International Journal of Bifurcation and Chaos, Vol. 20, No. 3 (2010) 679–685

DOI: 10.1142/S021812741002596X

THE STRUCTURE OF PHONOLOGICAL NETWORKS ACROSS MULTIPLE LANGUAGES

SAMUEL ARBESMAN
Department of Health Care Policy,
Harvard Medical School, 180 Longwood Avenue,
Boston, MA 02115, USA

STEVEN H. STROGATZ
Theoretical and Applied Mechanics,
223 Kimball, Cornell University,
Ithaca, NY 14853, USA

MICHAEL S. VITEVITCH
Psychology, 1415 Jayhawk Blvd.,
University of Kansas, Lawrence, KS 66045, USA

Received July 24, 2009; Revised August 5, 2009

The network characteristics based on the phonological similarities in the lexicons of several languages were examined. These languages differed widely in their history and linguistic structure, but commonalities in the network characteristics were observed. These networks were also found to be different from other networks studied in the literature. The properties of these networks suggest explanations for various aspects of linguistic processing and hint at deeper organization within the human language.

Keywords: Networks; language; phonology.

1. Introduction

The results of numerous graph-theoretic analyses suggest that a number of principles may influence the emergent structures found in a wide variety of complex systems, including information, social, technological and biological networks [Strogatz, 2001; Albert & Barabási, 2002; Newman, 2003]. These unifying characteristics include small-world properties, distinct community structure and scale-free distributions of the network connectivity.

Many aspects of language can be examined from a network perspective as well. Numerous studies have been conducted on semantic networks, where relationships in meaning have been created between words. These are often based on thesauri, word-associations in corpori or from academic databases [Ferrer & Ricard, 2001; Motter *et al.*, 2002]. In addition, linguistic networks have formed from orthographic similarities of words (how words are spelled) [Kello & Beltz, 2007]. Lastly, language can be viewed from the sounds of words (their phonological structure), where words that sound similar are neighbors. Although previous experiments have examined small portions of phonological networks (nearest neighbors of words) in the context of psycholinguistic theories of spoken word recognition [Luce & Pisoni, 1998], the first graph-theoretic analysis of an entire language network only appeared more recently [Vitevitch, 2008].

In these phonological networks, words in a language are represented as vertices or nodes, and an edge is placed between them if the words sound

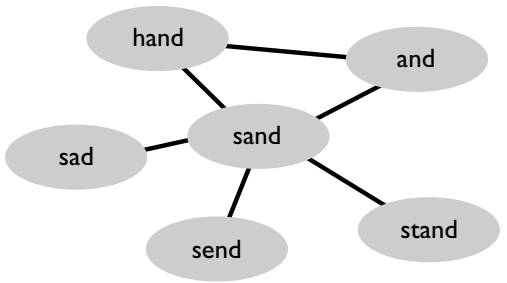


Fig. 1. A phonological network for five English words.

similar to each other (differing only by a single phoneme, or sound segment). For example, as shown in Fig. 1, vertices representing the words *hand*, *send*, *sad*, *and* and *stand* would all have edges connecting them to the vertex for the word *sand*. These phonological networks are especially intriguing to examine because psycholinguistic studies suggest that several characteristics of the network influence cognitive processing, such as word recognition and retrieval [Steyvers & Tenenbaum, 2005; Vitevitch, 2008]. In addition, much work has shown that degree influences word recognition [Luce & Pisoni, 1998], word production [Vitevitch, 2002], and word learning [Storkel *et al.*, 2006]. Furthermore, it has been recently demonstrated that clustering coefficients also influence production and recognition [Chan & Vitevitch, in press; Chan & Vitevitch, 2009].

In examining English, Vitevitch [2008] found that its phonological network had a small giant component (the largest connected portion of the graph), with many other smaller components ("islands"). This property is distinct from most other complex networks observed in the literature. In addition, the degree distribution (the distribution of the number of edges per node) was not well modeled by a scale-free distribution, or a power law, and was not well-approximated by a Poisson distribution. This is surprising since it is reasonable to expect a power law as in a preferential attachment model [Barabási & Albert, 1999], as it has been shown that words in the lexicon with high degree tend to have novel words attached to them more easily than words in the lexicon with low degree [Storkel *et al.*, 2006].

Here, we aimed explore the generality of these results, by doing the first comparative study of multiple languages, using phonological networks. Similar network characteristics across a variety of languages would hint toward principles that are common to all languages, whereas differences in network measures would provide a quantitative way to describe and categorize the languages of the world. For example, while relatively little cross-linguistic research has been done in psycholinguistics, it has been shown that degree has different influences in English compared to Spanish [Vitevitch & Rodríguez, 2005; Vitevitch & Stamer, 2006]. Due to this, it is important to evaluate other network characteristics across languages in order to get a better understanding of language processing in general. The findings in English and Spanish suggest that the same structure might have a different influence in a different language as a function of other "structural" characteristics that are not captured in current measures [Arbesman *et al.*, in preparation]. We examined some of the properties examined by Vitevitch in English, as well as a number of others, and found that phonological networks all have certain properties distinct from other types of complex networks (such as biological and social networks).

2. Methods

The network structure of selected languages was examined to determine the generality of the network characteristics previously observed in English [Vitevitch, 2008]. In addition to English, the following languages were examined: Spanish, Mandarin, Hawaiian and Basque (see Table 1). These languages are representative examples of different language families and are of wide variety in their linguistic properties.

English is an Indo-European language from the Germanic branch, whereas Spanish comes from the Romance branch of the Indo-European family of languages. Mandarin, a Sino-Tibetan language, differs from English, Spanish, Hawaiian and Basque in that it also uses tones to convey word meanings (e.g. "fan" with a high level tone means sail, with a rising tone means trouble, with a dipping tone means turn, and with a falling tone means rice). Tone was not included in the phonological transcriptions, however. Hawaiian is an Austronesian language with a phoneme inventory (the number of consonants and vowels in the language) that is smaller than those found in English, Spanish, Mandarin and Basque. Finally, Basque (or Euskara) is a linguistic isolate, meaning that it is not (or has not yet been identified as) a member of a given language family. Additional differences, such as those

Table 1. Summary information of phonological networks in several languages. GC stands for Giant Component and RN stands for Random Network. ASPL stands for Average Shortest Path Length.

	English	Spanish	Mandarin	Hawaiian	Basque
Network Size (number of words)	19,323	122,066	30,086	2,578	99,321
Giant Component Size (fraction)	6,498 (0.34)	44,833 (0.37)	19,712 (0.66)	1,406 (0.55)	35,173 (0.35)
Assortative Mixing by Degree (r)	0.657	0.762	0.654	0.556	0.719
ASPL	2.7	4.3	6.5	3.2	4.4
ASPL (GC)	6.1	10.3	10.1	5.5	10.4
ASPL of RN (using GC)	5.8	9.9	7.3	5.8	11.4
Clustering Coefficient	0.284	0.191	0.383	0.241	0.206
Clustering Coefficient of RN	8.35e-5	1.17e-5	8.55e-5	7.40e-4	1.21e-5
Transitivity	0.313	0.250	0.404	0.260	0.232
Ratio of Edges to Vertices	1.61	1.43	2.57	1.91	1.21
Ratio of Edges to Vertices (GC)	4.55	2.95	3.88	3.44	2.50

in morphology, exist among the languages that were selected for the present network analyses.

The phonological networks were constructed from a variety of sources. The English network contained the words from the Merriam-Webster Pocket Dictionary from 1964; this database has been used extensively in psycholinguistic studies [Luce & Pisoni, 1998]. The Hawaiian network was created in a similar manner using a Hawaiian Dictionary [Judd, 1980]. The words from the Spanish network consisted of the words in the LEXESP database [Sebastián-Gallés *et al.*, 2000], a large Spanish language corpus. The words in the Basque network were obtained in a manner similar to the words in the Spanish network [Perea *et al.*, 2006]. The Mandarin network uses the words from a database compiled in [Huang *et al.*, 1997].

3. Results

3.1. *Unique characteristics of the giant component*

3.1.1. *Giant component size*

The giant component sizes of the language networks were much smaller compared to other networks discussed in the literature. Typically, the giant component contains approximately 80–90% of the vertices [Newman, 2001]. However, in the present networks, the proportion of vertices in the giant component was much smaller, with some networks having less than 50% of the vertices in the giant component. The proportion of vertices in the giant components for comparably sized random networks, containing 70–80% of the vertices, are also larger than the values for the language networks [Callaway *et al.*, 2001]. This difference in giant component size suggests that these phonological networks may be more robust to node removal due to more tightly connected components, and indicates the prevalence of smaller components in the networks.

3.1.2. *Robustness to vertex removal*

In many systems studied to date, the mapping of node removal to its real-world equivalent is relatively intuitive (e.g. node removal in a network modeling an ecosystem is equivalent to a species becoming extinct). A reasonable linguistic analogue to node removal is that of the tip-of-the-tongue phenomenon, or similar situations where you know that you know the word (you have likely used it before), you have access to semantic information about it (you can describe it to people), yet the phonological word-form is temporarily inaccessible, since its activation does not cross the threshold for retrieval. It turns out that degree (i.e. neighborhood density) influences the likelihood that a word will be on the tip of the tongue [Vitevitch & Sommers, 2003], so understanding network structure and node removal might provide some insight into which words are likely to be vulnerable to the tip-of-the-tongue state, or suggest strategies to recover from it.

To evaluate the robustness of the networks, vertices were removed in two ways: at random, and in decreasing order by degree (number of edges connected to a vertex). These results are shown in Fig. 2. In scale-free networks, when vertices are randomly removed the mean shortest path length remains constant, whereas when vertices are removed in order of degree, the mean shortest path length increases dramatically [Newman, 2003]. In the language networks, however, both methods of

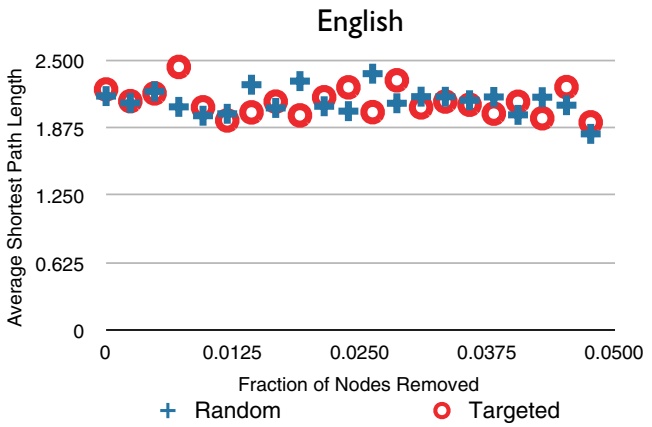


Fig. 2. An example run of node removal in English, either random or in a targeted fashion (in order by degree). Up to 5% of the nodes were removed, and all languages showed similar patterns to the above results. In addition, when the simulations were done only for the giant component, a similar constant, though elevated, value of the average shortest path length was found.

node removal resulted in little to no change in the mean shortest path lengths. The shortest path lengths were calculated using a sampling technique where 1000 nodes were chosen at random. Then, the distances to all other nodes (if part of the same component) were obtained and these path lengths were then all averaged, to give an estimate of the shortest path length. This sped up the calculations considerably. The extraordinary amount of robustness observed based on these common methods of node removal does seem intriguing and merits further examination.

3.1.3. *Assortative mixing*

In addition, we examined the assortative mixing by degree of the language networks, which is a measure of the correlation of degree between neighboring nodes. As seen in Table 1, all of the language networks had large and positive correlations of the degrees of connected vertices, indicating that high degree vertices tended to be connected to each other. Newman [2002] discussed how networks with assortative mixing by degree are more robust to vertex removal and percolate more easily (i.e. diseases or information spread easily) than networks with disassortative mixing. The high assortative mixing observed in the phonological networks is distinct from other types of networks: biological and technological networks often are disassortatively mixed, and social networks, which display assortative mixing, still have lower values of assortative mixing. Typical measures of assortativity for social networks are 0.1–0.3, and biological and technological networks are −0.1 to −0.2 [Newman, 2002]. On the other hand, phonological networks can be higher than 0.7.

High assortative mixing not only suggests robustness in the phonological networks, and highlights the resilience of lexical processing in the face of injury to the language related areas of the brain (i.e. stroke, or even the tip-of-the-tongue phenomenon discussed above), but it also has implications for the searchability of the phonological networks under intact conditions [Watts *et al.*, 2002]. This feature of the phonological network may contribute to the high rates of accuracy with which words are retrieved from the mental lexicon; one study estimated that healthy adult speakers make an error between 0.1–0.2% of the time they speak [Garnham *et al.*, 1981]. Lexical processing might proceed more slowly and errors in word retrieval might be more common if the phonological networks did not have such a robust structure. The phonological networks of patients with aphasia or other neurogenic disorders that disrupt language processing could be used to test this hypothesis.

3.2. *Small-world properties*

Although the languages differ in their history and linguistic characteristics, they all share a number of similarities in their network structure. An important commonality across the languages is that they all have the properties of a small-world network [Watts & Strogatz, 1998], that is, a high clustering coefficient and short vertex-to-vertex distance. The clustering coefficient can be calculated for each node (the average value of which is reported in Table 1), and is the fraction of neighbors of a given node that are neighbors with each other. It is also known as network density. The vertex-to-vertex distance, also known as the shortest path length, is the shortest number of hops in a network to go from one node to another. Since these networks have many components, the shortest path length from one node to another is only calculated for nodes that are in the same component [Newman, 2003]. In addition, the mean shortest path length was calculated just within the giant component of each language.

As seen in Table 1, the values for the clustering coefficient are many orders of magnitude larger than what would be expected from a comparably sized

random network — a network with the same number of nodes and edges — which can be calculated analytically [Watts & Strogatz, 1998]. The values of the clustering coefficient are also comparable to a similar measure referred to as transitivity, which is a more global measure of clustering [Newman, 2003].

On the other hand, the mean shortest path length of the language networks giant component, calculated using a random sample of 1000 nodes, was similar to the mean shortest path length for comparably sized random networks, and significantly shorter than the overall number of nodes in the network, as seen in Table 1 [Watts & Strogatz, 1998]. The statistics of the giant component were used for comparable random networks, because the overall ratio of edges to nodes is far lower than within the giant component, due to the large number of islands in the networks.

Since a small world structure is often a prerequisite for rapid search, and it is well-known that lexical retrieval processes are rapid and robust, it would be logical that the networks might be optimally structured for search. A clear future research direction is the examination of these networks for the properties, such as those discussed in [Kleinberg, 2000], that allow for rapid and robust search.

In addition, there exists the possibility of search (both in word recognition and production) within more than the single dimension of phonological similarity, which could include such additional dimensions as semantic similarity or syntatic relationships. This would lead to what would effectively be a fully connected network, where the other dimensions would guide search within the initial search space.

However, it must be noted that, unlike in social networks, where it is clear what a distance of three friends is, for example, it is not entirely clear what the qualitative difference is between a distance of 5 and 6 within phonological networks. This is important when looking at the average shortest path lengths of the giant components of the different language networks. For instance, is it relevant that this value for Mandarin (10.1) is twice that of Hawaiian (5.5)? While it is likely that this number is most relevant relative to the size of the entire network (they are all orders of magnitude smaller than the size of the lexica examined), these differences might hint at more significant distinctions between the languages examined.

The common occurrence of the small world property in networks observed in the literature may suggest that it is less a relevant property of language (since it is not unique to language) than simply an indicator that language is a fairly organic, unplanned construct. It is interesting, however, that the path length within a network appears to be an important property for language processing. A recent study [Yarkoni *et al.*, 2008] demonstrated that a measure related to path length in a phonological network (i.e. the minimum number of substitution, insertion, or deletion operations required to turn one word into another) influenced pronunciation times in visual word recognition tasks. Therefore, the relevance of different average path length across languages warrants further investigation.

3.3. *Degree distribution*

The degree distributions of scale-free networks obey a power law function, $P(z) \sim z^{-\alpha}$. In contrast to many observed networks, we find that the language networks deviate from this behavior. Instead, they are reasonably fit to truncated power laws, similar to scientific coauthorship networks [Newman, 2001], as seen in Table 2. A truncated power law, or a power law with an exponential cutoff, is defined as follows:

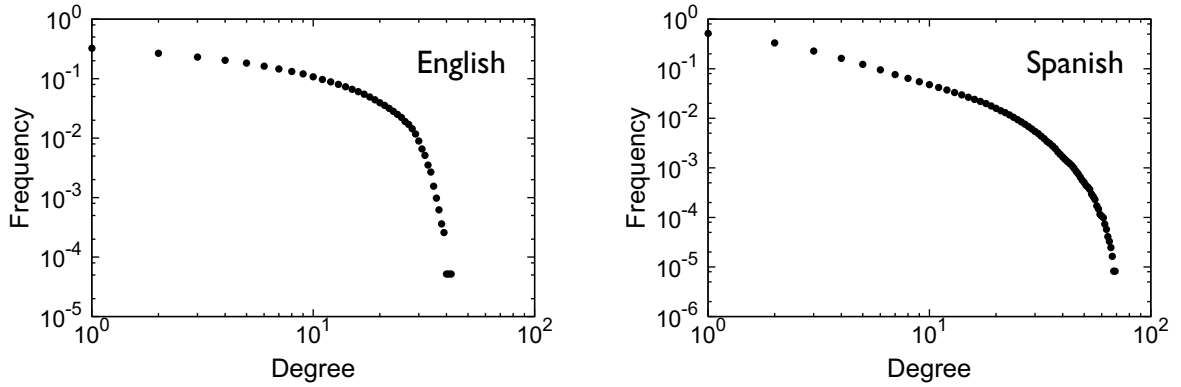
$$P(z) \sim z^{-\alpha} e^{-z/z_c} \tag{1}$$

Table 2 shows the parameters of the best fit of a truncated power law for the degree distribution of each language, as calculated by the methods found in [Clauset *et al.*, 2007]. All fits had p-values of less than 10^{-10}, in terms of the probability that they were better fit by a truncated power law than a traditional power law. In addition, as can be seen, Mandarin's fit is essentially an exponential distribution, with no power-law portion.

As mentioned earlier, it is reasonable to expect a power law as in a preferential attachment model

Table 2. Languages and best fit parameters for a truncated power law. All fits had p-values of less than 10^{-10}.

Language	Exponent (α)	Cutoff (z_c)
English	0.826	16.14
Spanish	0.815	7.06
Mandarin	−1.0	3.69
Hawaiian	0.270	7.34
Basque	0.575	4.56



(a) (b)

Fig. 3. The degree distributions of two of the language networks (English and Spanish), on a log–log scale. The final point for each distribution was not plotted, for legibility.

[Barabási & Albert, 1999], since it has been shown that words in the lexicon with high degree tend to have novel words attached to them more easily than words in the lexicon with low degree [Storkel *et al.*, 2006]. However, Amaral *et al.* [2000] found that if there is a constraint associated with the attachment of a new vertex (i.e. the vertex may only be able to accommodate a fixed number of edges), then a power law degree distribution, like that in the scale-free model proposed by Barabási and Albert [1999], is not likely to be observed. In the language networks (Fig. 3), a variety of constraints on word formation are present, such as the number of phonemes in the inventory of the language, the sequential arrangement of phonemes in words, the length of words, and the extent to which the language relies on morphemes (the smallest meaningful unit). All of these constraints limit the number of words that might be phonologically similar. Therefore, a truncated power law or similar distributions that decay faster than a traditional power law are reasonable as fits for the degree distributions in phonological networks.

4. Conclusion

The phonological networks of a variety of languages show a unique structure not found in other complex networks described in the literature. Despite coming from a diverse range of language families the networks all exhibited a common set of properties. Notably, the degree distribution is found to lie somewhere between a power law and an exponential distribution.

Furthermore, a small-world structure was observed, in conjunction with the distinguishing characteristic of the giant components as far smaller than typically observed. The small sizes of the giant component together with the strong assortative mixing by degree and the robustness of the network to the removal of vertices are suggestive to the resilience of language processing in the brain, although further study is necessary.

Together, these observed characteristics hint at some deeper organization within language. Despite surface differences among languages, there are important commonalities that have implications for the processing of language in humans. The intriguing characteristics of these networks merit further investigation from network scientists as well as psycholinguistic researchers.

Acknowledgments

Research supported in part by National Science Foundation grant DMS-0412757 to S. H. Strogatz, and in part by grants from the National Institutes of Health to the University of Kansas through the Schiefelbusch Institute for Life Span Studies (National Institute on Deafness and Other Communication Disorders (NIDCD) R01 DC 006472), the Kansas Intellectual and Developmental Disabilities Research Center (National Institute of Child Health and Human Development P30 HD002528), and the Center for Biobehavioral Neurosciences in Communication Disorders (NIDCD P30 DC005803) to M. S. Vitevitch.

References

Albert, R. & Barabási, A. L. [2002] "Statistical mechanics of complex networks," *Rev. Mod. Phys.* **74**, 47–97.

Amaral, L. A. N., Scala, A., Barthélémy, M. & Stanley, H. E. [2000] "Classes of small-world networks," *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152.

Arbesman, S., Strogatz, S. H. & Vitevitch, M. S. [in preparation] *Comparative Analysis of Networks of Phonologically Similar Words in English and Spanish.*

Barabási, A. L. & Albert, R. [1999] "Emergence of scaling in random networks," *Science* **286**, 509–512.

Callaway, D. S., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E. J. & Strogatz, S. H. [2001] "Are randomly grown graphs really random?" *Phys. Rev. E* **64**.

Chan, K. Y. & Vitevitch, M. S. [in press] "Network structure influences speech production," *Cogn. Sci.*

Chan, K. Y. & Vitevitch, M. S. [2009] "The influence of the phonological neighborhood clustering-coefficient on spoken word recognition," *J. Exper. Psychol. Human Percep. Perform.* **35**, 1934–1949.

Clauset, A., Shalizi, C. R. & Newman, M. E. J. [2007] "Power-law distributions in empirical data," arXiv:0706.1062v1 [physics.data-an].

Ferrer-i-Cancho, R. & Ricard, V. S. [2001] "The small world of human language," *Proc. R. Soc. Lond B* **268**, 2261–2265.

Garnham, A., Shillock, R., Brown, G., Mill, A. & Cutler, A. [1981] "Slips of the tongue in the London–Lund corpus of spontaneous conversation," *Linguistics* **19**, 805–817.

Huang, S., Bian, X., Wu, G. & McLemore, C. [1997] *LDC Mandarin Lexicon: Linguistic Data Consortium* (University of Pennsylvania).

Judd, H. P. [1980] *The Hawaiian Language and Hawaiian-English Dictionary: A Complete Grammar* (Hawaiian Service, Incorporated).

Kello, C. T. & Beltz, B. C. [2007] "Scale-free networks in phonological and orthographic wordform lexicons," in *Approaches to Phonological Complexity*, eds. Chitoran, I., Coupé, C., Marsico, E. & Pellegrino, F. (Mouton de Gruyter).

Kleinberg, J. M. [2000] "Navigation in a small world," *Nature* **406**, 845.

Luce, P. A. & Pisoni, D. B. [1998] "Recognizing spoken words: The neighborhood activation model," *Ear and Hearing* **19**, 1–36.

Motter, A. E., de Moura, A. P. S., Lai, Y.-C. & Dasgupta, P. [2002] "Topology of the conceptual network of language," *Phys. Rev. E* **65**, 065102.

Newman, M. E. J. [2001] "The structure of scientific collaboration networks," *Proc. Natl. Acad. Sci. USA* **98**, 404–409.

Newman, M. E. J. [2002] "Assortative mixing in networks," *Phys. Rev. Lett.* **89**, 208701.

Newman, M. E. J. [2003] "The structure and function of complex networks," *SIAM Rev.* **45**, 167–256.

Perea, M., Urkia, M., Davis, C. J., Agirre, A., Laseka, E. & Carreiras, M. [2006] "E-Hitz: A word frequency list and a program for deriving psycholingustic statistics in an agglutinative language (Basque)," *Behav. Res. Meth.* **38**, 610–615.

Sebastián-Gallés, N., Martí-Antonín, M. A., Carreiras-Valinã, M. F. & Cuetos-Vega, F. [2000] *Lexesp. Léxico informatizado del Espanõl.* Edicions de la Universitat de Barcelona.

Steyvers, M. & Tenenbaum, J. [2005] "The large scale structure of semantic networks: Statistical analyses and a model of semantic growth," *Cogn. Sci.* **29**, 41–78.

Storkel, H. L., Armbruster, J. & Hogan, T. P. [2006] "Differentiating phonotactic probability and neighborhood density in adult word learning," *J. Speech Lang. Hearing Res.* **49**, 1175–1192.

Strogatz, S. H. [2001] "Exploring complex networks," *Nature* **410**, 268–276.

Vitevitch, M. S. [2002] "The influence of phonological similarity neighborhoods on speech production," *J. Exper. Psychol.*: *Learn. Mem. Cogn.* **28**, 735–747.

Vitevitch, M. S. & Sommers, M. S. [2003] "The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults," *Mem. Cogn.* **31**, 491–504.

Vitevitch, M. S. & Rodríguez, E. [2005] "Neighborhood density effects in spoken word recognition in Spanish," *J. Multi. Commun. Disor.* **3**, 64–73.

Vitevitch, M. S. & Stamer, M. K. [2006] "The curious case of competition in Spanish speech production," *Lang. Cogn. Process.* **21**, 760–770.

Vitevitch, M. S. [2008] "What can graph theory tell us about word learning and lexical retrieval?" *J. Speech, Lang. Hear. Res.* **51**, 408–422.

Watts, D. J. & Strogatz, S. H. [1998] "Collective dynamics of 'small-world' networks," *Nature* **393**, 440–442.

Watts, D. J., Dodds, P. S. & Newman, M. E. J. [2002] "Identity and search in social networks," *Science* **296**, 1302–1305.

Yarkoni, T., Balota, D. A. & Yap, M. J. [2008] "Moving beyond Coltheart's N: A new measure of orthographic similarity," *Psychonomic Bull. Rev.* **15**, 971–979.