# 5

# SOCIAL STRUCTURE RELATES TO LINGUISTIC INFORMATION DENSITY

*David W. Vinson*
*Rick Dale*

## Abstract

Some recent theories of language see it as a complex and highly adaptive system, adjusting to factors at various time scales. For example, at a longer time scale, language may adapt to certain social or demographic variables of a linguistic community. At a shorter time scale, patterns of language use may be adjusted by social structures in real time. Until recently, datasets large enough to test how socio-cultural properties—spanning vast amounts of time and space—influence language change have been difficult to obtain. The emergence of digital computing and storage have brought about an unprecedented ability to collect and classify massive amounts of data. By harnessing the power of Big Data we can explore what socio-cultural properties influence language use. This chapter explores how social-network structures, in general, contribute to differences in language use. We analyzed over one million online business reviews using network analyses and information theory to quantify social connectivity and language use. Results indicate that perhaps a surprising proportion of variance in individual language use can be accounted for by subtle differences in social-network structures, even after fairly aggressive covariates have been added to regression models. The benefits of utilizing Big Data as a tool for testing classic theories in cognitive science and as a method toward guiding future research are discussed.

## Introduction

Language is a complex behavioral repertoire in a cognitively advanced species. The sounds, words, and syntactic patterns of language vary quite widely across human groups, who have developed different linguistic patterns over a long stretch of time and physical separation (Sapir, 1921). Explanations for this variation derive from two very different traditions. In the first, many language scientists have sought to abstract away from this observed variability to discern core characteristics of

language, which are universal and perhaps genetically fixed across people (from Chomsky, 1957 to Hauser, Chomsky, & Fitch, 2001). The second tradition sees variability as the mark of an intrinsically adaptive system. For example, Beckner et al. (2009) argue that language should be treated as responsive to socio-cultural change in real time. Instead of abstracting away apparently superficial variability in languages, this variability may be an echo of pervasive adaptation, from subtle modulation of real-time language use, to substantial linguistic change over longer stretches of time. This second tradition places language in the broader sphere of human behavior and cultural products in a time when environmental constraints have well known effects on many aspects of human behavior (see Triandis, 1994 for review).[1]

Given these explanatory tendencies, theorists of language can have starkly divergent ideas of it. An important next step in theoretical mitigation will be new tools and broad data samples so that, perhaps at last, analyses can match theory in extent and significance. Before the arrival of modern information technologies, a sufficient linguistic corpus would have taken years, if not an entire lifetime, to acquire. Indeed, some projects on the topic of linguistic diversity have this property of impressive timescale and rigor. Some examples include the Philadelphia Neighborhood Corpus, compiled by William Labov in the early 1970s, the *Ethnologue,* first compiled by Richard Pittman dating back to the early 1950s and the World Atlas of Language Structures (WALS), a collection of data and research from 55 authors on language structures available online, only six years ago (in 2008). Digitally stored language, and to a great extent accessible for analysis, has begun to exceed several exabytes, generated everyday online (Kudyba & Kwatinetz, 2014).[2] One way this profound new capability can be harnessed is by recasting current theoretical foundations, generalized from earlier small-scale laboratory studies, into a Big Data framework.

If language is pervasively adaptive, and is thus shaped by socio-cultural constraints, then this influence must be acting somehow in the present day, in real-time language use. Broader linguistic diversity and its socio-cultural factors reflect a culmination of many smaller, local changes in the incremental choices of language users. These local changes would likely be quite small, and not easily discerned by simple observation, and certainly not without massive amounts of data. In this chapter, we use a large source of language data, Yelp, Inc. business reviews, to test whether social-network structures relate in systematic ways to the language used in these reviews. We frame social-network variables in terms of well-known network measures, such as centrality and transitivity (Bullmore & Sporns, 2009), and relate these measures to language measures derived from information theory, such as information density and uniformity (Aylett, 1999; Jaeger, 2010; Levy & Jaeger, 2007). In general, we find subtle but detectable relationships between these two groups of variables. In what follows, we first motivate the broad theoretical framing of our Big Data question: What shapes

linguistic diversity and language change in the broad historical context? Following this we describe information theory and its use in quantifying language use. Then, we explain how social structure may influence language structure. We consider this a first step in understanding how theories in cognitive and computational social science can be used to harness the power of Big Data in important and meaningful ways (see Griffiths, 2015).

## *What Shapes Language?*

As described above, language can be cast as a highly adaptive behavioral property. If so, we would probably look to social, cultural or even ecological aspects of the environment to understand how it changes (Nettle, 1998; Nichols, 1992; Trudgill, 1989, 2011). Many studies, most over the past decade, suggest languages are dynamically constrained by a diverse range of environmental factors. Differences in the spread and density of language use (Lupyan & Dale, 2010), the ability of its users (Bentz et al., submitted; Christiansen & Chater, 2008; Dale & Lupyan, 2012; Ramscar, 2013; Wray & Grace, 2007) and its physical environment (Nettle, 1998; Everett, 2013) impact how a language is shaped online (Labov, 1972a, 1972b) and over time (Nowak, Komarova & Niyogi, 2002). These factors determine whether certain aspects of a language will persist or die (Abrams & Strogatz, 2003), simplify or remain complex (Lieberman, Michel, Jackson, Tang, & Nowak, 2007). Language change is also rapid, accelerating at a rate closer to that of the spread of agriculture (Gray & Atkinson, 2003; cf. Chater, Reali & Christiansen, 2009) than genetics. Using data recently made available from WALS and a recent version of the *Ethnologue* (Gordon, 2005), Lupyan and Dale (2010) found larger populations of speakers, spread over a wider geographical space, use less inflection and more lexical devices. This difference may be due to differences in communicating within smaller, "esoteric" niches and larger, "exoteric" niches (also see Wray & Grace, 2007), such as the ability of its speakers (Bentz & Winter, 2012; Lupyan & Dale, 2010; Dale & Lupyan, 2012) or one's exposure to a growing vocabulary (Reali, Chater & Christiansen, 2014).

Further evidence of socio-cultural effects may be present in real-time language usage. This is a goal of the current chapter—can we detect these *population-level* effects in a large database of language use? Before describing our study, we describe two key motivations of our proposed analyses: The useful application of (1) information theory in quantifying language use and (2) network theory in quantifying social structure.

## *Information and Adaptation*

Information theory (Shannon, 1948) defines the *second-order information* of a word as the negative log probability of a word occurring after some other word:

$$I(w_i) = -\log_2 p(w_i|w_{i-1})$$

The theory of uniform information density (UID; Levy & Jaeger 2007; Jaeger, 2010) states that speakers will aim to present the highest amount of information across a message at a uniform rate, so as to efficiently communicate the most content without violating a comprehender's channel capacity. Support for this theory comes from Aylett (1999), in an early expression of this account, who found that speech is slower when a message is informationally dense and Jaeger (2010), who found information-dense messages are more susceptible to optional word injections, diluting its overall density over time. Indeed, even word length may be adapted for its informational qualities (Piantadosi, Tily & Gibson, 2011).

In a recent paper, we investigated how a simple contextual influence, the intended valence of a message, influences information density and uniformity. While it is obvious that positive and negative emotions influence what words individuals use (see Vinson & Dale, 2014a for review), it is less obvious that the probability structure of language use is also influenced by one's intentions. Using a corpus of over two-hundred thousand online customer business reviews from Yelp, Inc., findings suggest that the information density of a message increases as the valence of that message becomes more extreme (positive or negative). It also becomes ~~more uniform (less variable)~~ as message valence becomes more positive (Vinson & Dale, 2014b). The results are commensurate with theories that suggest language use adapts to a variety of socio-cultural factors in real time.

In this chapter, we look to information-theoretic measures of these kinds to quantify aspects of language use, with the expectation that they will also relate in interesting ways to social structure.

## *Social Network Structure*

Another key motivation of our proposed analyses involves the use of network theory to quantify the intricate structural properties that connect a community of speakers (Christakis & Fowler, 2009; Lazer et al., 2009). Understanding how specific socio-cultural properties influence language can provide insight into the behavior of the language user herself (Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013). For instance, Kramer, Guilliory and Hancock (2014) having analyzed over six hundred thousand Facebook users, reported when a user's newsfeed was manipulated to show only those posts that were either positive or negative, a reader's own posts aligned with the emotional valence of their friends' messages. Understanding what a language *looks like* when under certain socio-cultural pressures can provide valuable insight into what societal pressures that help shape a language. Indeed, global changes to one's socio-cultural context, such as changes in the classification of severity of crime and punishment over time, are marked by linguistic change (Klingenstien, Hitchcock & DeDeo, 2014) while differences in the distance between socio-cultural niches are marked by differences in language use (Vilhena et al., 2014).

### *Current Study*

In the current study, we utilize the Yelp database as an arena to test how population-level differences might relate to language use. While previous work suggests online business reviews may provide insight into the psychological states of its individual reviewers (Jurafsky, Chahuneau, Routledge, & Smith, 2014), we expect that structural differences in one's social community as a whole, where language is crucial to conveying ideas, will affect language use. We focus on how a language user's social niche influences the amount and rate of information transferred across a message. Agent-based simulations (Chater et al., 2006; Dale & Luypan, 2012; Reali et al., 2014) and recent studies on the influences of interaction in social networks (Bond et al., 2012; Choi, Blumen, congleton, & Rajaram, 2014) indicate that the structure of language use may be influenced by structural aspects of a language user's social interactions. From an exploratory standpoint, we aim to determine if one's social-network structure predicts the probability structure of language use.

## Method

### *Corpus*

We used the Yelp Challenge Dataset (www.yelp.com/dataset.challenge), which, at the time of this analysis, contained reviews from businesses in Phoenix, Las Vegas, Madison, and Edinburgh. This includes 1,125,458 reviews from 252,898 users who reviewed businesses in these cities. The field entries for reviews included almost all the information that is supplied on the Yelp website itself, including the content of the review, whether the review was useful or funny, the star rating that was conferred upon the business, and so on. It omits a user's public username, but includes an array of other useful information, in particular a list of user ID codes that point to friends of a given user. Yelp users are free to ask any other Yelp user to be their friend. Friendship connections are driven by users' mutual agreement to become friends. These user ID codes allow us to iteratively build social networks by randomly choosing a user, and expanding the network by connecting friends and friends of friends, which we further detail below.

### *Linguistic Measures*

The first and simplest measure we explore in our analysis is the *number of words* in a given review, its word length. This surface measure is used as a basic but important covariate for regression analyses. Word length will define the bin count for entropy and other information analyses, and so directly impacts these measures.

The second measure we use is the *reviewer-internal entropy* (RI-Ent) of a reviewer's word use. This marks the discrete Shannon entropy of a reviewer's overall

word distribution. If reviewers use many different words, entropy would be high. If a reviewer reuses a smaller subset of words, the entropy of word distribution would be low, as this would represent a less uniform distribution over word types.

A third measure is the average information encoded in the reviewer's word use, which we'll call *average unigram information* (AUI). Information, as described above, is a measure of the number of bits a word encodes given its frequency in the overall corpus. Reviews with higher information use less frequent words, thus offering more specific and less common verbiage in describing a business.

A fourth measure is one more commonly used in studies of informational structure of language, which we'll call the *average conditional information* (ACI). This is a bit-based measure of a word based on its probability conditioned on the prior word in the text. In other words, it is a measure of the bits encoded in a given bigram of the text. We compute the average bits across bigrams of a review, which reflect the uniqueness in word combinations.[3]

Finally, we extract two crude measures of information variability by calculating the standard deviation over AUI and ACI, which we call *unigram informational variability* (UIV) and *conditional informational variability* (CIV), respectively. Both measures are a reflection of how stable the distribution is over a reviewer's average unigram and bigram bit values. These measures relate directly to uniform information density (see Levy & Jaeger, 2007; Jaeger 2010). A very uniform distribution of information is represented by a stable mean and lower UIV/CIV;

**TABLE 5.1** Summary of information theoretic measures quantifying language in reviews.

| *Measure* | *Description* | *Definition* |
|---|---|---|
| RI-Ent | Reviewer-internal entropy | $RI-ENT_j = -\sum_{i=1}^{N} p(w_i\|R_j) log_2 p(w_i\|R_j)$ |
| AUI | Average unigram information | $AUI_j = \frac{1}{N} \sum_{i=1}^{N} log_2 p(w_i)$ |
| ACI | Average conditional information | $ACI_j = -\frac{1}{N-1} \sum_{i=1}^{N} log_2 p(w_i\|w_{i-1})$ |
| UIV | Unigram informational variability | $UIV_j = \sigma(UI_j)$ |
| CIV | Conditional informational variability | $CIV_j = \sigma(\mathbf{CI}_j)$ |

$N$ = number of words in a review; $p(w)$ = probability of word $w$; $w_i$ = $i^{th}$ word of a review; $\mathbf{UI}_j$ = set of unigram information scores for each word of a given review; $\mathbf{CI}_j$ = set of conditional information scores for each word of a given review.

a review with unigram or bigram combinations that span a wide range of informativeness induces a wider range of bit values, and thus a higher UIV/CIV (less uniform density). A summary of these measures appears in Table 5.1. Punctuation, stop words and spacing were removed using the tm package in R before information-theoretic measures were obtained.[4]

## *Social Networks*

One benefit of the Big Data approach we take in this chapter is that we can pose our questions about language and social structure using the targeted unit of analysis of social networks themselves. In other words, we can *sample networks* from the Yelp dataset directly, with each network exhibiting network scores that reflect a certain aspect of local social structure. We can then explore relationships between the information-theoretic measures and these social-network scores.

We sampled 962 unique social networks from the Yelp dataset, which amounted to approximately 38,000 unique users and 450,000 unique reviews. Users represent nodes in social networks and were selected using a selection and connection algorithm also shown in Figure 5.1. We start by choosing a random user who has

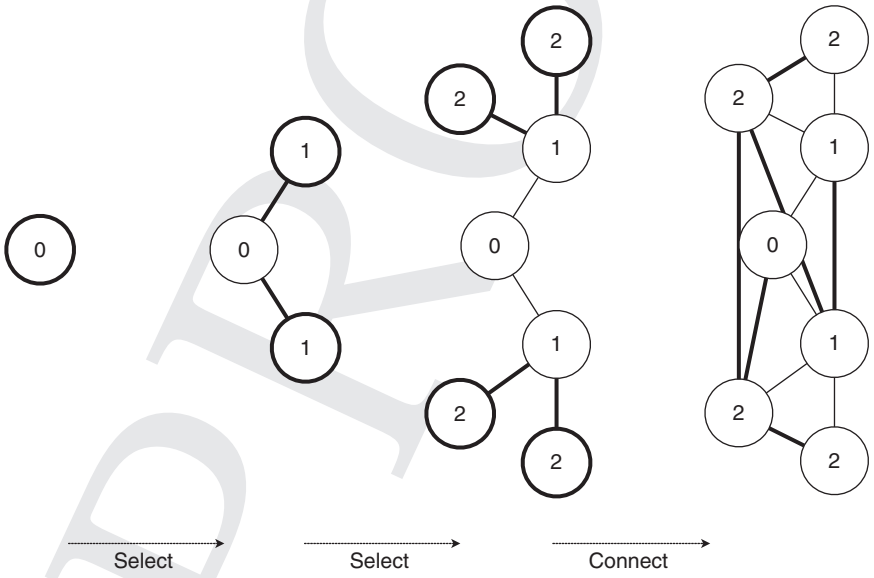


**FIGURE 5.1** "0"-degree was chosen at random among those with between 11 and 20 friends. We then connected these individuals. Then, from ten randomly chosen friends of the seed node, we chose up to ten friends of friends and connected them. Following this, we interconnected the whole set. Note that this is a simplified visualization of the process, as friends and friends of friends were chosen up to a count of ten, which produces much larger networks (visualized in examples below).

between 11 and 20 friends in the dataset (we chose this range to obtain networks which were not too small or too large as to be computationally cumbersome). After we chose that user, we connected all his or her friends and then expanded the social network by one degree; randomly selecting ten of his or her friends and connecting up to ten of his or her friend's friends to the network. We then interconnected all users in this set (shown as the first-degree nodes and connections in Figure 5.1). We conducted this same process of finding friends of these first-degree nodes, and then interconnected *those* new nodes of the second degree. In order to make sure networks did not become too large, we randomly sampled up to ten friends of each node only. Fifty percent of all networks fell between 89 and 108 reviewers in size, and the resulting nets reveal a relatively normal distribution of network metrics described in the next section.

### *Network Measures*

A variety of different network measures were used to quantify the structure of each network. We consider two different categories of network structures: simple and complex. A summary of all seven (two simple, five complex) network measures appears in Table 5.2. We used two simple network measures: The number of reviewers in a network, or *nodes,* and the number of friendship connections between reviewers, or *edges.*

We considered five complex network measures. The first measure, network *degree*, is the ratio of edges to nodes. This provides a measure of connectivity across a network. High-degree networks have a higher edge-to-node ratio than lower degree networks.

The second measure, network *transitivity*, determines the probability that two adjacent nodes are themselves connected (sometimes termed the "clustering coefficient"). Groups of three nodes, or triples, can either be closed (e.g. fully connected) or open (e.g. two of the three nodes are not connected). The ratio of closed triples to total triples provides a measure of the probability that adjacent nodes are themselves connected. High transitivity occurs when the ratio of closed-to-open triples is close to one.

A third measure, network *betweenness,* determines the average number of shortest paths in a network that pass through some other node. The shortest path of two nodes is the one that connects both nodes with the fewest edges. A pair of nodes can have many shortest paths. A node's betweenness value is the sum of the ratio of a pair of node's shortest paths that pass through the node, over the total number of shortest paths in the network. We determined network betweenness by taking the average node betweenness for all nodes in the network. Effectively, this provides a measure of network efficiency. The higher a network's betweenness, the faster some bit of new information can travel throughout the network.

**TABLE 5.2** Summary of the measures quantifying a network's structure.

| *Measure* | *Definition* | *Discription* |
|---|---|---|
| Nodes | *Nodes* | Number of individuals in the network |
| Edges | *Edges* | Number of node to node connections; vertices in a network |
| Degrees | $\frac{Edges}{Nodes}$ | The ratio of connections to nodes in a network |
| Transitivity/Clustering Coefficient | $\frac{N\ closed\ Triples}{N\ triples}$ | The average number of completely connected triples given the total number of triples in a network. |
| Betweenness | $\frac{\sum\left\{\sum_{s\neq t\neq v}\frac{SP_{st}(V)}{SP_{sp}}\right\}}{N}$ | $SP_{st}$ is the number of total shortest paths from node $s$ to node $t$. $SP_{st}(V)$ is the number of shortest paths from $s$ to $t$ that pass through node $V$. The sum for all shortest paths for all nodes determines the betweenness of node $V$. We take the average betweenness of each node for all nodes $N$ in a network. |
| Centrality | $C_z = \frac{\sum_{i=1}^{N}\lvert C_x(n*) - C_x(n_i)\rvert}{Max\sum_{i=1}^{N}\lvert C_x(n*) - C_z(n_i)}$ | $C_x$ is the graph level centrality defined as the sum of the absolute difference between the observed maximum central node $C_x(n^*)$ and all other node centrality measures $C_x(n_i)$ over the theoretical maximum centrality of a network with the same number of nodes. As, this is a measure of the maximum possible centrality and actual centrality, graph level centrality will always fall between 0 (low centrality) and 1 (high centrality). |

Continued

**TABLE 5.2** (cont).

| *Measure* | *Definition* | *Discription* |
|---|---|---|
| Scale Free | $f(x) = x^{-\sigma}$ | $\alpha$ is the exponent characterizing the power law fit predicted by the degree distribution □. $\alpha$ is always greater than 1 and typically falls within the range of $2 < \alpha < 3$, but not always. |

A fourth measure stems from node *centrality* which determines the number of connections a single node has with other nodes. Centrality can also be determined for the whole network, known as *graph centrality*. Graph centrality is the ratio of the sum of the absolute value of the centrality of each node, over the maximum possible centrality of each node (Freeman, 1979). Node centrality is greatest when a single node is connected to all other nodes, whereas graph centrality is greatest when all nodes are connected to all other nodes. Information is thought to travel faster in high-centrality networks. Here we use graph centrality only. From this point on we will refer to to *graph centrality* simply as *centrality.* Network *betweenness* and network *centrality* share common theoretical assumptions, but quantify different structural properties of a network.

Our fifth and final measure determines whether the connections between nodes in a network share connectivity at both local and global scales. Scale free networks display connectivity at all scales, local and global, simultaneously (Dodds, Watts, & Sabel, 2003). A network is *scale free* when its degree distribution (i.e., the number of edge connections per each node) fits a power law distribution. Networks that are less scale free are typically dominated by either a local (a tightly connected set of nodes) or global connectivity (randomly connected nodes). Networks that exemplify differences in complex structures are presented in Figure 5.2.

## *Additional Measures*

Individual reviews were not quantified independently. Instead, all reviews from a single individual were concatenated into one document. This allowed for information-theoretic measures to be performed over a single user's total set of reviews. The average information of a network was then computed by taking the average across all individuals (nodes) in the network. Such an analysis affords testing how the structure of an individual's social network impacts that individual's

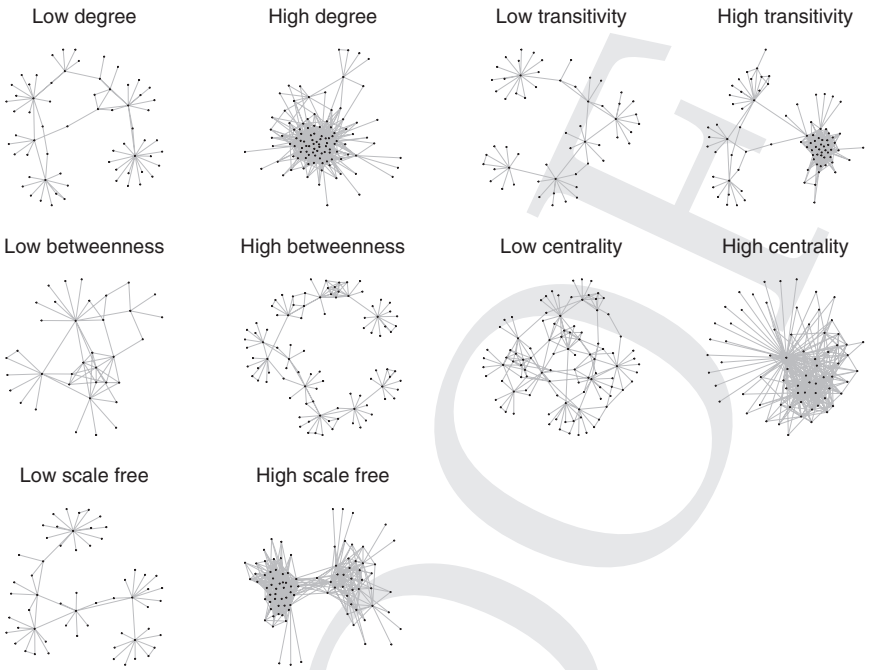


**FIGURE 5.2** Example Yelp networks with high/low structural properties.

overall language use. However, due to the nature of how our information-theoretic measures were determined, individuals who wrote well over one hundred reviews were treated the same as those who wrote merely one. This introduces a possible bias since information measures are typically underestimated when using non-infinite sample sizes (as in the case of our information measures). While we control for certain measures such as the average reviewer's total review length and network size, additional biases may occur due to the nature of how each measure was determined (e.g. averaging across reviewers with unequal length reviews). To address these concerns two additional measures, (1) a *gini coefficient* and (2) a random review baseline—to assess the reliability of our analyses—were used. They are described below.

*Gini Coefficient*. The Gini coefficient (range = [0,1]) was originally developed to assess the distribution of wealth across a nation. As the coefficient approaches zero, wealth is thought to approach greater equality. As the coefficient approaches one, more of the nation's wealth is thought to be shared among only a handful of its residents. We use the Gini coefficient to assess the distribution of reviews across a network. Since each node's reviews were concatenated, given only one value for each information-theoretic measure, certain reviewer's measures will be more representative of the linguistic distributions. The Gini coefficient provides

an additional test as to whether a network's *average information* is influenced by the network's distribution of reviews.

*Random Review Baseline.* A random network baseline provides a baseline value to compare coefficient values from true networks. Baseline information measures were computed by randomly sampling (without replacement) the same number of reviews written by each reviewer. For example, if a reviewer writes five reviews, then five random reviews are selected to take their place. These five reviews are then deleted from the pool of total reviews used throughout all networks. This ensures that the exact same reviews are used in baseline reviews. We did not go so far as to scramble tokens *within* reviews. While this would provide a sufficient baseline, obtaining the true information-theoretic measures of each review, without token substitution, provides a more conservative measure. The random review baseline was used to compare all significant true network effects as an additional measure of reliability.

## *Broad Expectations and Some Predictions*

This chapter presents a broad thesis, and we test it using the Yelp data: Social-network structure will relate in interesting ways to patterns of language use. Although this is a broad expectation, it is not a specific prediction, and we do not wish to take a strong stance on specific hypotheses here, giving the reader the impression that we had conceived, in advance, the diverse pattern of results that we present below. Instead, Big Data is providing rich territory for new exploration. The benefit of a Big Data approach is to identify interesting and promising new patterns or relationships and, we hope, encourage further exploration. As we show below, even after controlling for a variety of collinearities among these measures, the broad thesis holds. Regression models strongly relate social-network and information-theoretic measures. Some of the models show proportion variance accounted for at above 50 percent. Despite this broad exploratory strategy, a number of potential predictions naturally pop out of existing discussion of information transmission in networks. We consider three of these here before moving into the results.

One possibility is that more scale-free networks enhance information transmission, and thus would concomitantly increase channel capacity of nodes in the network. One might suppose that in the Yelp context, a more local scale-free structure might have a similar effect: An efficient spread of information may be indicated by a wide diversity of information measures, as expressed in the UIV and CIV measures.

A second prediction—not mutually exclusive from the first—draws from the work on conceptual entrainment and imitation in psycholinguistics, is that densely connected nets may induce the least AUI. If nodes in a tightly connected network infect each other with common vocabulary, then this would reduce the local

information content of these words, rendering them less unique and thus show the lowest entropy, AUI, and so on. One may expect something similar for transitivity, which would reflect the intensity of local mutual interconnections (closed triples).

However, the reverse is also possible. If language users are more densely connected it may be more likely that they have established a richer common ground overall. If so, language use may contain more information-dense words specific to a shared context (Doyle & Frank, submitted). A fourth prediction is that more network connectivity over a smaller group (higher-network degree) may afford more complex language use, and so lead to higher AUI and ACI.

A final prediction comes from the use of information theory to measure the rate of information transmission. When a speaker's message is more information-dense, it is more likely that it will also be more uniform. Previous findings show speakers increase their speech rate when presenting low information-dense messages, but slow their speech rate for information-dense messages (Pellegrino, Coupé, & Marisco, 2011). It may be that any social structure that leads to increases in information density simultaneously decreases information variability.

## Results

### *Simple Measures*

The confidence intervals (99.9 percent CI) of five multiple regression models, where nodes, edges and the Gini coefficient were used to predict each information-theoretic measure, are presented in Table 5.3. We set a conservative criterion for significance ($p < 0.001$) for all analyses. Only those analyses that were significant are presented. Crucially, all significant effects of independent variables were individually compared to their effects on the random review baseline. To do this, we treated the random review baseline and true network reviews as two levels of the same variable: "true_baseline". Using linear regression we added an interaction term between independent network variables and the true_baseline variable. A significant interaction is demarcated by "†" in Tables 5.3 and 5.4. The effects of these network variables on information-theoretic measures are significantly different in true networks compared to baseline networks. This helps to ensure that our findings are not simply an artifact of our methodology.

All variables were standardized (scaled and shifted to have $M = 0$ and $SD = 1$). Additionally, the number of words (length) was log transformed due to a heavy-tailed distribution. All other variables were normally distributed. Because length correlates with all information-theoretic measures and UIV and CIV correlate with the mean of AUI and ACI, respectively (due to the presence of a true zero), the mean of each information measure was first predicted by length while UIV and CIV were also predicted by AUI and ACI. The residual variability of these linear regression models was then predicted by nodes, edges and the Gini coefficient. The purpose of residualization is to further ensure that observed

**TABLE 5.3** Lexical measures as predicted by *nodes, edges and Gini coefficient*.

| | *Nodes* | *Edges* | *Gini coef* | *F-statistic* |
|---|---|---|---|---|
| *Length* | *n.s.* | (0.09, 0.27) | (0.30, 0.43) | $F(3, 958) = 125$<br>$R^2 = 0.28$, $R^2_{adj} = 0.28$ |
| $RI - Ent_{residual}$ | *n.s.* | $(0.10, 0.15)^{\dagger}$ | $(-0.16, -0.12)^{\dagger}$ | $F(3, 958) = 446.6$<br>$R^2 = 0.58$, $R^2_{adj} = 0.58$ |
| $AUI_{residual}$ | $(-0.05, -0.01)^{\dagger}$ | $(0.10, 0.12)^{\dagger}$ | $(-0.12, -0.09)^{\dagger}$ | $F(3, 958) = 391$<br>$R^2 = 0.55$, $R^2_{adj} = 0.55$ |
| $ACI_{residual}$ | *n.s.* | $(0.07, 0.12)^{\dagger}$ | $(-0.04, -0.01)$ | $F(3, 958) = 154.1$<br>$R^2 = 0.33$, $R^2_{adj} = 0.32$ |
| $UIV_{residual}$ | $(-0.01, -0.001)$ | $(0.001, 0.01)^{\dagger}$ | (004, 0.01) | $F(3, 958) = 39.58$<br>$R^2 = 0.11$, $R^2_{adj} = 0.11$ |
| $CIV_{residual}$ | n.s. | $(-0.003, 0)$ | (0.002, 0.004) | $F(3, 958) = 27.99$<br>$R^2 = 0.08$, $R^2_{adj} = 0.08$ |

Only the Mean and 99.9 percent Confidence Intervals for each IV with $p < 0.001$ are presented. The "†" symbol denotes all network effects that were significantly different from baseline network effects ($p < .001$). Multiple linear regressions were performed in R: lm(DV∼Nodes+Edges+Gini).

interactions are not due to trivial collinearity between simpler variables (length, nodes) and ones that may be subtler and more interesting (CIV, centrality, etc.).[5]

The number of nodes provides a crude measure of network size, edges, network density and the Gini coefficient (the distribution of reviews across the network). Importantly, no correlation exists between the Gini coefficient with either edges or nodes. And, although a strong correlation exists between nodes and edges ($r$ = 0.67, $t$ (960) = 28.23, $p < 0.0001$), in only two instances, AUI and UIV, did nodes account for some portion of variance. As nodes increased, average unigram information, along with average unigram variability, decreased. However, only the relationship between nodes and average unigram information was significantly different between the true network and the random review baseline. In all cases, save conditional information variability (CIV), a significant proportion of variance in information measures was accounted for by edges, and in all but length and CIV, the relationship between information measures and edges was significantly different between the true network and the random review baseline (Figure 5.3(a) presents an example interaction plot between ACI, edges and true_baseline measures). Finally, the Gini coefficient accounted for a significant portion of variance for all information measures, but only for RI-Ent and AUI did it have a significantly different relationship between the true network and the random review baseline. One explanation may be that more unique language use may naturally occur when more individuals contribute more evenly to the conversation. Another possibility is that networks with less even review distributions are more likely to

**TABLE 5.4** Information-theoretic measures predicted by complex network measures.

| | *Degree* | *Transitivity* | *Betweenness* | *Centrality* | *Scale Free:* $\alpha$ | *F-statistic* |
|---|---|---|---|---|---|---|
| *Length* | *n.s.* | (0.20, 0.42) | (-0.28, -0.08) | (0.01, 0.16) | *n.s.* | $F(5, 956) = 33.3$<br>$R^2 = 0.15, R^2_{adj} = 0.14$ |
| $RI\text{-}Ent_{residual}$ | (0.04, 0.13)† | (-0.11, -0.03) | *n.s.* | *n.s.* | *n.s.* | $F(5, 956) = 21.47$<br>$R^2 = 0.10, R^2_{adj} = 0.10$ |
| $AUI_{residual}$ | (0.02, 0.09)† | *n.s.* | *n.s.* | (0.004, 0.05)† | *n.s.* | $F(5, 956) = 10.36$<br>$R^2 = 0.05, R^2_{adj} = 0.05$ |
| $ACI_{residual}$ | (0.01, 0.07)† | (-0.07, -0.01)† | *n.s.* | *n.s.* | *n.s.* | $F(5, 956) = 11.38$<br>$R^2 = 0.06, R^2_{adj} = 0.05$ |
| $UIV_{residual}$ | *n.s.* | *n.s.* | *n.s.* | *n.s.* | *n.s.* | *n.s.* |
| $CIV_{residual}$ | *n.s.* | *n.s.* | *n.s.* | *n.s.* | *n.s.* | *n.s.* |

Only the mean and 99.9 percent confidence intervals for each IV with $p < 0.001$ are presented. All reported values were significant ($p < 0.001$). The "†" symbol denotes all network effects significantly different from baseline network effects.

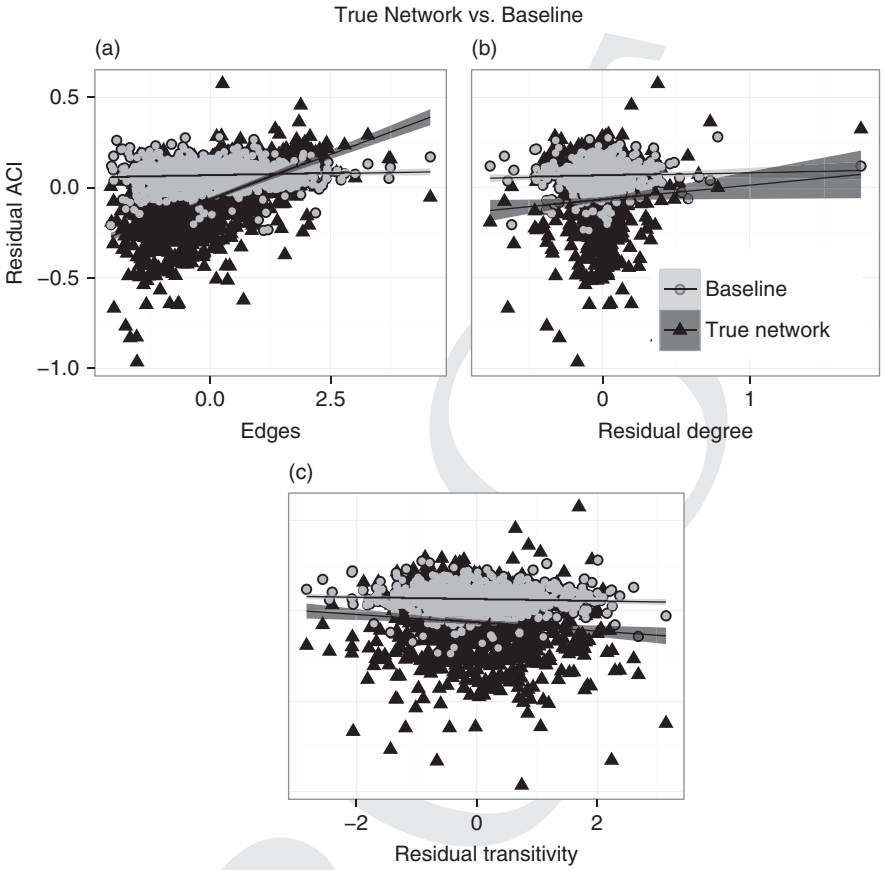


**FIGURE 5.3** Network measures for true and baseline networks by across conditional information-density. All four plots show significant interactions for variables in True networks compared to baseline networks. Linear regression models with interaction terms were used in R:lm(DV IV+true_baseline+IV*true_baseline).

have more reviews, suggesting that a larger number of reviewer's language use is more representative of the overall linguistic distribution of reviews. A simple linear regression analysis reveals the Gini coefficient accounts for a significant portion of variance in the total number of reviews in each network ($R^2_{adj} = 0.21$, $F[1,960] = 249.6$, $p < 0.001$, 99.9 percent CI [0.25, 0.38]), increasing as the number of reviews increases.

We interpret these results cautiously considering it is a first step toward understanding what aspects of a network relate to language use. The results suggest changes in population size and connectivity occur alongside changes in the structure of language use. Speculatively, the individual language user may be influenced by the size and connectivity of her network. When the size of his or

her network increases, the words he or she uses may be more frequent. However when connectivity increases, the words he or she uses may be of low frequency, and therefore more information dense. This supports current work that shows how a shared common ground may lead to an increase in information dense word use (Doyle & Frank, submitted). This is further explored in the discussion.

Although we find significant effects, how network size and connectivity influence information density and channel capacity, and how different ways of interpreting information (as we have done here) interact with simple network measures is unclear. Generally, these results suggest that word choice may relate to social-network parameters.

### *Complex Measures*

The complex network measures, *centrality*, *degree* and *scale free*, were log transformed for all analyses due to heavy-tailed distributions. Given the larger number of network variables and their use of similar network properties such as number of nodes or edges, it is possible that some complex network measures will be correlated. To avoid any variance inflation that may occur while using multiple predictors, we determined what variables were collinear using a variance inflation factor (vif) function in R. We first used nodes and edges to predict the variance of each complex network measure. We factor this out by taking the residual of each model and then used the vif function from the R library car to determine what complex network measures exhibited collinearity. Using a conservative VIF threshold of five or less (see Craney & Surls, 2002; Stines, 1995 for review) we determined that no complex network measures used in our model was at risk of any collinearity that would have seriously inflated the variance.[6] All VIF scores were under the conservative threshold for all complex network variables and are therefore not reported. Residuals of complex network measures, having factored out any variance accounted for by nodes and edges, were used to predict each information-theoretic measure presented in Table 5.4.

One or more complex measures accounted for a significant proportion of variance in each information density measure. Certain trends are readily observed across these models. Specifically, word length increased as network transitivity and centrality increased and decreased as network betweenness increased, however no network measure effects were significantly different from random review baseline effects (significance marked by the "?" symbol in Table 5.4). Additionally, RI-Ent and AUI and ACI increased as network degree increased accounting for ~5–10 percent of the variance in each measure. The relationship between network degree and corresponding information measures in true networks was significantly different from baseline. This was also the case for network centrality for both AUI and network transitivity for ACI. Figure 5.3 presents interaction plots for residual ACI by degree (b) and residual transitivity (c) between true and random

review baseline networks. Complex network measures did not share a significant relationship with UIV or CIV.

It is clear that certain network structures predict differences in information density measures even after stringent controls were applied to both information and network measures. Specifically, support for higher information dense messages may be the result of networks that exhibit high global connectivity driven by increases in specific network properties, namely, network degree and centrality. This further supports previous work showing that a shared common ground may bring about higher information-dense language use. Specifically, networks that exhibit a more centralized nucleus and are more densely connected (higher degree) may be more likely to share a similar common ground among many members of the group. If so, a shared common ground may result in more unique language use. Yet, networks that exhibit close, niche-like groupings exemplified by high network transitivity, may infect its members with the same vocabulary, decreasing the overall variability in language use. Further analysis is necessary to unpack the relationship that different social-network structures have with language use.

## *Discussion*

We find pervasive relationships between language patterns, as expressed in information-theoretic measures of review content, and social network variables, even after taking care to control for collinearity. The main findings are listed here:

1. Reviewers used more information-dense words (RI-Ent, AUI) and bigrams (ACI) in networks with more friendship connections.
2. Reviewers used more information-dense words (RI-Ent, AUI) in networks that have a lower Gini coefficient; networks where reviews were more evenly distributed.
3. Reviewers use more information-dense words (RI-Ent, AUI) and bigrams (ACI) as network degree (ratio of friendships connections to number of individuals in the network) increased and as individuals in the network were grouped more around a single center (AUI only).
4. Reviewers used fewer information-dense bigrams as the number of local friendship connections increased (e.g. network transitivity).
5. Unigram information variability (UIV) was higher with higher connectivity; channel capacity was less uniform in networks with more friendship connections.

The predictions laid out at the end of the Methods section are somewhat borne out. Scale-free networks do not appear to have a strong relationship among information-theoretic scores, but networks that exhibit higher transitivity do lead to lower information-dense bigrams (though not significant for any other

information measure) and, while more connections lead to higher information density, they do not lead to lower information variability. Indeed, when considering the last finding, the opposite was true: Networks with higher connectivity used more information-dense words at a more varied rate. Although this was not what we predicted, it is in line with previous work supporting the notion that certain contextual influences afford greater resilience to a varied rate of information transmission (Vinson & Dale, 2014b). In this case, more friendship connections may allow for richer information-dense messages to be successfully communicated *less uniformly*.

We found support for two predictions: (1) high transitivity networks lead to less information-dense bigram language use and (2) high degree networks tend to exhibit higher information density. In addition, more centralized networks also lead to higher information-dense unigram language use. The first prediction suggests that networks where more local mutual interconnections exist may be more likely to infect other members with similar vocabulary. That is, more local connectivity may lead to more linguistic imitation or entrainment. Here we merely find that the structure of reviewers' language use is similar to one another. It is possible that similarities in linguistic structure reveal similarities in semantic content across connected language users, but future research is needed to support this claim.

Support for the second prediction suggests that users adapt their information-dense messages when they are more highly connected. This effect can be explained if we assume that certain social network structures afford groups the ability to establish an overall richer common ground. Previous work shows that increased shared knowledge leads to more information dense messages (Doyle & Frank, submitted; Qian & Jaeger, 2012). It may be that increases in network degree and centrality enhance network members' abilities to establish a richer common ground, leading to more information dense messages. One possibility may be that certain networks tend to review similar types of restaurants. Again, further exploration into how the number of friendship connections, network degree and centrality impact information density and variability is needed to determine the importance of specific network properties in language use. Figure 5.4(a–c) provide example networks that exhibit low, middle and high network ACI given the specific network structures that predict ACI above (e.g. increases in network degree and decreases in transitivity).

## General Discussion

In this chapter we explored how language use might relate to social structure. We built 962 social networks from over 200,000 individuals who collectively wrote over one million online customer business reviews. This massive, structured dataset allowed testing how language use might adapt to structural differences in social networks. Utilizing Big Data in this way affords assessing how differences in one's

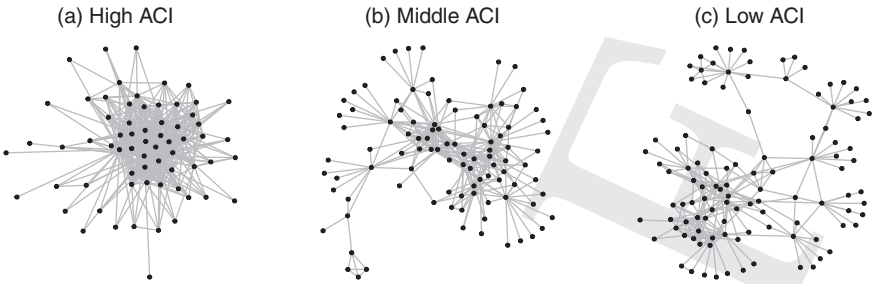


**FIGURE 5.4** Yelp networks occurring at the tails of certain complex network measure distributions (as specified in text) presenting ideal conditions for language use exhibiting high (a) middle (b) and low (c) average condition information (ACI).

local social environment might relate to language use and communication more generally. Our findings suggest that as the connectivity of a population increases, speakers use words that are less common. Complex network variables such as the edge-to-node ratio, network centrality, and local connectivity (e.g. transitivity) also predict changes in the properties of words used. The variability of word use was also affected by simple network structures in interesting ways. As a first exploration our findings suggest local social interactions may contribute in interesting ways language use. A key strength of using a Big Data approach is in uncovering new ways to test theoretical questions about cognitive science, and science in general. Below we discuss how our results fit into the broader theoretical framework of understanding what shapes language.

When controlling for nodes, edges and review length, many $R^2$ values in our regression models were lowered. However, finding that some variability in language use is accounted for by population connectivity suggests language use may be partly a function of the interactions among individuals. Both network degree, centrality and transitivity varied in predictable ways with information measures. Mainly, as the number of connections between nodes increased and as the network became more centralized the use of less frequent unigrams (AUI) increased. Interestingly, networks that exhibit high connectivity and greater centrality may have more long-range connections. A growing number of long-range connections may lead to the inclusion of individuals that would normally be farther away from the center of the network. Individuals in a network with these structural properties may be communicating more collectively; having more readily established a richer common ground. If so, higher information density is more probable, as the communication that is taking place can become less generic and more complex. Additionally, networks with higher local connectivity, or high transitivity tend to use more common language, specifically bigrams. This again may be seen as supporting a theory of common ground, that individuals with more local connectivity are more likely to communicate using similar terminology, in this

case, bigrams. Using a Big Data approach it is possible to further explore other structural aspects of one's network that might influence language use.

While we merely speculate about potential conclusions, it is possible to obtain rough measures of the likelihood of including more individuals at longer ranges. Specifically, a network's diameter—the longest stretch of space between two individual nodes in any network—may serve as a measure of the distance that a network occupies in socio-cultural space. This may be taken as a measure of how many strangers are in a network, with longer diameters being commensurate with the inclusion of more strangers.

It may be fruitful to explore the impact of a single individual on a network's language use. We do not yet explore processes at the individual level, opting instead to sample networks and explore their aggregate linguistic tendencies. Understanding the specifics of individual interaction may be crucial toward understanding how and why languages adapt. We took an exploratory approach and found general support for the idea that network structure influences certain aspects of language use, but we did not look for phonological or syntactic patterns; in fact our analysis could be regarded as a relatively preliminary initial lexical distribution analysis. However, information finds fruitful application in quantifying massive text-based datasets and has been touted as foundational in an emerging understanding of language as an adaptive and efficient communicative system (Jaeger, 2010; Moscoso Del Prado Martín, Kostíc, & Baayen, 2004). In addition, previous work investigating the role of individual differences in structuring one's network are important to consider. For instance, differences in personality, such as being extroverted or introverted, are related to specific network-level differences (Kalish & Robbins, 2006). It is open to further exploration as to how information *flows* take place in networks, such as through hubs and other social processes. Perhaps tracing the origin of the network by determining the oldest reviews of the network and comparing these to the network's average age may provide insight into the importance of how certain individuals or personalities contribute to the network's current language use.

We see the current results as suggestive of an approach toward language as an adaptive and complex system (Beckner et al., 2009; Lupyan & Dale, in press). Our findings stand alongside previous research that reveals some aspect of the structure of language adapts to changes in one's socio-cultural context (Klingenstein et al., 2014; Kramer et al., 2014; Lupyan & Dale, 2010; Vilhena et al., 2014). Since evolution can be thought of as the aggregation of smaller adaptive changes taking place from one generation to the next, finding differences in language within social networks suggests languages are adaptive, more in line with shifts in social and cultural structure than genetic change (Gray & Atkinson, 2003; cf. Chater et al., 2008). The results of this study suggest that general language adaptation may occur over shorter time scales, in specific social contexts, that could be detected in accessible Big Data repositories (see, e.g. recently, Stoll, Zakharko, Moran,

Schikowski, & Bickel, 2015). The space of communicable ideas may be more dynamic, adapting to both local and global constraints at multiple scales of time. A deeper understanding of why language use changes may help elucidate what ideas can be communicated when and why. The application of sampling local social networks provides one method toward understanding what properties of a population of speakers may relate to language change over time—at the very least, as shown here, in terms of general usage patterns.

Testing how real network structures influence language use is not possible without large amounts of data. The network sampling technique used here allows smaller networks to be sampled within a much larger social-network structure. The use of Big Data in this way provides an opportunity to measure subtle and intricate features whose impacts may go unnoticed in smaller-scale experimental datasets. Still, we would of course recommend interpreting initial results cautiously. The use of Big Data can provide further insight into the cognitive factors contributing to behavior, but can only rarely be used to test for causation. To this point, one major role the use of Big Data plays in cognitive science, and one we emphasize here, is its ability to provide a sense of direction and a series of new hypotheses.

## Acknowledgments

We would like to thank reviewers for their helpful and insightful commentary. This work was supported in part by NSF grant INSPIRE-1344279.

## Notes

1 This description involves some convenient simplification. Some abstract and genetic notions of language also embrace ideas of adaptation (Pinker & Bloom, 1990), and other sources of theoretical subtlety render our description of the two traditions an admittedly expository approximation. However, the distinction between these traditions is stark enough to warrant the approximation: The adaptive approach sees all levels of language as adaptive across multiple time scales, whereas more fixed, abstract notions of language see it as only adaptive in a restricted range of linguistic characteristics.

2 Massive online sites capable of collecting terabytes of metadata per day have only emerged in the last 10 years: Google started in 1998; Myspace 2003; Facebook 2004; Yelp 2004; Google+ 2011. Volume, velocity and variety of incoming data are thought to be the biggest challenges toward understanding Big Data today (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012).

3 Previous research calls this Information Density and uses this as a measure of Uniform Information Density. We use the name Average Conditional Information given the breadth of information-theoretic measures used in this study.

4 Note: AUI and ACI were calculated by taking only the unique n-grams.
5 Our approach toward controlling for collinearity by residualizing variables follows that of previous work (Jaeger, 2010). However, it is important to note the process of residualizing to control for collinearity is currently in debate (see Wurm & Fisicaro, 2014 for review). It is our understanding that the current stringent use of this method is warranted provided it stands as a first pass toward understanding how language use is influenced by network structures.
6 The variance inflation acceptable for a given model is thought to be somewhere between five and ten (Craney & Surles, 2002). After the variance predicted by nodes and edges was removed from our analyses, no complex network measure reached the variance inflation threshold of five.

## References

Abrams, D. M., & Strogatz, S. H. (2003). Linguistics: Modelling the dynamics of language death. *Nature*, *424*(6951), 900.

Aylett, M. P. (1999). Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and syllabic duration. *Proceedings of ICPhS–99*, San Francisco.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences, 17*(7), 348–360.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., … Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language learning, 59*(s1), 1–26.

Bentz, C., Vererk, A., Douwe, K., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PloS one*, 10(6), e0128254.

Bentz, C., & Winter, B. (2013). Languages with more second language speakers tend to lose nominal case. *Language Dynamics and Change, 3*, 1–27.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415), 295–298.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience, 10*(3), 186–198.

Chater, N., Reali, F., & Christiansen, M. H. (2009). Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences, 106*(4), 1015–1020.

Choi, H. Y., Blumen, H. M., Congleton, A. R., & Rajaram, S. (2014). The role of group configuration in the social transmission of memory: Evidence from identical and reconfigured groups. *Journal of Cognitive Psychology, 26*(1), 65–80.

Chomsky, N. A. (1957) *Syntactic Structures*. New York: Mouton.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509.

Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives.* New York, NY: Little, Brown.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering 14*(3), 391–403.

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems, 15*, 1150017/1–1150017/16.

Dale, R., & Vinson, D. W. (2013). The observer's observer's paradox. *Journal of Experimental & Theoretical Artificial Intelligence, 25*(3), 303–322.

Dodds, P. S., Watts, D. J., & Sabel, C. F. (2003). Information exchange and the robustness of organizational networks. *Proceedings of the National Academy of Sciences*, *100*(21), 12516–12521.

Doyle, G., & Frank, M. C. (2015). Shared common ground influences information density in microblog texts. In Proceedings of NAACL-HLT. pp. 1587–1596.

Ember, C. R., & Ember, M. (2007). Climate, econiche, and sexuality: influences on sonority in language. *American Anthropologist, 109(*1), 180–185.

Everett, C. (2013). Evidence for direct geographic influences on linguistic sounds: The case of ejectives. *PloS one, 8*(6), e65275.

Freeman, L. C. (1979). Centrality in social Networks conceptual clarification. *Social Networks, 1*(3), 215–239.

Gordon, R. G. (2005). *Ethnologue: Languages of the World, 15th Edition*. Dallas, TX: SIL International.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*(6965), 435–439.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition, 135*, 21–23.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve?. *Science, 298*(5598), 1569–1579.

Jaeger, F. T. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*(1), 23–62.

Jurafsky, D., Chahuneau, V., Routledge, B. R., & Smith, N. A. (2014). Narrative framing of consumer sentiment in online restaurant reviews. *First Monday, 19*(4).

Kalish, Y., & Robins, G. (2006). Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure. *Social Networks, 28*(1), 56–84.

Klingenstein, S., Hitchcock, T., & DeDeo, S. (2014). The civilizing process in London's Old Bailey. *Proceedings of the National Academy of Sciences, 111*(26), 9419–9424.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences,* 111(14), 8788–8790.

Kudyba, S., & Kwatinetz, M. (2014). Introduction to the Big Data Era. In S. Kudyba (Ed.), *Big Data, Mining, and Analytics: Components of Strategic Decision Making* (pp. 1–15). Boca Ratan, FL: CRC Press.

Labov, W. (1972a). Language in the inner city: Studies in the Black English vernacular (Vol. 3). Philadelphia, PA: University of Pennsylvania Press.

Labov, W. (1972b). *Sociolinguistic patterns* (No. 4). Philadelphia, PA: University of Pennsylvania Press.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., & Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. *Science*, *323*(5915), 721.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, and T. Hoffman (Eds.), *Advances in neural information processing systems (NIPS) 19*, pp. 849–856. Cambridge, MA: MIT Press.

Lieberman, E., Michel, J. B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature, 449*(7163), 713–716.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one, 5*(1), e8559.

Lupyan, G., & Dale, R. (2015). The role of adaptation in understanding linguistic diversity. In R. LaPolla, & R. de Busser (Eds.), *The Shaping of Language: The Relationship between the Structures of Languages and their Social, Cultural, Historical, and Natural Environments.* (pp. 289–316). Amsterdam, The Netherlands: John Benjamins Publishing Company.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big Data. The management revolution. *Harvard Bus Rev, 90*(10), 61–67.

Moscoso del Prado Martı́n, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition, 94*(1), 1–18.

Nettle, D. (1998). Explaining global patterns of language diversity. *Journal of Anthropological Archaeology, 17*(4), 354–374.

Nichols, J. (1992). *Linguistic diversity in space and time*. Chicago, IL: University of Chicago Press.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature, 417*(6889), 611–617.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language, 87*(3), 539–558.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences, 108*(9), 3526–3529.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences, 13*(4), 707–727.

Qian, T., & Jaeger, T. F. (2012). Cue effectiveness in communicatively efficient discourse production. *Cognitive Science, 36*(7), 1312–1336.

Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, *46*(4), 377–396.

Reali,F., Chater, N., & Christiansen, M. H. (2014). The paradox of linguistic complexity and community size. In E.A. Cartmill, S. Roberts, H. Lyn & H. Cornish (Eds.), The evolution of language: *Proceedings of the 10th International Conference* (pp. 270–277). Singapore: World Scientific.

Sapir, E., Language: *An Introduction to the Study of Speech* (Harcourt, Brace and company, 1921).

Shannon C. A. (1948) A mathematical theory of communications. *Bell Systems Technical Journal, 27*(4): 623–656.

Stine, R. A. (1995). Graphical interpretation of variance inflation factors. *The American Statistician*, *49*(1), 53–56.

Stoll, S., Zakharko, T., Moran, S., Schikowski, R., & Bickel, B. (2015). Syntactic mixing across generations in an environment of community-wide bilingualism. *Frontiers in Psychology*, 6, 82.

Triandis, H. C. (1994). *Culture and social behavior*. New York, NY: McGraw-Hill Book Company.

Trudgill, P. (1989). Contact and isolation in linguistic change. In L. Breivik & E. Jahr (Eds.), *Language change: Contribution to the study of its causes*, pp. 227–237. Berlin: Mouton de Gruyter.

Trudgill, P. (2011). *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford, UK: Oxford University Press.

Vilhena, D. A., Foster, J. G., Rosvall, M., West, J. D., Evans, J., & Bergstrom, C. T. (2014). Finding cultural holes: how structure and culture diverge in networks of scholarly communication. *Sociological Science, 1*, 221–238.

Vinson, D. W., & Dale, R. (2014a). An exploration of semantic tendencies in word of mouth business reviews. *In Proceedings of the Science and Information Conference (SAI), 2014* (pp. 803–809). IEEE.

Vinson, D. W., & Dale, R. (2014b). Valence weakly constrains the information density of messages. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1682–1687). Austin, TX: Cognitive Science Society.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua, 117*(3), 543–578.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72*, 37–48.