next fewest long words. Similarly, compared to the other wordform lexicons, the French and English wordform lexicons contain considerably more words with 2, 3, or 4 phonemes, and considerably fewer words with 7 phonemes or more. Again, the Dutch wordform lexicon has the next most short words, and the next fewest long words. We would therefore expect pseudolexicon networks matched to the French and English word length distributions to be the most densely connected, followed by Dutch — and this is indeed what we see in the network statistics. As Figure 3.2 shows, over a range of lexicon sizes, the pseudolexicons with the most edges, the greatest fraction of nodes in the giant component, and the highest average clustering coefficient are those matched to the English and French word length distributions, followed by those matched to the Dutch word length distributions.

3.4.2 Comparison of pseudolexicons with fixed word length distribution

3.4.2.1 Method

We also investigated the extent to which the size of the phoneme inventory influences the network statistics. A second set of pseudolexicons was generated, this time using common word-length distributions and individually matching the phoneme inventory sizes to the real lexicons. The common word length distribution was arbitrarily chosen to be that of the Portuguese lemmas lexicon. Pseudocode for the procedure used to generate this second set of pseudolexicons is given in Algorithm 3. Pseudolexicons with various numbers of nodes were produced by matching the word-length distributions to each of the frequency-thresholded sublexicons of the real Portuguese lemmas lexicon. For each combination of lexicon size and phoneme inventory size, the procedure was repeated twenty times using different random seeds. Phonological networks were constructed from the resulting pseudolexicons, and the average clustering coefficient, number of edges, and fraction of nodes in the giant component of each resulting

\(^1\)The full Polish lemma lexicon contains only 6024 words, so this is the largest lexicon size at which we can compare all of the lexicons.
Figure 3.1: Distributions of the lengths of the ∼6000 most frequent words in each lexicon.
Chapter 3. Comparing phonological networks across languages

Figure 3.2: Network statistics of pseudolexicons with a fixed phoneme inventory size and lexicon-specific word-length distributions. Each data point is averaged across twenty pseudolexicons generated with different random seeds.
network were computed using the NetworkX python package. These statistics were averaged over each of the twenty random pseudolexicons with the same number of nodes and phoneme inventory size, and plotted against the number of nodes. Since they all have the same word length distribution, any differences in the network statistics of same-sized pseudolexicons from this second set are purely due to their different phoneme inventory sizes.

Algorithm 3 Construct a pseudolexicon with a specified word-length distribution and the same phoneme inventory size as a specified lexicon

**Input:** a phoneme inventory size \( I \), and a frequency distribution of word-lengths \( W \)

**Output:** a pseudolexicon \( PL \)

1: \( PL \leftarrow \) empty list
2: for all word lengths \( \ell \in W \) do
3: \( n \leftarrow W[\ell] \)
4: for \( n \) iterations do
5: \( \text{pseudo}_w \leftarrow \) empty string
6: while \( \text{pseudo}_w \) is the empty string or \( \text{pseudo}_w \in PL \) do
7: \( \text{partial}\text{pseudo}_w \leftarrow \) empty string
8: while length of \( \text{partial}\text{pseudo}_w < \ell \) do
9: \( p \leftarrow \) random integer in the range \((0,I)\)
10: append \( p \) to \( \text{partial}\text{pseudo}_w \)
11: end while
12: \( \text{pseudo}_w \leftarrow \text{partial}\text{pseudo}_w \)
13: end while
14: append \( \text{pseudo}_w \) to \( PL \)
15: end for
16: end for
17: return \( PL \)

### 3.4.2.2 Results

As Figure 3.3 shows, the phoneme inventory size does influence the network statistics, with smaller phoneme inventories resulting in more edges, a greater fraction of nodes in the giant component, and a higher average clustering coefficient. However, the effect sizes are an order of magnitude smaller than those of the word-length distribution.
Chapter 3. Comparing phonological networks across languages

Figure 3.3: Network statistics of pseudolexicons with a fixed word-length distribution and lexicon-specific phoneme inventory sizes. Each data point is averaged across twenty pseudolexicons generated with different random seeds.
3.4.3 Discussion

The plots in Figures 3.2 and 3.3 are an empirical validation of the theoretical argument that the word length distribution and phoneme inventory size of a random lexicon influence its phonological network statistics. Figure 3.2 indicates that random lexicons whose word length distributions are heavily biased towards short words tend to be denser than random lexicons which have the same number of nodes and phoneme inventory size, but contain relatively more long words. Figure 3.3 indicates that random lexicons with small phoneme inventories tend to be denser than random lexicons which have the same number of nodes and word-length distribution, but larger phoneme inventory sizes. As expected, the effect of the word length distribution is much greater than the effect of the phoneme inventory size. Based on these results, we can predict that real English or French lexicons will have higher average clustering coefficients, higher edge counts, and a greater fraction of nodes in the giant component than equivalently sized Basque or Spanish lexicons, simply by virtue of their word length distributions. We can also make predictions about how the size of the lexicon, determined by the frequency threshold, might affect the network statistics. If the words in real lexicons were sampled uniformly at random from the space of possible words, we would expect the average clustering coefficient and the fraction of nodes in the giant component to decrease as the frequency threshold is increased, just as they do in Figures 3.2 and 3.3. What would really hint at deeper organisation within human language would be cases where the network statistics of the real lexicons deviate from these expectations. One prediction we can make about how the statistics of real lexicons are likely to deviate from those of their corresponding random lexicons is that lexicons which include inflectional variants will be more dense relative to their corresponding pseudolexicons than lexicons which are restricted to lemmas will. This is because inflectional variants are likely to be phonological neighbours of the lemmas from which they are derived, as well as to other variants of the same lemma, and to equivalent variants of their own lemmas’ neighbouring lemmas—consider for example the Spanish lemmas *comer* (/komeR/) and *correr* (/koreR/), and their inflectional variants *como* (/komo/), *comes* (/komes/), *come* (/kome/), *corro* (/koro/), *corres* (/kores/), *corre* (/kore/), and so on. Taking this line of reasoning further, we can also predict that this difference will be more pronounced for languages which have more inflectional variants per lemma, such as languages with rich nominative case systems.
3.5 Robust comparison of phonological networks across eight languages

Having demonstrated empirically that the topology of a phonological network is influenced both by its word-length distribution and by its phoneme inventory size, we generated a third set of pseudolexicons, which matched the sizes, phoneme inventory sizes, and word-length distributions of the corresponding real lexicons. For each of the resulting real and random phonological networks, we computed various network statistics and plotted these against the number of nodes in the network. The resulting plots enable us to compare network statistics across different languages, different lexicon types, and different lexicon sizes—and importantly, they enable us to compare the differences in values between each of these real lexicons and their corresponding size-, phoneme-inventory-size-, and word-length-distribution-matched pseudolexicons.

3.5.1 Comparison with Arbesman et al. (2010)

A summary of network statistics reported by Arbesman et al. (2010) for their English, Spanish, Mandarin, Hawaiian, and Basque lexicons is reproduced in Table 3.4, along with the corresponding values for our English lemma, Spanish wordform, and Basque wordform lexicons.² Since our English lemmas lexicon is much larger than the English lexicon that Arbesman et al. used, Table 3.4 also gives statistics computed on a similarly-sized subgraph, extracted using a frequency threshold of 10. Note however that our Spanish wordforms lexicon is much smaller than that used by Arbesman et al, and hence it could not be matched in size. Our Basque wordforms lexicon was derived from the same database as theirs, but contains slightly more words.³ Since our Basque lexicons are almost identical, the network statistics ought to be near-identical too—and as Table 3.4 shows, this is indeed the case. The network statistics for our full-size English lemmas lexicon, on the other hand, are not such a good match with those reported by Arbesman et al. The edge-to-node ratio and fraction of nodes in the giant component, for instance, are considerably lower than in any of the lexicons analysed by Arbesman et al. Note however, that the statistics of the similarly-sized subgraph of our English lemmas network are somewhat closer to those of the English lexicon anal-

²We do not compare the average clustering coefficients, since Arbesman et al. did not specify which definition of the average clustering coefficient they used.

³Presumably Arbesman et al. must have discarded some words from the original database, for reasons unknown.
This confirms our hypothesis that the size of the lexicon can have a substantial effect on the network statistics. Nevertheless, although our Spanish wordforms lexicon is less than half the size of the one that Arbesman et al. analysed, its statistics are quite similar. Thus, the size of a lexicon does not appear to influence the network statistics in a straightforward, predictable way across all languages. This is our main motivation for comparing plots of network statistics computed on lexicons of various sizes, as opposed to static values which are effectively single points sampled from these plots.

### 3.5.2 Average word length

Plots of the average number of phonemes per word in each lexicon are shown in Figure 3.4. French and English lexicons have the shortest average word lengths, while the Spanish lexicon has the longest. It may seem surprising that the average word length tends to be shorter when inflectional variants are included than when lexicons are restricted to lemmas, since one would expect most inflectional variants to be slightly longer than the lemmas from which they are derived, given that inflection is often marked by affixation. It makes sense, however, when we consider that frequency is inversely related to word length, and that a given lemma usually has multiple inflectional variants: in a lexicon of a given size, the inclusion of several frequent four-phoneme variants of a frequent three-phoneme lemma means there is less room for less frequent, six- or seven-phoneme lemmas.

Plots of the average word length in the giant component are shown in Figure 3.5. For all eight languages, the giant components of the real lexicons include longer words than those of their random counterparts, and the gap between the real and random lexicons widens as the lexicon size is increased. The average word length in the giant component is longer when inflectional variants are included, and it is interesting to note that the gap between lemma and wordform lexicons is widest for Polish and Basque – which have extensive grammatical case systems; narrowest for English, French, and Dutch – which have little or no case marking; and somewhere in between for German – which has a moderate amount of case marking (Iggesen, 2013). This supports our prediction that wordform lexicons would be more dense relative to their corresponding pseudolexicons than lemma lexicons would, and that this difference would be more pronounced for languages which have more inflectional variants per lemma.
### Table 3.4: Comparison of phonological network statistics reported by Arbesman et al. with those computed on our data (shown in boldface).
Adapted from (Arbesman et al., 2010, p.681).

<table>
<thead>
<tr>
<th></th>
<th>English lemmas</th>
<th>Spanish wordforms</th>
<th>Mandarin</th>
<th>Hawaiian</th>
<th>Basque wordforms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Arbesman et al. full lexicon subgraph</td>
<td>Arbesman et al. full lexicon</td>
<td>Arbesman et al.</td>
<td>Arbesman et al.</td>
<td>Arbesman et al. full lexicon</td>
</tr>
<tr>
<td>Number of nodes</td>
<td>19,323</td>
<td>44,912</td>
<td>122,066</td>
<td>42,461</td>
<td>30,086</td>
</tr>
<tr>
<td>Fraction of nodes in GC</td>
<td>0.34</td>
<td>0.24</td>
<td>0.37</td>
<td>0.35</td>
<td>0.66</td>
</tr>
<tr>
<td>Assortativity coefficient</td>
<td>0.657</td>
<td>0.728</td>
<td>0.762</td>
<td>0.740</td>
<td>0.654</td>
</tr>
<tr>
<td>ASPL (GC)</td>
<td>6.1</td>
<td>7.5</td>
<td>6.8</td>
<td>10.6</td>
<td>10.1</td>
</tr>
<tr>
<td>Transitivity</td>
<td>0.313</td>
<td>0.312</td>
<td>0.323</td>
<td>0.225</td>
<td>0.404</td>
</tr>
<tr>
<td>Edge-to-node ratio</td>
<td>1.61</td>
<td>1.05</td>
<td>1.47</td>
<td>1.25</td>
<td>2.57</td>
</tr>
<tr>
<td>Edge-to-node ratio (GC)</td>
<td>4.55</td>
<td>3.89</td>
<td>4.14</td>
<td>2.43</td>
<td>3.88</td>
</tr>
</tbody>
</table>
Figure 3.4: **Average word length** Solid lines show the average word length as a function of lexicon size for each lemma lexicon, and dashed lines show the average word length as a function of lexicon size for each wordform lexicon. By design, pseudolexicons have identical word-length distributions, and hence identical average word lengths, to the relevant real lexicons.
Fig. 3.5: **Average word length in the giant component** The thick, coloured lines show the average word length in the giant component as a function of lexicon size, for each real lexicon. The thin, black lines show the average word length in the giant component averaged across twenty random pseudolexicons, and the shaded regions indicate the range of values across the twenty pseudolexicons. Solid lines are for lemmas; dashed lines are for all wordforms.
3.5.3 Number of edges

From the log-log plots in Figure 3.6 we can see that the number of edges tends to be approximately proportional to the number of nodes, in both real and random phonological networks. However, across all eight languages, the constant of proportionality is considerably greater in the real networks than in their random counterparts. This reflects the fact that phonotactic constraints and morphological paradigms are at play in real lexicons, meaning words are more similar to one another than if they were really chosen at random from the space of all possible words. The constant of proportionality in the real networks tends to be slightly greater when inflectional variants are included than when the lexicon is restricted to lemmas – but this is generally not the case for the pseudolexicon networks. Furthermore, the gap between lemma and wordform lexicons is once again widest for the languages which have the richest nominative case systems, lending further support to our predictions.

3.5.4 Fraction of nodes in the giant component

As Figure 3.7 shows, the fraction of nodes in the giant component is larger in each of the real lexicons than would be expected if the words had been sampled at random. However, the difference in values between the real lexicons and their corresponding pseudolexicons is considerably greater for some languages, such as French and English, than for others, most notably Polish. Interestingly, while for English, French, and Dutch the fraction of nodes in the giant component tends to decrease as more infrequent words are included, this relationship is reversed for Spanish. This does not appear to be solely a consequence of differences in word-length distributions, since in the random pseudolexicon which matches the Spanish word-length distribution, the fraction of nodes in the giant component does tend to decrease with increasing lexicon size. These results illustrate how the size of the lexicon can make a considerable difference to the qualitative characteristics of a phonological network: at a lexicon size of around 33,000 words, the giant component of the Spanish wordforms network contains 33% of the words, while the giant component of the Dutch wordforms network contains only 27% of the words; yet at a lexicon size of around 8000 words, the giant component of the Spanish wordforms network contains just 23% of the words, while the giant component of the Dutch wordforms network contains as many as 37% of the words. Hence if we only compared the 33,000-word lexicons we would conclude that Spanish wordforms were more densely connected than Dutch wordforms; whereas if
Figure 3.6: **Number of edges** The thick, coloured lines show the number of edges as a function of lexicon size, for each real lexicon. The thin, black lines show the number of edges averaged across twenty random pseudolexicons, and the shaded regions indicate the range of values across the twenty pseudolexicons. Solid lines are for lemmas; dashed lines are for all wordforms.
we only compared the 8000-word lexicons we would come to the opposite conclusion. Furthermore, the inclusion or exclusion of inflectional variants also makes a considerable difference to the fraction of nodes in the giant component, and the magnitude of this difference varies across languages, which again supports our predictions from Section §3.4.3.

3.5.5 Fraction of edges in the giant component

Figure 3.8 shows the fraction of edges in the giant component. While the shapes of the pseudolexicon curves vary considerably, in most of the real lexicons the fraction of edges in the giant component is relatively stable across different lexicon sizes. A notable exception is Polish, for which the trajectories in the real lexicons are similar to those in the pseudolexicons, but the values are consistently lower. Hence while English, Dutch, and German lexicons have a greater fraction of nodes in their giant components than would be expected given their word length distributions, Polish lexicons have a smaller fraction than expected. In Spanish and Portuguese lexicons, the fraction of edges in the giant component is lower than expected when only relatively frequent words are included in the lexicon, but greater than expected when many infrequent words are included.

3.5.6 Fraction of nodes which are isolates

Figure 3.9 shows the fraction of nodes which are isolates. Notice that although the gradients differ, in all pseudolexicons the fraction of nodes which are isolates increases as the size of the lexicon is increased. For real lexicons, however, the inclusion or exclusion of inflectional variants makes a considerable difference to the shape of the curve, which cannot be attributed to differences in the word-length distributions since we don’t see these differences in the pseudolexicons. In wordform lexicons, the fraction of nodes which are isolates generally tends to decrease as more infrequent words are included. In lemma lexicons, however, there are differences across languages: in the English and Dutch lemma lexicons, and in the French and German lemma lexicons to a lesser extent, the fraction of nodes which are isolates tends to increase as more infrequent words are included; whereas it tends to decrease in the Portuguese lemma lexicon, and remains relatively stable in the Basque lexicon.

Taking Figures 3.7, 3.8, and 3.9 together, we see that in the French, English, and Dutch wordform lexicons, around 90% of the edges and 30-50% of the nodes make up
Figure 3.7: **Fraction of nodes in the giant component.** The thick, coloured lines show the fraction of nodes in the giant component as a function of lexicon size, for each real lexicon. The thin, black lines show the fraction of nodes in the giant component averaged across twenty random pseudolexicons, and the shaded regions indicate the range of values across the twenty pseudolexicons. Solid lines are for lemmas; dashed lines are for all wordforms.
Figure 3.8: **Fraction of edges in the giant component.** The thick, coloured lines show the fraction of edges in the giant component as a function of lexicon size, for each real lexicon. The thin, black lines show the fraction of edges in the giant component averaged across twenty random pseudolexicons, and the shaded regions indicate the range of values across the twenty pseudolexicons. Solid lines are for lemmas; dashed lines are for all wordforms.
the giant component, and the remaining 10% of the edges are distributed among the 20-30% of nodes that are neither isolates nor part of the giant component. In German and Basque wordform lexicons, around 70-80% of edges and 30% of nodes make up the giant component, and the remaining 20-30% of edges are distributed among 25-40% of the nodes. In the Spanish wordforms lexicon, 60-70% of edges and 20-40% of nodes make up the giant component, and the remaining 40-30% of edges are distributed among 30-40% of the nodes. Finally, in the Polish wordforms lexicon, 40-60% of edges and 15-20% of nodes make up the giant component, and the remaining 40-60% of edges are distributed among 40-50% of the nodes. In other words, French, English and Dutch wordform lexicons have relatively large, dense giant components, but nodes outside of their giant components are very sparsely connected. German and Basque wordform lexicons have slightly less dense giant components, and slightly more edges among the nodes outwith them. The Spanish wordforms lexicon has a small giant component and many isolates when it is restricted to relatively frequent words, but as more infrequent words are added, the proportion of nodes in the giant component grows, and the proportion of nodes which are isolates decreases. Compared to English, French, and Dutch, the Spanish wordform lexicon consistently has a greater density of edges among nodes outside of the giant component. The Polish wordforms lexicon has a small but dense giant component, and a high density of edges outside of the giant component.

3.5.7 Average shortest path length in the giant component

Although we can see from Figure 3.7 that the differences in giant component sizes between real and random lexicons are considerably greater for English and French than for Polish, Figure 3.10 shows that the average shortest path lengths in the giant components of real French and English lexicons are much more similar to their random counterparts than is the case for Polish. That is to say, even though the English and French giant components are much larger than we would expect from random phonological networks with the same word-length distributions, their average shortest path lengths are very similar to those of the random lexicons’ smaller giant components. The Polish giant components on the other hand, are relatively similar in size to those of their corresponding random lexicons, but have significantly longer average shortest path lengths. This suggests that the Polish lexicon is less of a ‘small-world’ than the English or French lexicons.
Figure 3.9: **Fraction of nodes which are isolates.** The thick, coloured lines show the fraction of nodes which are isolates as a function of lexicon size, for each real lexicon. The thin, black lines show the fraction of nodes which are isolates averaged across twenty random pseudolexicons, and the shaded regions indicate the range of values across the twenty pseudolexicons. Solid lines are for lemmas; dashed lines are for all wordforms.
Figure 3.10: **Average shortest path length in the giant component.** The thick, coloured lines show the average shortest path length in the giant component as a function of lexicon size, for each real lexicon. The thin, black lines show the average shortest path length in the giant component averaged across twenty random pseudolexicons, and the shaded regions indicate the range of values across the twenty pseudolexicons. Solid lines are for lemmas; dashed lines are for all wordforms.
3.5.8 Average degree in the giant component

The average degree in the giant component, shown in Figure 3.11, is higher for most of the real lexicons than for their random counterparts. However, for Portuguese and Polish it is similar or lower in the real lexicons compared with the pseudolexicons. For Basque, the values for the real and pseudolexicons converge as the frequency threshold is increased, whereas for English and Dutch they diverge.

3.5.9 Assortative mixing by degree

The coefficient of assortative mixing by degree, shown in Figure 3.12, tends to be slightly higher in the real lexicons than in their random counterparts, but only for Spanish is this consistently the case across all lexicon sizes. It most notably dips below the pseudolexicon average for Portuguese lemmas and Polish wordforms, which potentially could explain the lower-than-expected average degrees in the giant components of these lexicons: in a network with high assortative mixing by degree, the high degree nodes will cluster together, forming a densely connected giant component; whereas in a network with lower assortative mixing by degree, some of the high degree nodes might form separate, smaller islands, and hence the giant component will not have such a dense core.

3.5.10 Transitivity

The transitivity, shown in Figure 3.13, tends to be slightly lower in the real lexicons than in their random counterparts. As Stella and Brede (2015) have suggested, this points to the possibility of a constraint on word confusability - i.e. a pressure to avoid words having too many phonological neighbours which are distinguished by the same segment.

3.6 Conclusion

We have developed a new methodology for comparing phonological networks across languages, and have shown how the resulting plots can facilitate more nuanced analyses than can be achieved by simply comparing static numerical values. We have demonstrated empirically that both the distribution of word lengths and the size of the phoneme inventory influence the fraction of nodes in the giant component, the num-
Figure 3.11: Average degree in the giant component. The thick, coloured lines show the average degree in the giant component as a function of lexicon size, for each real lexicon. The thin, black lines show the average degree in the giant component averaged across twenty random pseudolexicons, and the shaded regions indicate the range of values across the twenty pseudolexicons. Solid lines are for lemmas; dashed lines are for all wordforms.
Figure 3.12: **Assortative mixing by degree** The thick, coloured lines show the coefficient of assortative mixing by degree as a function of lexicon size, for each real lexicon. The thin, black lines show the coefficient of assortative mixing by degree averaged across twenty random pseudolexicons, and the shaded regions indicate the range of values across the twenty pseudolexicons. Solid lines are for lemmas; dashed lines are for all wordforms.
Figure 3.13: **Transitivity** The thick, coloured lines show the transitivity as a function of lexicon size, for each real lexicon. The thin, black lines show the transitivity averaged across twenty random pseudolexicons, and the shaded regions indicate the range of values across the twenty pseudolexicons. Solid lines are for lemmas; dashed lines are for all wordforms.
ber of edges, and the average clustering coefficient in phonological networks whose nodes are sampled uniformly at random from the space of possible words. We predicted that real lexicons which include inflectional variants would be more densely connected relative to random lexicons with the same phoneme inventory size and word length distribution than would equivalently sized lexicons with the inflectional variants removed— and that these differences would be more pronounced for languages with richer inflectional paradigms. These predictions were borne out in the results of our comparative analysis; and other, more intriguing differences across languages were also revealed. For example, while the fraction of nodes in the giant component decreases as more and more infrequent words are added to the Dutch wordforms lexicon, it increases as more and more infrequent words are added to the Spanish wordforms lexicon. This indicates that the Dutch lexicon has a core of densely interconnected frequent words and a periphery of more sparsely connected, less frequent words; while in the Spanish lexicon, frequent and infrequent words are more evenly distributed. Additionally, English, French, and Dutch wordform lexicons are composed of large, dense, giant components and a periphery of very sparsely connected nodes, whilst Spanish and Polish wordform lexicons have relatively similar edge counts inside and outside of the giant component. A potential implication of these differences is that speakers of French, English, and Dutch might use different lexical processing mechanisms than speakers of Spanish and Polish.

Our new methodology is not a hard-and-fast solution to the problem of comparing lexicons of different sizes: since we only have 6024 Polish lemmas in total, it is still difficult to compare the Polish lemma lexicon with other lexicons for which we have much more data. For instance, the fraction of nodes in the giant component of the Polish lemmas lexicon just begins to increase slightly at the right-hand edge of the plot (Figure §3.7), but we’re none the wiser as to whether it would continue on that trajectory if we added more words, and end up looking similar to the Spanish wordforms plot; or whether the fraction of nodes in the giant component might remain relatively stable or even begin to decrease if we continued adding more infrequent words. Nevertheless, comparisons between series of phonological network statistics are clearly much more informative than comparisons between singular data points, which have been the status quo up until now.

A fruitful area for further work would be to apply this novel methodological approach to a more diverse set of languages. It seems likely that the phonological network statistics of languages whose morphology is both highly synthetic and highly fusional
would be more densely connected, relative to random pseudolexicons, than would the statistics of languages whose morphology is relatively analytic and isolating— but the languages in our dataset, of which all but one are from the Indo-European family, do not differ greatly along these dimensions. Another area for further work would be to try to determine the expected values of statistics of random phonological networks analytically, given a particular phoneme inventory size and word-length distribution, to avoid the need to explicitly construct the pseudolexicons.
Chapter 4

Comparing phonological networks across age groups

4.1 Background

A central question in first language acquisition research concerns the nature of children’s lexical representations. According to the lexical restructuring hypothesis (Met-sala, 1997), children’s early lexical representations are relatively coarse-grained or ‘holistic’, but as their vocabularies grow and become more densely packed with similar-sounding words, the need to discriminate between these similar-sounding words pressures children to re-analyse their lexical representations in terms of more fine-grained, segmental units. This hypothesis depends on the premise that children’s lexicons do in fact become more densely packed as they acquire more words. Studies investigating the density of children’s developing lexicons have produced conflicting results, with some appearing to provide evidence in support of the lexical restructuring hypothesis, and others appearing to counter it.

Charles-Luce and Luce (1990) compared expressive lexicons of 5-year-old and 7-year-old children with an adult lexicon derived from a dictionary. They computed the phonological neighbourhood densities \(^1\) of each of the 3, 4, and 5 phoneme words in the children’s lexicons, and then computed the phonological neighbourhood densities of these same words with respect to the adult lexicon. They then compared the distributions of phonological neighbourhood densities across the three lexicons, separately for each of the three word lengths. They found that words tended to have more phono-

\(^1\) i.e. the number of words having a Levenstein-edit distance of 1 from the target word – this is equivalent to a word’s degree, if we think of the lexicon as a phonological network.
logical neighbours in the adult lexicons than in the children’s lexicon, and furthermore that they tended to have more neighbours in the 7-year-old lexicon than in the 5-year-old lexicon. These findings were consistent for three, four, and five phoneme words. They therefore concluded that children’s lexicons become more densely packed as they grow.

Following up on this study, Coady and Aslin (2003) compared expressive and receptive lexicons of children aged between 27 and 42 months with an adult lexicon. They first used the same procedure as Charles-Luce and Luce (1990), and obtained consistent results: words had fewer phonological neighbours in the children’s lexicons than in the adult lexicon. However, they then calculated the ratios of the neighbourhood sizes to the vocabulary size, in order to control for the fact that the adult lexicon is much larger than the children’s lexicons. When phonological neighbourhood density was expressed as a proportion of the lexicon, the pattern of results reversed, and the children’s lexicons appeared denser than those of adults.

Inspired by the phonological network analyses of (Vitevitch, 2008) and (Arbesman et al., 2010), Carlson et al. (2011) used a variety of metrics to compare the global phonological network structures of lexicons drawn from corpora of child speech (from children aged between 14 and 46 months), child-directed speech (from the caregivers of those children), and adult-directed speech (from a different sample of adults). Their results for the distribution of neighbourhood sizes were consistent with those of Coady and Aslin: when normalised for the size of the lexicon, the proportion of words with large neighbourhoods was higher in the child speech and child-directed speech lexicons than in the adult-directed speech lexicon. They also found that compared to adult-directed speech, the phonological networks of child speech and child-directed speech had higher edge-to-node-ratios, higher average clustering coefficients and transitivity, higher assortative mixing by degree, and a greater proportion of nodes in the largest connected component. Furthermore, the child speech lexicon had a shorter average shortest path length than the child-directed speech and adult-directed speech lexicons. These findings all indicate that children’s lexicons are denser than adults’ lexicons; not sparser. Carlson et al. suggest that this denser structure may arise from biases in the early stages of acquisition which favour lexicons that are stable and efficiently searchable, since the observed ‘small-world’ properties and high assortative mixing by degree have been associated with rapid transmission of information (Watts and Strogatz, 1998), (Newman, 2003) and robustness in the face of random node removal (Arbesman et al., 2010). As Carlson et al. acknowledged, relating the structure
Chapter 4. Comparing phonological networks across age groups

of children’s lexicons to efficient search and stability “is dependent on a theory of children’s lexical acquisition and processing mechanisms that allow them to exploit the network properties we have discussed” (Carlson et al., 2011)[p.106].

In §4.2, we will argue that a much more parsimonious explanation for the denser structure of children’s lexicons lies in the distribution of word lengths. While Charles-Luce and Luce (1990) and Coady and Aslin (2003) compared the neighbourhood density distributions for 3, 4, and 5-phoneme words separately, (Carlson et al., 2011) only analysed networks constructed from the complete, un-segregated lexicons, arguing that they were “interested in the impact of phonological connectivity across the entire lexicon, rather than in connectivity among, e.g. 3-phoneme words”. The problem with this approach is that, as Coady and Aslin (2003) have shown, average neighbourhood size decreases as word length increases, and children’s lexicons have a smaller proportion of longer words compared to adults’ lexicons. Thus the fact that various network measures show children’s lexicons to be denser than adults’ lexicons may simply be due to the fact that children’s lexicons contain a disproportionate number of shorter words. One could of course hypothesise that children favour lexicons with a high proportion of short words precisely because this induces a denser phonological network structure which is advantageous for lexical processing— but more parsimoniously, one could instead hypothesise that short words are preferentially acquired because – by virtue of their shortness – they are easier to hold in working memory, easier to process, and easier to articulate. Also, frequent words tend to be shorter, and so early lexicons may contain many short words simply because children are exposed to these words most often.

4.2 Comparison of child, caregiver, and adult phonological networks

Carlson et al. (2011) have suggested that compared to phonological networks derived from corpora of adult-directed speech (ADS), phonological networks derived from corpora of child-directed speech (CDS) and children’s speech (CS) have topological properties more indicative of ‘small-world structure’, and that the growth process underlying child phonological networks may be biased to favour searchability and stability in the developing mental lexicon. However, since we have shown in Chapter 3 that word-length distributions which are biased towards short words result in phono-
logical networks having larger giant components, more edges, and a higher average clustering coefficient, an alternative hypothesis is that CS and CDS phonological networks exhibit these characteristics simply because children favour short words. In order to test this hypothesis, we replicate Carlson et al’s (2011) study, but in addition to directly comparing the raw lexicons extracted from the corpora, we also compare sublexicons matched in size and word-length distribution. If the differences in the network statistics are purely a consequence of the different word-length distributions, then these differences should disappear when we compare samples which all have the same word-length distribution.

4.2.1 Data

The ADS lexicon was drawn from the Buckeye corpus of adult interviews (Pitt et al., 2007). Although this is the same source corpus that Carlson et al. (2011) used, there are some discrepancies in our data: Carlson et al. (2011) report that their ADS lexicon contained 10,964 orthographic word types after filtering out word types not listed in a pronunciation dictionary, yet before any filtering we counted only 9,577 distinct orthographic word types. Carlson et al. (2011) drew their CS and CDS lexicons from a large longitudinal corpus of spontaneous speech between parents and children. As we do not have access to the specific corpus they used, we instead used CS and CDS lexicons extracted from the CHILDES database (MacWhinney, 2000). For the CDS lexicon, we used the CHILDES parental database (Li et al., 2000), which consists of word types and frequency counts from the utterances of parents, caregivers, and experimenters extracted from 27 English corpora in the CHILDES database. For the CS lexicon, we extracted all unique word types uttered by any target child aged between 14 and 46 months in any of the North American English corpora of the CHILDES database. The full list of corpora from which we extracted the CS lexicon is given in Appendix B.

Following Carlson et al. (2011), phonological transcriptions for each lexicon were obtained from the CMU pronouncing dictionary (Weide, 1998), and word types not listed in the dictionary were removed. Table 4.1 gives the number of distinct phonological word types in each lexicon, after removing orthographic types not listed in the CMU pronouncing dictionary. Also displayed in Table 4.1 are the number of distinct orthographic forms in each lexicon, and the number of distinct orthographic forms in the lexicons analysed by Carlson et al. (2011), for comparison. Our lexicon constuc-
Table 4.1: Number of distinct phonological types in each of our lexicons, number of distinct orthographic types in each of our lexicons, and number of distinct orthographic types in each of the lexicons analysed by Carlson et al. (all following removal of words not occurring in CMU dictionary).

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Phon.types in our data</th>
<th>Orth.types in our data</th>
<th>Orth.types in Carlson et al. (2011)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>10,320</td>
<td>10,667</td>
<td>6,882</td>
</tr>
<tr>
<td>CDS</td>
<td>11,631</td>
<td>12,320</td>
<td>12,584</td>
</tr>
<tr>
<td>ADS</td>
<td>8,911</td>
<td>9,187</td>
<td>10,964</td>
</tr>
</tbody>
</table>

4.2.2 Raw lexicons

4.2.2.1 Word-length distributions

The word-length distributions for each lexicon are displayed in Figure 4.1. The CS lexicon is the most biased towards shorter words, and the ADS lexicon the least.
4.2.2.2 Network statistics

Figure 4.2 shows the values of six network statistics, measured in the whole network and/or the giant component as applicable, for the CS, CDS, and ADS lexicons. Although Carlson et al. (2011) reported values for the average shortest path length in the whole network, it is not clear how these were computed or how they should be interpreted (in fact, the values they report for the average shortest path length in the whole network appear to be identical to the values for the average shortest path length in the giant component), so we do not attempt to replicate them here. Our results corroborate the finding that CS phonological networks have higher edge-to-node-ratios, higher clustering coefficients, higher transitivity, a higher percentage of nodes in the giant component, and shorter average shortest path lengths than ADS phonological networks — however, in contrast to Carlson et al. (2011)’s results, the statistics of our CDS phonological network tend to be more similar to those of ADS than CS. This may be because our CDS lexicon was drawn from the set of all utterances produced by parents, caregivers or experimenters in the original CHILDES corpora, and not all of these utterances will actually have been directed at the child – meaning our CDS lexicon may be somewhat adulterated with adult-directed speech.

4.2.3 Sublexicons with fixed word-length distribution

4.2.3.1 Extraction of sublexicons

A target word-length distribution was obtained by taking each word length that has a non-zero frequency in all three lexicons, and setting the frequency for that word length in the target word-length distribution to be the lowest frequency with which it occurs in any of the three raw lexicons. This ensures that all three lexicons contain enough words of each word length to be able to produce a sublexicon which has the target word-length distribution. Figure 4.3 shows the word-length distributions of the raw CS, CDS, and ADS lexicons, with the target word-length distribution overlaid. The target word-length distribution contains fewer short words than the raw CS lexicon, fewer long words than the raw ADS lexicon, and fewer words of every length than the CDS lexicon. If the relatively high network statistic values of the raw CS lexicon are a consequence of it containing more short words, then the network statistics for a CS sublexicon matched to this target word-length distribution ought to be lower. Likewise, if the relatively low network statistic values of the raw ADS lexicon are a consequence
Figure 4.2: Network statistics measured in the whole network (left) and the giant component only (right)
of it containing more long words, then the network statistics for an ADS sublexicon matched to the target word-length distribution ought to be higher.

![Graph showing word-length distributions in each of the raw lexicons (orange bars), with the target word-length distribution overlaid (purple bars).]

Figure 4.3: Word-length distributions in each of the raw lexicons (orange bars), with the target word-length distribution overlaid (purple bars)

The following procedure was used to extract sublexicons with the target word-length distribution from the CS, CDS, and ADS lexicons: Let $L$ be the set of all word lengths that occur in the target word-length distribution. Then for each word length $l$ such that $l \in L$, let $f_l$ be the frequency count of that word length in the target word-length distribution, and add to the sublexicon the $f_l$ most frequent words from the raw lexicon that have word-length $l$. This procedure yields sublexicons of size 7,314 words.

4.2.3.2 Network statistics

Figure 4.4 allows us to compare the values of the network statistics for the raw lexicons with the network statistics for the word-length–distribution–matched sublexicons. As predicted by our hypothesis that the denser structure of the CS lexicon is a consequence of it containing more short words, the values of the network statistics are much more similar across the three lexicons when their word-length distributions are matched (with the notable exception of the transitivity, for which the values actually become more disparate). Furthermore, as we predicted on the basis of the comparison of the target word-length distribution to the raw word-length distributions, statistics for which the raw CS lexicon has relatively high values (other than transitivity) become much lower in the CS sublexicon, while in the whole network at least, statistics for
which the raw ADS lexicon has relatively low values become higher in the ADS sublexicon. When the word-length distributions are matched, the network statistic values for CS and CDS tend to be equivalent to or even lower than those of ADS — that is to say, if the statistics do differ, they actually pattern in the opposite direction to how they pattern in the raw lexicons, so that the ADS sublexicon could be said to be the most dense.

To confirm that the statistics became more similar in the sublexicons specifically because their word-length distributions are matched, and not simply because their sizes are matched, Figure 4.5 compares the values of the network statistics for the raw lexicons with the network statistics for a different set of sublexicons, matched in size but not word-length-distribution. These sublexicons were constructed by extracting the 7,314 most frequent words from the corresponding raw lexicon. The statistics of the size-matched sublexicons do not deviate much from those of the raw lexicons, and where they do deviate, they generally all deviate in the same direction and by a similar magnitude. We can therefore conclude that most of the differences we see in the network statistics of the CS, CDS, and ADS lexicons are driven by their different word-length distributions.

4.2.4 Discussion

The statistics of the giant component in the ADS lexicon change very little when nodes are removed in order to match its word-length distribution to the common target — but notably, the percentage of nodes in the giant component increases. The giant component cannot have grown, since nodes were removed from the network rather than added to it, so this indicates that the majority of removed nodes were taken from outside the giant component. The edge-to-node-ratio in the ADS network as a whole also increases, and this indicates that the number of nodes must have decreased more than the number of edges did — meaning that some of the nodes that were removed must have been isolates. It is interesting to note that the transitivity barely changes, yet the average clustering coefficient increases. Recall from Section 2.1.2 that transitivity is defined as

\[ T = \frac{3\delta}{\tau}, \]

where \( \delta \) is the number of triangles in the network and \( \tau \) is the number of triples in the network. The fact that the transitivity remains almost exactly the same indicates that the ratio of triples to triangles in the ADS network has barely changed. How then, has
Figure 4.4: Network statistics measured in the whole network (left) and the giant component only (right) of the raw CS, CDS, and ADS lexicons (black squares) and their word-length–distribution–matched sublexicons (red crosses)
Figure 4.5: Network statistics measured in the whole network (left) and the giant component only (right) of the raw CS, CDS, and ADS lexicons (black squares) and their size-matched sublexicons (red crosses)
the average clustering coefficient increased? Recall also that we calculate the average clustering coefficient using the equation

$$\langle c \rangle = \frac{1}{|V|} \sum_{v \in V} c(v),$$

setting $c(v)$ to zero for nodes $v$ having $k(v) < 2$. This means that nodes with one or zero neighbours are included in the denominator but do not contribute to the sum. Hence decreasing the number of nodes that have one or zero neighbours increases the value of the average clustering coefficient, but doesn’t affect the transitivity since it doesn’t change the number of triangles or triples in the network. This implies that most of the nodes removed from the ADS lexicon had one or no neighbours— an implication which is supported by the fact that the degree assortativity coefficient in the network as a whole decreases, but the degree assortativity coefficient in the giant component does not. If the original network contained lots of two-node components, this would have inflated its assortativity coefficient as all of those tiny components would be perfectly assortatively mixed, with each node having exactly one neighbour who also had exactly one neighbour. Hence, removing one or both nodes from lots of these components would reduce the assortativity coefficient. Taken together then, the results indicate that the words which were removed from the ADS lexicon in order to match its word-length distribution to the common target were among the most isolated nodes in the original network. This supports our hypothesis that the relative sparsity of the ADS lexicon is a natural consequence of it containing relatively more long words.

It is interesting that all of the statistics become more similar across lexicons when the word-length-distributions are matched, except for the transitivity. Whereas it decreases slightly in the CS network and remains unchanged in the ADS lexicon, the transitivity increases markedly in the CDS network. This increase in transitivity indicates that when the word-length-distribution of the CDS network is matched to the common target, the decrease in the number of triples is considerably greater than the decrease in the number of triangles. The percentage of nodes in the giant component decreases slightly which indicates that some nodes are removed from the giant component, but the increase in transitivity indicates that they are mostly removed from the peripheries, rather than the densely clustered core. Again, this makes perfect sense in light of the fact that the many of removed words were of intermediate lengths— and so we would expect them to be better connected than the very longest words, but not as densely connected as the very shortest words.

Finally, the statistics of the word-length-distribution-matched sublexicon of the CS
lexicon also confirm our expectations. We hypothesised that the original CS lexicon
was the densest simply because it contained the highest proportion of short words,
and the results support this hypothesis. When short words are removed in order to
match the word-length-distribution to the common target, the average shortest path
length in the giant component increases, despite the fact that it consists of considerably
fewer nodes. The edge-to-node-ratio, average clustering coefficient, and transitivity
also decrease, both in the giant component and in the network as a whole. These
results indicate that nodes were removed not just from the peripheries, but also from
the dense core.

4.2.5 Conclusion

The phonological network statistics of our raw child and adult lexicons corroborate
those reported by Carlson et al. (2011): the CS network has a higher edge-to-node
ratio, a higher average clustering coefficient, higher transitivity, a greater fraction of
nodes in the giant component, and a shorter average short path length than the ADS
network. These results resoundingly indicate that child lexicons are denser than adult
lexicons, and we do not dispute that this is the case. However, in light of what we
now know about the influence of the word length distribution on phonological network
topology, we hypothesised that the denser structure of children’s productive lexicons
relative to those of adult lexicons could be parsimoniously explained by their shorter
word length distributions. To test this hypothesis, we compared the phonological net-
work statistics of samples from our CS, CDS, and ADS lexicons which all had the same
word length distribution. When the word length distribution was controlled, the phono-
logical network statistics were much more similar across the three lexicons. This was
not the case when we only controlled the number of words, indicating that the word
length distributions were indeed the driving force behind the differences. One notable
exception to this pattern was the transitivity, which actually became more different
across the three lexicons when their word length distributions were controlled. In con-
sidering why this should be the case, we amassed further support for the notion that
shorter words tend to form the densely connected core of a phonological lexicon, while
longer words attach to the periphery, and very long words tend to be very sparsely con-
nected. This pattern appears to hold true across lexicons derived from child speech,
child-directed speech, and adult-directed speech, which only serves to underscore the
point that we need not appeal to anything other than the word length distributions to
explain their structural differences.
Chapter 5

Conclusion

We have demonstrated empirically that the word-length distribution and phoneme inventory size influence the topology of a phonological network. We hypothesised that differences in phonological network structure across lexicons derived from corpora of child, child-directed, and adult-directed speech could be parsimoniously explained by differences in their word-length distributions, and validated this hypothesis by showing that when their word-length distributions were controlled, phonological network statistics were much more similar across the three lexicons.

We also predicted that the statistics of phonological networks which included inflectional variants would deviate more from random baselines when inflectional variants were included than when the lexicon was restricted to lemmas, and that the magnitudes of these differences would be greater in languages with richer nominative case paradigms. These predictions were borne out in our comparative analysis of phonological networks across eight languages. We argued that differences in the ways in which lexicons are sampled would confound comparisons of their phonological network structures, and by comparing the phonological network statistics of different languages across a range of lexicon sizes, we have demonstrated that the statistics of phonological networks built from samples of relatively frequent words can have markedly different values than those of phonological network built from larger samples which include more infrequent words. Interestingly, the trajectories of the network statistics as more and more infrequent words are included differ considerably across languages, and not all of these differences can be attributed to differences in word-length distributions and phoneme inventory sizes. Hence while our main objective was to show that different lexicon sizes and the inclusion or exclusion of inflectional variants confound the comparison of phonological networks, it seems promising that
further consideration of the different ways in which these factors affect the network statistics of different languages could help shed more light on linguistic typology or lexical processing.
Appendix A

A

A.1 Tables of frequency thresholds and resulting sublexicon sizes
<table>
<thead>
<tr>
<th>English Lemmas</th>
<th>English Wordforms</th>
<th>French Lemmas</th>
<th>French Wordforms</th>
<th>Dutch Lemmas</th>
<th>Dutch Wordforms</th>
<th>Spanish Wordforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>Size</td>
<td>Threshold</td>
<td>Size</td>
<td>Threshold</td>
<td>Size</td>
<td>Threshold</td>
</tr>
<tr>
<td>none</td>
<td>44841</td>
<td>none</td>
<td>43641</td>
<td>4</td>
<td>52661</td>
<td>none</td>
</tr>
<tr>
<td>0</td>
<td>32964</td>
<td>0</td>
<td>34638</td>
<td>0.13</td>
<td>32430</td>
<td>32</td>
</tr>
<tr>
<td>1</td>
<td>30181</td>
<td>0.01</td>
<td>30379</td>
<td>0.16</td>
<td>29934</td>
<td>37</td>
</tr>
<tr>
<td>2</td>
<td>28254</td>
<td>0.02</td>
<td>28876</td>
<td>0.2</td>
<td>27854</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td>26798</td>
<td>0.04</td>
<td>26532</td>
<td>0.24</td>
<td>26379</td>
<td>46</td>
</tr>
<tr>
<td>4</td>
<td>25545</td>
<td>0.06</td>
<td>25465</td>
<td>0.27</td>
<td>25292</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>24445</td>
<td>0.09</td>
<td>24315</td>
<td>0.31</td>
<td>24019</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>22702</td>
<td>0.14</td>
<td>21917</td>
<td>0.37</td>
<td>22491</td>
<td>42</td>
</tr>
<tr>
<td>7</td>
<td>20671</td>
<td>0.19</td>
<td>20441</td>
<td>0.47</td>
<td>20462</td>
<td>49</td>
</tr>
<tr>
<td>8</td>
<td>16428</td>
<td>0.41</td>
<td>16394</td>
<td>0.81</td>
<td>16279</td>
<td>69</td>
</tr>
<tr>
<td>9</td>
<td>13886</td>
<td>0.67</td>
<td>13845</td>
<td>1.16</td>
<td>13757</td>
<td>98</td>
</tr>
<tr>
<td>10</td>
<td>12110</td>
<td>0.95</td>
<td>12051</td>
<td>1.53</td>
<td>11968</td>
<td>126</td>
</tr>
<tr>
<td>11</td>
<td>10864</td>
<td>1.21</td>
<td>10842</td>
<td>1.87</td>
<td>10759</td>
<td>154</td>
</tr>
<tr>
<td>12</td>
<td>9190</td>
<td>1.73</td>
<td>9170</td>
<td>2.57</td>
<td>9123</td>
<td>178</td>
</tr>
<tr>
<td>13</td>
<td>7562</td>
<td>2.58</td>
<td>7547</td>
<td>3.68</td>
<td>7511</td>
<td>293</td>
</tr>
<tr>
<td>14</td>
<td>5128</td>
<td>3.58</td>
<td>5123</td>
<td>7.23</td>
<td>5098</td>
<td>491</td>
</tr>
<tr>
<td>15</td>
<td>4001</td>
<td>8.33</td>
<td>3998</td>
<td>10.8</td>
<td>3981</td>
<td>668</td>
</tr>
<tr>
<td>16</td>
<td>3315</td>
<td>11.47</td>
<td>3313</td>
<td>14.3</td>
<td>3295</td>
<td>842</td>
</tr>
<tr>
<td>17</td>
<td>2839</td>
<td>14.56</td>
<td>2835</td>
<td>17.83</td>
<td>2824</td>
<td>1013</td>
</tr>
<tr>
<td>18</td>
<td>2228</td>
<td>21.19</td>
<td>2225</td>
<td>25.35</td>
<td>2216</td>
<td>1355</td>
</tr>
<tr>
<td>19</td>
<td>1694</td>
<td>31.66</td>
<td>1692</td>
<td>36.6</td>
<td>1688</td>
<td>1843</td>
</tr>
</tbody>
</table>

Table A.1: Frequency thresholds applied to the English, French, Dutch, and Spanish lexicons, and the resulting sublexicon sizes. Where cells are empty, there was no frequency threshold that could produce a sublexicon of a similar size to the corresponding English Lemmas sublexicon.
Table A.2: Frequency thresholds applied to the German, Portuguese, Polish, and Basque lexicons, and the resulting sublexicon sizes. Where cells are empty, there was no frequency threshold that could produce a sublexicon of a similar size to the corresponding English Lemmas sublexicon.
### Appendix B

#### B.1 List of CHILDES corpora from which the Child Speech lexicon was extracted

<table>
<thead>
<tr>
<th>Bates</th>
<th>Feldman</th>
<th>Peters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernstein</td>
<td>Gleason</td>
<td>Post</td>
</tr>
<tr>
<td>Bliss</td>
<td>Goad</td>
<td>Providence</td>
</tr>
<tr>
<td>Bloom70</td>
<td>HSLLD</td>
<td>Rollins</td>
</tr>
<tr>
<td>Bloom73</td>
<td>Haggerty</td>
<td>Sachs</td>
</tr>
<tr>
<td>Bohannon</td>
<td>Higginson</td>
<td>Snow</td>
</tr>
<tr>
<td>Braunwald</td>
<td>Inkelas</td>
<td>Soderstrom</td>
</tr>
<tr>
<td>Brent</td>
<td>Kuczaj</td>
<td>Stanford</td>
</tr>
<tr>
<td>Brown</td>
<td>MacWhinney</td>
<td>Suppes</td>
</tr>
<tr>
<td>Clark</td>
<td>McCune</td>
<td>Tardif</td>
</tr>
<tr>
<td>ComptonPater</td>
<td>McMillan</td>
<td>Valian</td>
</tr>
<tr>
<td>Cornell</td>
<td>Morisset</td>
<td>VanHouten</td>
</tr>
<tr>
<td>Davis</td>
<td>Nelson</td>
<td>VanKleeck</td>
</tr>
<tr>
<td>Demetras1</td>
<td>NewEngland</td>
<td>Warren</td>
</tr>
<tr>
<td>Demetras2</td>
<td>PaidoEnglish</td>
<td>Weist</td>
</tr>
</tbody>
</table>
Bibliography


