Poznań Studies in Contemporary Linguistics 50(2), 2014, pp. 197–205

doi:10.1515/psicl-2014-0014

NETWORK SCIENCE AS A METHOD OF MEASURING LANGUAGE COMPLEXITY

MICHAEL VITEVITCH
University of Kansas, Lawrence
mvitevit@ku.edu

ABSTRACT

The physical, mathematical, and information sciences have developed a number of ways to measure complexity and complex systems in the social, biological, and physical domains. One way of measuring complex systems that might be useful to language scientists is the set of tools from the interdisciplinary field known as network science. A number of studies that have used the tools of network science to examine various aspects of language and language processing are summarized. It is acknowledged that much work must be done to use the tools of network science to address the debate about the (equal) complexity of languages. However, this work may prove useful to language scientists interested in the (equal) complexity of languages, as well as in other topics about language. Furthermore, the distinct structural characteristics observed in networks of several languages to date may also prove useful to network scientists as they try to understand how certain structural characteristics influence network dynamics in other domains. Language scientists are urged to embrace the techniques of network science to address the question of the complexity of languages.

KEYWORDS: Language complexity; network science; complex network; small world network.

In the debate about the (equal) complexity of languages, a common question that is asked is: how does one actually measure the complexity of a language? Various approaches have been taken to measure the complexity of the phonological, morphological, and syntactic components of language. For example, in terms of phonological complexity one can measure and compare across languages the typical length of words in a language, the size of the phoneme inventory in a language, and the typical syllable structures in a language. One can also compare across languages metrical characteristics (e.g., Fenk-Oczlon and Fenk 2010). However, the measures that one makes in the

phonological system are very different from those measures that one makes in the metrical system, or the morphological system, or the syntactic system, making it difficult to compare across the systems in a like for like way. In what follows, we will explore an alternative approach that may allow language scientists to make measurements in the phonological, morphological, and syntactic components of language that can be compared in a like for like way. Furthermore, these same measures can be performed across languages enabling language scientists to compare the relative complexity of languages.

A number of physical, mathematical, and information sciences have focused on how to measure complexity and complex systems in the social, biological, and physical domains. These fields have developed several ways to measure complexity including Kolmogorov complexity, Lyapunov exponents, and entropy, but it is not clear how well-suited many of these measures of physical or informational complexity are to measuring certain aspects of language. There is, however, one approach that is used to measure complexity in the social, biological, physical, and other domains that may prove useful in examining certain aspects of language complexity, namely, the statistical and mathematical tools developed in the interdisciplinary field known as network science.

In network science (for an introduction, see Newman 2010), nodes (or vertices) are used to represent individual entities, and connections (edges or arcs) are used to represent relationships between entities, forming a web-like structure, or network, of the entire system. In the simplest case, the connections denote bi-directional relationships, in which case the connections are referred to as edges, but connections can also denote uni-directional relationships, in which case the connections are referred to as arcs. For example, a social group could be represented as a network with a node representing each person in the group. If a relationship between people is bi-directional – for example, I call you a friend, and you call me a friend – then edges can be placed between nodes that are "friends" with each other. However, if relationships are uni-directional – for example, I would loan some money to you, but you would not be willing to loan some money to me – then arcs (sometimes called directed links) would be placed between nodes in the social system.

Network science has been used to measure various parts of the language system. For example, Mukherjee et al. (2009), created a network to explore how the phoneme inventories of the languages of the world self-organize, leading to similar patterns across languages. In this network a node corresponded to a consonant found in the languages of the world, and a connection was placed between nodes if those consonants occurred together in a

language. In this case the connection was weighted to indicate the number of languages in which those two nodes co-occurred.

One can also construct a language network such that each node represents an individual word-form. Connections between nodes in this case could indicate that two word-forms are phonologically similar to each other (Vitevitch 2008) or orthographically similar to each other (Kello and Beltz in press). A useful bibliography of research using networks to study various aspects of language can be found at: http://www.lsi.upc.edu/~rferrericancho/linguistic_and_cognitive_networks.html. A quick review of that website indicates that many of the components of language that are typically part of the debate on the (equal) complexity of languages, such as morphology (Liu and Xu 2011) and syntax (Amancio et al. 2012), can be and have been explored with the network science approach. (For discussion of the assumptions that accompany using nodes, edges, and a network as a representational framework in various domains, see Butts 2009.)

A central tenet of network science is that the dynamics of a system are constrained by the structure of the system (as represented by a network; Watts and Strogatz 1998). Networks have been used in a number of domains to examine the dynamics of those complex systems. For example, networks have been used to understand the spread of information or disease in social groups (Kamp et al. 2013).

One can also examine how changes to the network influence the dynamics of a system. For example, Montoya and Solé (2002) created a network of an ecosystem to examine how the extinction of a given species might affect the rest of the ecosystem. In that network, nodes in the network represented the different animal species in the ecosystem, and arcs (also called directed links) were used to represent the predator–prey relationship between animal species (i.e., who eats who). Montoya and Solé observed that the extinction of certain species would have little effect on the larger ecosystem, whereas the extinction of certain other species would be devastating not only to the species directly connected to them, but the loss of those species would have significant repercussion throughout the rest of the ecosystem.

Although the network metaphor is somewhat intuitive, the interdisciplinary field of network science is more than just a metaphor because it offers a number of mathematical and computational tools that can be used to measure complex systems. Measurements can be made at various scales of a complex system, including the micro-level, the macro-level, and the meso-level.

At the micro-level, individual agents in a complex system can be examined. Common measures at the micro-level include degree, clustering coeffi-

cient, and several measures of centrality. Degree is the number of connections that a node has to other nodes. The clustering coefficient measures the extent to which the neighbors of a given node are also neighbors of each other (for a more precise definition see Watts and Strogatz 1998). There are several measures of centrality – degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality – but each attempts to capture in some way the relative importance of a node in a network.

At the macro-level, measurements can be made that describe the over-all structure of the system. One common approach to describing the overall structure of the network is to compute "average" measures (i.e., the mean) across the entire network for degree, shortest path length (i.e., how many connections must one traverse to get from node A to node B), and clustering coefficient in order to classify the overall structure of the system as exhibiting small-world characteristics. In a small-world network one observes in that network: (1) an average path length that is similar to the average path length of a network with the same number of nodes, but with connections placed among the nodes at random, and (2) an average clustering coefficient that is much larger than the average clustering coefficient of a network with the same number of nodes, but with connections placed among the nodes at random.

Another common class of network is known as a scale-free network. In a scale free network one observes that there are many nodes with few connections, and a small number of nodes with many, many connections (such that a frequency distribution of the degree of each nodes follows a power-law with an exponent that falls between 2–3). For a brief review of network science and its application to the cognitive sciences, especially to language, see Baronchelli et al. (2013).

At the meso-level measurements describe the network at scales between the micro- and macro-levels. Typically at the meso-level one uses community detection algorithms to look for sub-groups that may exist within a larger system. For a study examining the community structure found in a network of phonological word forms see Siew (2013). Interestingly, Siew (2013) suggested that the communities found in the network of phonological word forms (i.e., the network examined in Vitevitch 2008) may prevent activation from spreading throughout the entire lexicon during the recognition of spoken words. That is, activation of a target word may easily spread to similar sounding competitors in the same community, but that activation may not spread as easily to words in other communities, thereby restricting the number of potential lexical competitors that must be evaluated during word recognition.

Although network science does not provide a single, global measure of complexity it can be used to measure different systems within a language (e.g., phonology, morphology, syntax), and the micro-, meso-, and macro-levels of those systems in a like for like manner. In addition, one can also use network science to make like for like comparisons across languages. Arbesman et al. (2010) did such a cross-language analysis of the phonological word-form networks of English, Spanish, Mandarin, Hawaiian, and Basque. (In the phonological word-form networks a node corresponded to a word-form, and a connection was placed between nodes that were phonologically similar to each other as defined in Vitevitch 2008.) Although the number of languages examined was small, the languages differed from each other in a number of interesting ways, including the language family that each language belongs to, the typical length of words in each language, the size of the phoneme inventory, canonical syllable structure, and morphological productivity, among other things.

Some of the network measures reported in Arbesman et al. (2010) are reproduced in the present paper in Table 1 for the convenience of the reader. Of the measures presented in Table 1 assortative mixing by degree (Vitevitch et al. 2014), path length (see Iyengar et al. 2012; and Vitevitch et al. in press), and clustering coefficient (Chan and Vitevitch 2009, 2010) have been shown in various psycholinguistic experiments and analyses to influence lexical processes such as word retrieval during spoken word recognition and spoken word production.

In comparing a given network measure in Table 1 across the different languages one might observe that there is some variability even in this limited sample of languages. Consider the average-shortest-path length in the giant component of the five languages. Hawaiian has a value of 5.5 (meaning that when any two nodes in the giant component are selected, it takes, on average, 5.5 links to get from one node to the other), whereas Spanish has a value of 10.3.

To better assess the differences across languages in a given network measure it is imperative to make these measurements in many more languages than the handful of languages sampled in Arbesman et al. (2010) in order to obtain a distribution for a given network measure that could be used for comparison and statistical analysis. Such a distribution would provide necessary information about how numerically different two values need to be in order to be considered statistically different from each other. In analyses such as this, network science could be used to address arguments regarding the absolute complexity of various languages.

Table 1. A selection of the data presented in Arbesman et al. (2010). GC = giant component.

	English	Spanish	Mandarin	Hawaiian	Basque
Network Size (number of words)	19,323	122,066	30,086	2,578	99,321
Giant Component Size (proportion)	6,498 (0.34)	44,833 (0.37)	19,712 (0.66)	1,406 (0.55)	35,173 (0.35)
Mixing by Degree	+0.66	+0.76	+0.65	+0.56	+0.72
Average Shortest-Path-Length (GC)	6.1	10.3	10.1	5.5	10.4
Clustering Coefficient	0.28	0.19	0.38	0.24	0.21

Other work has examined further the implications for psycholinguistic processing of several network measures. Work in my lab has demonstrated that the (local) clustering coefficient influences speech perception (Chan and Vitevitch 2009), speech production (Chan and Vitevitch 2010), certain aspects of short- and long-term memory (Vitevitch et al. 2012), and word-learning (Goldstein and Vitevitch 2014). We have also examined how assortative mixing by degree (Vitevitch et al. 2014), and path length (Vitevitch et al. in press) influence lexical processes such as word retrieval during spoken word recognition and spoken word production. However, there remain a number of other network measures that may also be shown to influence processing in some way, and may prove useful for increasing our understanding of language processing in general. Furthermore, psycholinguistic studies of how various network science measures influence language processing might provide a bridge between the idea of absolute complexity and the idea of relative or functional complexity differences among languages.

Future studies would need to look at a variety of network science measures, how those measures related to psycholinguistic processing, and at the processing consequences for a difference in those measures across languages. For example, does the difference in average-shortest-path length – returning to our observation about Hawaiian and Spanish – mean that it might take longer to retrieve a word in a language with a longer average-shortest-path length than in a language with a shorter average-shortest-path length?

Crucially, the debate about the (equal) complexity of languages tends to focus on issues related to morphology and syntax. Network science measures of various aspects of morphology and syntax would need to be made across languages to be able to contribute to the debate on the (equal) complexity of

languages. Thus, there is some additional work required before the tools of network science can be used to fully address the question of the (equal) complexity of languages.

However, the work that must be done to use the tools of network science to compare different languages may also be useful to language scientists attempting to address questions in addition to the (equal) complexity of languages. Consider the clustering coefficient, which measures the extent to which phonological neighbors of a word are also neighbors with each other. The clustering coefficient ranges from 0 (meaning none of the neighbors of a word are neighbors with each other) to 1 (meaning all of the neighbors of a word are neighbors with all of the other neighbors). However, the average values for the clustering coefficient in Table 1 are distributed around .2. Obtaining these network measures on a larger sample of languages may help language scientists better define the limits of human language, and further distinguish human language from other complex forms of communication.

Making various network measures across a number of languages is, of course, very time-consuming, but such information may also prove useful to network scientists. For example, the values for a given network measure in Table 1 fall within a limited range of the possible values that could be observed for that measure. Consider mixing by degree, which is determined by the Pearson correlation coefficient of the degrees at the end of an edge. Recall that the correlation coefficient ranges from −1 to +1, but the values observed in Table 1 are all positive and close to .5. As noted in Arbesman et al. (2010), the values observed in the language networks are higher than the values that are typically observed in other networks found in the real world. The networks of these languages provide network scientists with examples of systems that exhibit certain distinctive structural characteristics whose unique consequences on processing could be explored via computer simulation (or via psycholinguistic experiments as in Vitevitch et al. 2014).

The question of how to compare in a like for like way across language networks that vary in size and other characteristics may also challenge network scientists to develop new techniques to facilitate such comparisons. These techniques would not only be useful for comparing amongst networks of words, but could also be used to compare amongst networks that represent any kind of system. Some work on the issue of comparing the structure of two or more networks has been done by Faust and Skvoretz (2002), by Robins et al. (2007), and others, but there appears to be much work still to do on this issue. Thus, the benefits of carrying out the additional work of measuring a larger sample of languages (or language networks) might be manifold.

As attested by some of the work listed in the reference section of this brief article, the tools of network science have been used to examine syntax and morphology, making network analyses a potentially useful contribution to the debate about the (equal) complexity of languages. As attested by other work listed in the reference section of this brief article – the attentive reader will notice the prevalence of physical science journals in the reference section – the debate about the (equal) complexity of languages has caught the interest and attention of researchers from a variety of fields, not just Linguistics. As language scientists we should learn how to use the tools of network science as well as the other methods developed by physical, mathematical, and information scientists to measure complexity to contribute to the debate about the (equal) complexity of languages in a linguistically-informed way. We should not cede this debate to researchers in other fields.

REFERENCES

Amancio, D.R., M.G.V. Nunes, O.N. Oliveira Jr and L. da F. Costa. 2012. "Extractive summarization using complex networks and syntactic dependency". *Physica A,* 391. 1855–1864.

Arbesman, S., S.H. Strogatz and M.S. Vitevitch. 2010. "The structure of phonological networks across multiple languages". *International Journal of Bifurcation and Chaos* 20. 679–685.

Baronchelli, A., R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chater and M.H. Christiansen. 2013. "Networks in cognitive science". *Trends in Cognitive Sciences* 17. 348–360.

Butts, C. 2009. "Revisiting the foundations of network analysis". *Science* 325. 414–416.

Chan, K.Y. and M.S. Vitevitch. 2009. "The Influence of the phonological neighborhood clustering-coefficient on spoken word recognition". *Journal of Experimental Psychology: Human Perception & Performance* 35. 1934–1949.

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

Faust, K. and J. Skvoretz. 2002. "Comparing networks across space and time, size and species". *Sociological Methodology* 32. 267–299.

Fenk-Oczlon, G. and A. Fenk. 2010. "Measuring basic tempo across languages and some implications for speech rhythm". *Proceedings of the 11th Annual Conference of the International Speech Communication Association* (INTERSPEECH 2010), Makuhari, Japan. 1537–1540.

Goldstein, R. and M.S. Vitevitch. 2014. "The influence of clustering coefficient on word-learning". Unpublished manuscript submitted for publication.

Iyengar, S.R.S., C.E.V. Madhavan, K.A. Zweig and A. Natarajan. 2012. "Understanding human navigation using network analysis". *Topics in Cognitive Science* 4. 121–134.

Kamp, C., M. Moslonka-Lefebvre, S. Alizon. 2013. "Epidemic spread on weighted networks". *PLoS Computational Biology* 9(12): e1003352.

Kello, C.T. and B.C. Beltz. In press. "Scale-free networks in phonological and orthographic wordform lexicons". In: Chitoran, I., C. Coupé, E. Marsico and F. Pellegrino (eds.), *Approaches to phonological complexity.* Berlin: Mouton de Gruyter.

Liu, H. and C. Xu. 2011. "Can syntactic networks indicate morphological complexity of a language?" *Europhysics Letters* 93. 28005.

Montoya, J.M. and R.V. Solé. 2002. "Small world patterns in food webs". *Journal of Theoretical Biology* 214. 405–412.

Mukherjee, A., M. Choudhury, A. Basu and N. Ganguly. 2009. "Self-organization of the sound inventories: analysis and synthesis of the occurrence and co-occurrence networks of consonants". *Journal of Quantitative Linguistics* 16. 157–184.

Newman, M.E.J. 2010. *Network science: An introduction*. Oxford: Oxford University Press.

Robins, G., P. Pattison, Y. Kalish and D. Lusher. 2007. "An introduction to exponential random graph p* models for social networks". *Social Networks* 29. 173–191.

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

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

Vitevitch, M.S., K.Y. Chan and R. Goldstein. 2014. "Insights into failed lexical retrieval from network science". *Cognitive Psychology* 68. 1–32.

Vitevitch, M.S., K.Y. Chan and S. Roodenrys. 2012. "Complex network structure influences processing in long-term and short-term memory". *Journal of Memory & Language* 67. 30–44.

Vitevitch, M.S., R. Goldstein and E. Johnson. In press. "Path-length and the misperception of speech: Insights from Network Science and Psycholinguistics". In: Mehler, A., P. Blanchard, B. Job and S. Banish (eds.), *Towards a theoretical framework for analyzing complex linguistic networks.* (*Understanding Complex Systems* series.) Heidelberg: Springer.

Watts, D.J. and S.H. Strogatz. 1998. "Collective dynamics of 'small-world' networks". *Nature* 393. 409–410.

Address correspondence to:
Michael Vitevitch
University of Kansas
1415 Jayhawk Blvd.
Lawrence, Kansas 66045
United States
mvitevit@ku.edu