Graph-theoretic Properties of the Class of Phonological Neighbourhood Networks

Rory Turnbull
Newcastle University
rory.turnbull@newcastle.ac.uk

Abstract

This paper concerns the structure of phonological neighbourhood networks, which are a graph-theoretic representation of the phonological lexicon. These networks represent each word as a node and links are placed between words which are phonological neighbours, usually defined as a string edit distance of one. Phonological neighbourhood networks have been used to study many aspects of the mental lexicon and psycholinguistic theories of speech production and perception. This paper offers preliminary graph-theoretic observations about phonological neighbourhood networks considered as a class. To aid this exploration, this paper introduces the concept of the *hyperlexicon*, the network consisting of all possible words for a given symbol set and their neighbourhood relations. The construction of the hyperlexicon is discussed, and basic properties are derived. This work is among the first to directly address the nature of phonological neighbourhood networks from an analytic perspective.

1 Motivation

Recent work in phonological psycholinguistics has investigated the structure of the lexicon through the use of phonological neighbourhood networks (Chan and Vitevitch, 2010; Turnbull and Peperkamp, 2017; Siew, 2013; Siew and Vitevitch, 2020; Shoemark et al., 2016). A phonological neighbourhood network is a representation of the lexicon where each word is treated as a node and a link is placed between nodes if and only if those two nodes are phonological neighbours. Two words are neighbours if their string edit distance, in terms of phonological representation, is one. In other words, the neighbours of a word w are all the words that can be formed by the addition, deletion, or substitution of a single phoneme from w. The neighbourhood relation is symmetric (if w is a neighbour of w', then w' is necessarily a neighbour of w), intransitive (if w is a neighbour of w', and w' is a neighbour of w'', it is not necessarily the case that w is a neighbour of w''), and anti-reflexive (w cannot be a neighbour of itself).

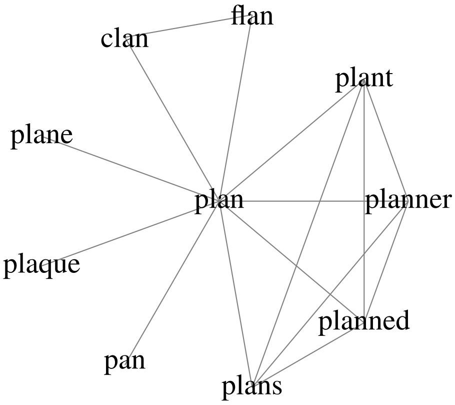

Figure 1: Example phonological neighbourhood network centred around the English word *plan*. Note that some neighbours of a word are neighbours of each other. Adapted from Turnbull and Peperkamp (2017).

Figure 1 shows an abbreviated phonological neighbourhood network for some words of English. One advantage of this representation is that it permits analysis with the methods of network science and graph theory, and work so far has shown a good deal of promise in modeling psycholinguistic properties of the lexicon with these methods (Chan and Vitevitch, 2010; Vitevitch, 2008). A common analysis technique within network science is to compare a given network with a randomly generated one that has the same number of nodes and links. Notable features of the target network relative to the random network are likely due to intrinsic properties of the target network, rather than chance. From this structure one can then infer details about the organising principles that generated the network originally.

For phonological neighbourhood networks, however, this method is often inappropriate, as many

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 233–240
Online Event, June 10, 2021. ©2021 Association for Computational Linguistics

logically possible network structures are not possible phonological neighbourhood networks. This fact is because the links between nodes—the neighbourhood relations—are intrinsic to the definitions of the nodes themselves. Changing a link between nodes necessarily means changing the content of a node, which then could entail other changes to other links. This problem was highlighted by Turnbull and Peperkamp (2017),[1] who instead chose to randomly generate *lexicons* and derive networks from those lexicons. However, randomly generated lexicons do not guarantee the same number of links will be present in the resulting network, making it difficult to compare like with like. For this reason, studying phonological neighbourhood networks as a class, and discovering their defining characteristics, is an important methodological goal for psycholinguists.

This research therefore seeks to answer the following broad questions: What are the distinctive characteristics of phonological neighbourhood networks, including their definitions in terms of edge sets and vertex sets, their extremal properties, and characterization of forbidden subgraphs? Is there an effective and efficient method by which phonological neighbourhood graphs can be distinguished from other graphs? The present paper lays the mathematical foundations for future investigations of both of these questions.

2 Preliminaries

This section briefly defines the basic mathematical definitions and operations used in the remainder of the paper. The reader is referred to standard textbooks in graph theory, such as Trudeau (1993) or Diestel (2005), for more details. As mathematical terminology and notation can vary between subfields, alternative names and characterizations of some objects are mentioned in the ensuing sections, but they are not strictly necessary to understand the arguments of this paper.

Networks can be modeled as mathematical objects known as *graphs*, which consist of vertices (nodes) and edges (links). Let G be an undirected graph with no self-loops with vertex set $V(G)$ and edge set $E(G)$. Let K_n denote the complete graph with n vertices and all possible edges.

A graph H is said to be a *subgraph* of a graph G if $V(H) \subseteq V(G)$ and $E(H) \subseteq V(G)$, that is, if the the edges and vertices of H are subsets of those of G. A subgraph H is an *induced subgraph* of G if every edge in $E(G)$ whose endpoints are both in $V(H)$ is present in $E(H)$. In other words, an induced subgraph can be obtained by the process of removing vertices (and any incident edges) from a graph, but not removing edges on their own. Figure 2 provides illustrative examples.

The *diamond* is K_4 with one edge removed. A *circle* C_k has the set of nodes $\{1, 2, ..., k\}$ and edge set $\{\{1, 2\}, \{2, 3\}, ..., \{(k-1), k\}, \{k, 1\}\}$. (Circle graphs that are induced subgraphs of a larger graph are also known as *k-holes*.) Figure 3 depicts the diamond and C_5.

A *star* S_k is a graph with one central vertex which is connected to k other unique vertices. No other vertices or edges exist. Figure 4 depicts the stars S_3 (also known as a *claw*), S_4, and S_6.

The *Cartesian product* $A \times B$ of two sets A and B is defined as

$$A \times B = \{(a, b) | a \in A, b \in B\}, \qquad (1)$$

that is, the Cartesian product of A and B is the set of all ordered pairs where the first element is a member of A and the second element is a member of B. For example, the Cartesian product of $\{a, b, c\}$ and $\{x, y\}$ is $\{(a, x), (a, y), (b, x), (b, y), (c, x), (c, y)\}$.

The Cartesian product $G \square H$ of two graphs G and H has the vertex set

$$V(G \square H) = V(G) \times V(H). \qquad (2)$$

A given vertex (a, x) is linked with another vertex (b, y) if $a = b$ (the first elements are identical) and $\{x, y\} \in E(H)$ (the second elements are linked in H), or if $x = y$ (the second elements are identical) and $\{a, b\} \in E(G)$ (the first elements are linked in G). To aid understanding, Figure 5 depicts an example of the Cartesian product of two graphs, G and H. Graph G has $V(G) = \{a, b, c\}$ and $E(G) = \{\{a, b\}, \{b, c\}\}$. Graph H has $V(H) = \{x, y\}$ and $E(H) = \{\{x, y\}\}$. Observe how G and H can be seen in $G \square H$ as two orthogonal dimensions. Note also that the total number of vertices in $G \square H$ is equal to the product of the number of vertices in G and H.

We further denote the *Cartesian exponent* of a graph G as

$$G^{\square n} = \underbrace{G \square G \square G \ldots G}_{n}, \qquad (3)$$

[1] See also Gruenenfelder and Pisoni (2009) for related concerns.

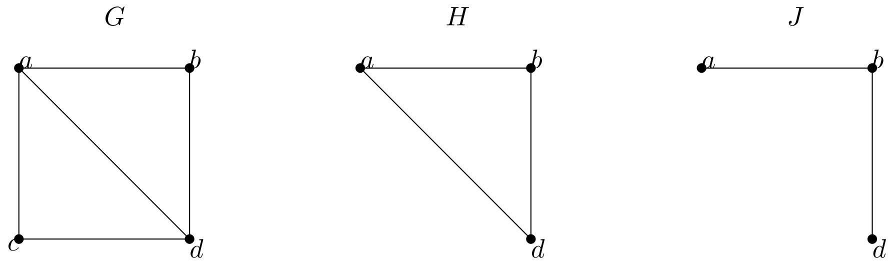

Figure 2: Three graphs. H is an induced subgraph of G formed through the removal of vertex c and its incident edges. J is also a subgraph of G, but it is not an induced subgraph due to the fact that the edge between vertices a and d is missing.

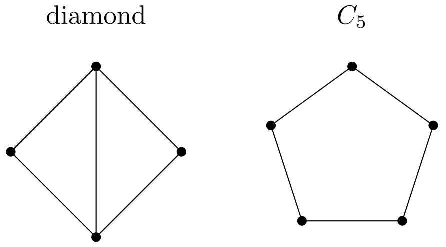

Figure 3: The diamond graph and C_5.

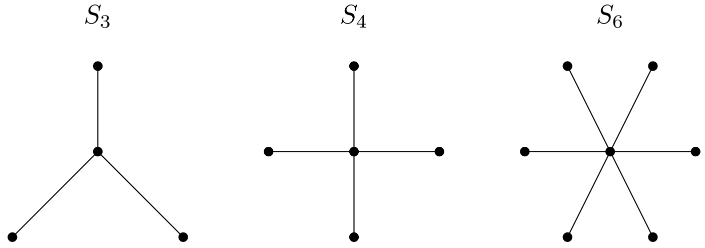

Figure 4: Star graphs S_3, S_4, and S_6.

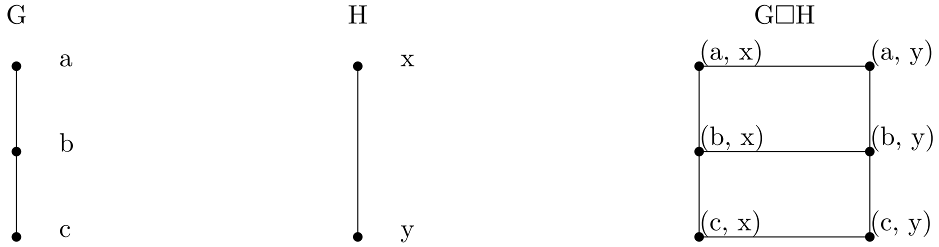

Figure 5: Graphs G and H and the Cartesian product $G\Box H$.

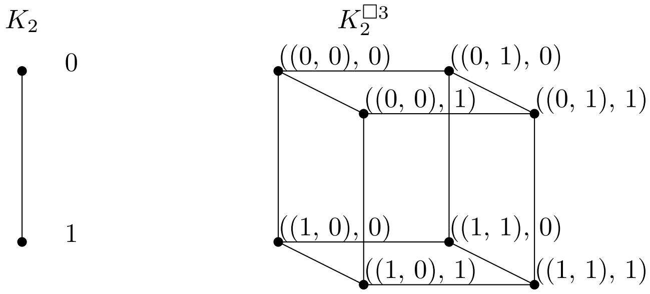

Figure 6: The complete graph K_2 and Cartesian exponent $K_2^{\square 3}$, i.e. $K_2 \square K_2 \square K_2$.

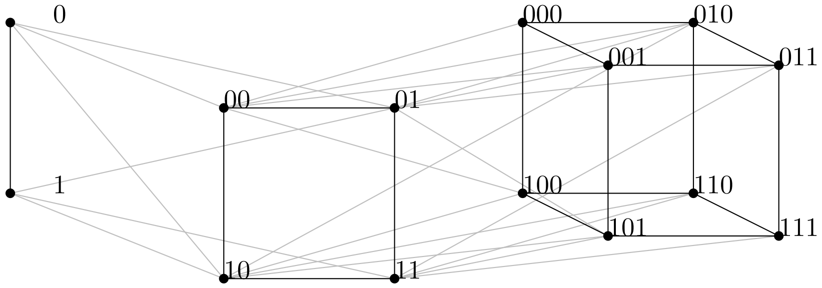

Figure 7: The hyperlexicon $\mathcal{H}(2, \{1, 2, 3\})$. Here the alphabet is defined as $\{0, 1\}$ but any set of two symbols is possible. Edges between layers (i.e. phoneme additions/deletions) are drawn in grey; edges within layers (i.e. phoneme substitutions) are drawn in black.

that is, the Cartesian product of G with itself $n-1$ times. Figure 6 depicts the graphs K_2 and $K_2^{\square 3}$.

It can be seen that $K_2^{\square 2}$ is a square, $K_2^{\square 3}$ is a cube, and $K_2^{\square n}$ is an n-dimensional hypercube.[2] Likewise, the vertex labels of $K_m^{\square n}$ are equivalent to all strings of length n drawn from an alphabet of m symbols. The edges of $K_m^{\square n}$ are equivalent to the neighbourhood relations of such strings. These facts establish the basis upon which we can use these tools to model phonological neighbourhood networks.

3 The Hyperlexicon

In this paper we introduce the concept of the *hyperlexicon*. A hyperlexicon $\mathcal{H}(\phi, L)$ is defined as the phonological neighbourhood network generated from all possible string sequences of lengths $\{\ell_1, \ldots \ell_n\}$ for $\ell \in L$ over an alphabet of length ϕ.

Figure 7 depicts the hyperlexicon of all 'words' of length 1, 2, and 3, over the alphabet of `0` and `1`. Stella and Brede (2015) observed that the set of all possible phoneme sequences (i.e. the hyperlexicon) is composed of multiple 'layers', each corresponding to a distinct member of L. This layered structure can be clearly seen in Figure 7. Edges within a layer correspond to neighbours by substitution, while edges between layers correspond to neighbours by deletion or insertion. Note further that each layer is isomorphic to $K_\phi^{\square \ell}$, the ℓ^{th} Cartesian exponent of the complete graph with ϕ vertices.[3]

Imagine now a hypothetical lexicon consisting of the words `1`, `00`, `10`, and `110`. The phonological neighbourhood network of this lexicon is depicted in Figure 8, overlaid on the hyperlexicon from Figure 7. It can be seen that this lexicon's network is an induced subgraph of the hyperlexicon.

Indeed, phonological neighbourhood networks are necessarily induced subgraphs of the hyperlexicon. For example, the English lexicon consists of strings of varying lengths, with the set of English phonemes as its 'alphabet'. There are some strings of English phonemes which are not part of the English lexicon—i.e. nonwords such as *blick* and *pmisgkr*. The set of words in the English lexicon, then, is a subset of the set of all logically possible strings of English phonemes. A hyperlexicon corresponds to the neighbourhood network derived from a set of all logically possible strings

[2] More generally, $K_m^{\square n}$ is an $m \times m$ Rook's graph in n dimensions.

[3] Each layer can also be characterized as an expansion of the hypercube graph Q_ℓ, or as a Hamming graph $\text{Ham}(\ell, \phi)$.

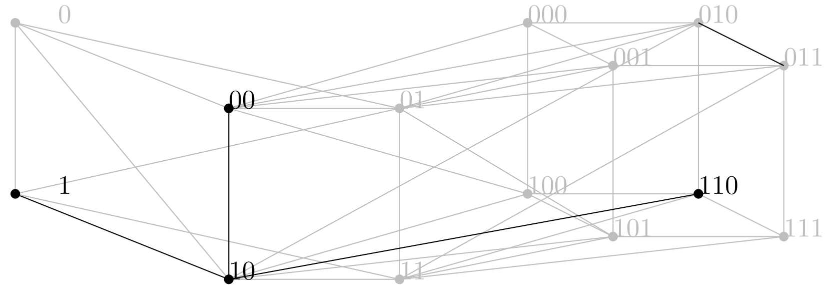

Figure 8: The phonological neighbourhood network (in black) of the lexicon `1`, `00`, `10`, and `110`, depicted as a subgraph of the hyperlexicon $\mathcal{H}(2, \{1, 2, 3\})$ (in grey).

of length L, given some set of phonemes of length ϕ. Any subset of this set of strings will correspond to an induced subgraph of the hyperlexicon. Any phonological neighbourhood network, then, with words of lengths in L constituted from ϕ distinct phonemes, is necessarily an induced subgraph of the hyperlexicon $\mathcal{H}(\phi, L)$. Studying properties of the hyperlexicon therefore gives us insight into the possible structures of phonological neighbourhood networks.

The vertex set of the hyperlexicon is given by

$$V(\mathcal{H}(\phi, L)) = \bigcup_{\ell \in L} V(K_\phi^{\square \ell}), \tag{4}$$

that is, the set union of each layer's vertices. The number of vertices of $\mathcal{H}(\phi, L)$ is the sum of the size of each layer, which is

$$\sum^{L} \phi^{\ell}. \tag{5}$$

If L is contiguous, $\mathcal{H}(\phi, L)$ is necessarily connected (i.e. there is exactly one connected component); if L is not contiguous,[4] then $\mathcal{H}(\phi, L)$ has multiple connected components.

4 The Edges between the Layers of the Hyperlexicon

Defining the edge set of $\mathcal{H}(\phi, L)$ is less straightforward than the vertex set and is not fully solved. Within each layer, the edges are the same as in the graph $K_\phi^{\square \ell}$. Between the layers the situation is considerably more complex. To begin, we first determine the number of possible unique neighbours for any word. For a word of length ℓ in a language with ϕ distinct phonemes, neighbours are generated through the addition, deletion, or substitution of a single phoneme. The number of possible neighbours can be shown to depend upon word length ℓ, alphabet size ϕ, and the number of pairs of adjacent identical phonemes, described below.

4.1 Substitutions

It is straightforward to demonstrate that there are

$$\phi\ell - \ell \tag{6}$$

possible substitutions. This statement follows from the fact that neighbourhood is an anti-reflexive relation, so vacuously substituting a phoneme for itself will not generate a neighbour.

4.2 Additions

The number of additions can be derived from the fact that each of ϕ symbols can be added to $\ell + 1$ positions, which gives $\phi(\ell + 1)$. However, for each insertion position, one of these ϕ phonemes will result in a string which is identical to an insertion of the same phoneme at a different location. For example, prefixing `a` onto the beginning of `ab` is equivalent to inserting `a` into the middle of `ab`: they both result in `aab`. The number of additions is therefore

$$\phi(\ell + 1) - \ell \tag{7}$$

which simplifies to Equation (6) plus ϕ:

$$\phi\ell - \ell + \phi. \tag{8}$$

4.3 Deletions

The number of deletions is not constant and depends upon the structure of the word. For example, although there are three distinct deletion positions in a possible word `aaa`, all three of them lead to the same unique word `aa`; so practically speaking

[4] Such a scenario is plausible for languages with strict phonotactics requiring an obligatory onset and forbidding codas, i.e. all syllables must be CV. For such languages, $L = \{2, 4, 6, 8, \dots\}$. Hua (Blevins, 1995) and Senufo (Kientz, 1979) have been reported to have this kind of syllable structure.

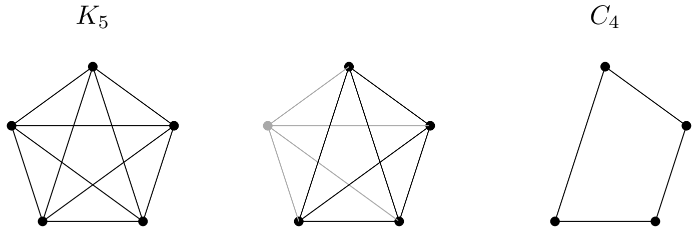

Figure 9: In attempting to generate C_4 (right) from K_5 (left), a single vertex must be removed. However, as shown by the middle graph, removal of a single vertex (in grey) results in a graph with too many edges. This is true no matter which vertex we choose to remove. C_4 is therefore not an induced subgraph of K_5; we can call C_4 a *forbidden subgraph* of K_5.

there is only one possible deletion. On the other hand, all three possible deletions on abc result in three unique strings (namely, bc, ac, ab), so it has three deletions.

The actual number of possible deletions depends on the number of pairs of identical adjacent symbols.[5] Pairs of identical adjacent symbols act as a single symbol for the purposes of counting possible deletion sites. The number of deletions is therefore

$$\ell - a, \tag{9}$$

where a is the number of pairs of adjacent identical symbols in the word. For example abba has only 3 possible deletions, despite being of length 4. (For this word, a deletion at position 2 is equivalent to a deletion at position 3; the sequence bb can be essentially treated as a single symbol for the purposes of counting deletion sites.)

All words allow at least 1 deletion, and some words allow as many as ℓ deletions.

4.4 Vertex Degree

From the sections above, it follows that each vertex in $\mathcal{H}(\phi, L)$ has $\phi\ell - \ell$ edges to other nodes in the same layer as it. If there is a higher layer, then each node also has $\phi\ell - \ell + \phi$ edges leading to nodes in that layer. If there is a lower layer, then each node has between 1 and ℓ edges leading to that layer.

5 Forbidden Subgraphs

Finally, we begin to attempt to characterize the class of hyperlexicons in terms of forbidden subgraphs. A forbidden subgraph of G is any graph which is not isomorphic to any induced subgraph of G. For example, there is no induced subgraph of K_5 which is isomorphic to C_4. This fact is illustrated in Figure 9. C_4 is therefore a *forbidden subgraph* of K_5. Graph structures which are impossible within a hyperlexicon are also impossible within real phonological networks, because real phonological networks are induced subgraphs of a hyperlexicon. Understanding the forbidden subgraphs of a hyperlexicon therefore allows us to understand possible natural language networks.

5.1 Forbidden Subgraphs of individual layers

A hyperlexicon is composed of layers. Each layer is $K_\phi^{\square\ell}$, the ℓth cartesian exponent of K_ϕ. For the special case of $\ell = 2$ (i.e. words of length two), these graphs have been studied in the mathematical literature under the names of *Rooks' graphs*, *gridline graphs*, *adjacency graphs*, and *graphs of* $(0, 1)$ *matrices*. Peterson (2003) studied these graphs in cases where $\ell > 2$, and established that the diamond and C_5 are among the forbidden subgraphs of $K_\phi^{\square\ell}$.

The 3-star S_3 has been shown to be a forbidden subgraph of $K_\phi^{\square 2}$ (Hedetniemi, 1971). More generally, no layer at length ℓ has $S_{\ell+1}$ as an induced subgraph. This observation follows from the pigeonhole principle: the first ℓ vertices of $S_{\ell+1}$ can be found in the ℓ dimensions of the graph. The final vertex must be in one of the dimensions already considered, and therefore must be adjacent to an existing vertex. This leads to a triangle, meaning the induced subgraph is no longer a star.

These structures, forbidden from each individual layer of the hyperlexicon, are not forbidden from the hyperlexicon as a whole. Within the hyperlexicon $\mathcal{H}(3, \{1, 2, 3\})$ we observe both the diamond and C_5; see Figures 10 and 11.[6] Similarly, for a hy-

[5] Using the terminology of combinatorics on words, a "pair of identical adjacent symbols" can be understood as a square of length 2.

[6] We have been unable to find any induced C_5 in cases where $\phi < 3$. While this conjecture might be of mathematical interest, it is not relevant to our main use-case of phonological

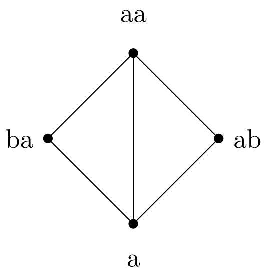

Figure 10: The diamond graph as an induced subgraph of a hyperlexicon.

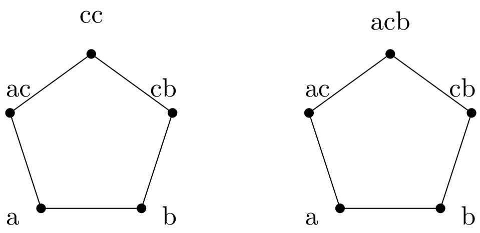

Figure 11: C_5 as two induced subgraphs of a hyperlexicon.

perlexicon with $\max(L) = 4$, the star S_5 is present as an induced subgraph, as shown in Figure 12.

Since these structures cannot occur within each layer, it follows that their existence within a hyperlexicon must necessarily span more than one layer. Indeed, we hypothesize that in the case of $S_{\ell+1}$ where $\ell = \max(L)$, this structure must necessarily span three layers.

5.2 Forbidden Subgraphs of the Entire Hyperlexicon

No hyperlexicon has $K_{\phi+2}$ as an induced subgraph. K_ϕ exists, as this constitutes the 'dimensions' of each layer. From K_ϕ it is possible to induce $K_{\phi+1}$ by adding a vertex from one layer down. For example, the string a is adjacent to aa, ab, ac, and so on. However there is no other vertex in the lower layer which is adjacent to all of a's neighbours and to a itself. $K_{\phi+2}$ is therefore not an induced subgraph of the hyperlexicon.

6 Conclusion

This paper has reviewed the basic structure of hyperlexicon graphs. Induced subgraphs of hyperlexicon graphs typify the class of phonological neighbourhood networks. It is hoped that the preliminary results presented here will spur further work on the nature of phonological neighbourhood networks as formal objects. This work in turn has methodological implications for evaluating and measuring phonological neighbourhood networks derived from natural languages.

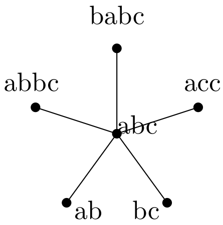

Figure 12: S_5 as an induced subgraph of a hyperlexicon with $\max(L) = 4$.

Acknowledgements

I am grateful to Paul Tupper and several anonymous reviewers for their comments on an earlier draft. All errors and omissions are my own.

References

Juliette Blevins. 1995. The syllable in phonological theory. In John Goldsmith, editor, *Handbook of phonological theory*, pages 206–244. Blackwell.

K Y Chan and M S Vitevitch. 2010. Network structure influences speech production. *Cognitive Science*, 34(4):685–697.

Reinhard Diestel. 2005. *Graph Theory*, 3rd edition. Springer, New York, NY.

Thomas M Gruenenfelder and David B Pisoni. 2009. The lexical restructuring hypothesis and graph theoretic analyses of networks based on random lexicons. *Journal of Speech, Language, and Hearing Research*, 52(2):596–609.

Stephen T Hedetniemi. 1971. Graphs of (0,1)-matrices. In M Capobianco, J B Frechen, and M Krolik, editors, *Recent Trends in Graph Theory*, pages 157–171. Springer, Berlin.

Albert Kientz. 1979. *Dieu et les génies: Récits étiologiques senoufo (Côte-d'Ivoire)*. Centre national de la recherche scientifique, Paris.

Dale Peterson. 2003. Gridline graphs: a review in two dimensions and an extension to higher dimensions. *Discrete Applied Mathematics*, 126:223–239.

Philippa Shoemark, Sharon Goldwater, James Kirby, and Rik Sarkar. 2016. Towards robust crosslinguistic comparisons of phonological networks. In *Proceedings of the 14th ACL SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 110–120.

networks, as all known natural languages possess considerably more than 3 phonemes.

Cynthia S Q Siew. 2013. Community structure in the phonological network. *Frontiers in Psychology*, 4:553.

Cynthia S Q Siew and Michael S Vitevitch. 2020. Investigating the influence of inverse preferential attachment on network development. *Entropy*, 22(9):1029.

Massimo Stella and Markus Brede. 2015. Patterns in the English language: Phonological networks, percolation and assembly models. *Journal of Statistical Mechanics: Theory and Experiment*, 5:P05006.

Richard J Trudeau. 1993. *Introduction to Graph Theory*. Dover, New York, NY.

Rory Turnbull and Sharon Peperkamp. 2017. What governs a language's lexicon? determining the organizing principles of phonological neighbourhood networks. In Hocine Cherifi, Sabrina Gaito, Walter Quattrociocchi, and Alessandra Sala, editors, *Complex Networks & Their Applications V*, volume 693 of *Studies in Computational Intelligence*, pages 83–94. Springer, Cham, Switzerland.

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