COMPLEXITY INTERNATIONAL

Volume 12

Patterns in syntactic dependency networks from authored and randomised texts

Markus Brede[1] and David Newth[1]
[1]*CSIRO Centre for Complex Systems Science,*
GPO Box 284 Canberra, ACT 2601, AUSTRALIA
Email: markus.brede@csiro.au, david.newth@csiro.au

Abstract
The syntactic relationships between words allow a communicator to express a virtually endless array of thoughts by a finite set of elements. The co-occurrence of words in a sentence reflects the syntactic dependency between words, and can be represented as a directed graph. In this account we compiled the grammar dependency networks of 86 texts from 11 well known English authors. In an analysis of the common and specific features of these networks we try to attribute network properties to individual authors. A pointwise defined measure shows no significant groups which could be identified with authors. Further, a comparison to randomized versions of the same texts shows a systematic, but very small difference between networks constructed for the originals and the randomisations, respectively. This suggests, that the scale-free and small world-like nature of these networks can be explained by an underlying regularity in the word frequency distribution, known as Zipf's law. A stochastic model, which allows the construction of networks for arbitrary word frequency distributions, illustrates this idea.

1. Introduction

Human language has the very distinct property of being able to communicate an endless array of thoughts through the use of a finite set of elements (Chomsky 2002). The power of language can—at least in part—be attributed to the flexibility given to each word by the rules of assemblage or syntax. In short a syntax is a set of rules for combining words into logical phrases and sentences. Such rules ultimately define explicit syntactic relationships among words (Chomsky 1965).

For some time now language has been regarded as a complex adaptive system (Gell-mann 1994; Pinker 1997). Over time, the meanings, spelling, accepted usage and even the syntax evolve to reflect trends, and social norms. Language usage, also varies between authors, social status and education, all of which influence an individual's ability to assemble, and colour written and spoken prose.

Recently the application of network theory to an array of complex systems has revealed that they share a number of common topological properties (for a summary see, (Albert & Barabási 2002; Albert & Barabási 2002)). The application of network theory based approaches to grammar dependency networks, have provided new insights into semantic networks (Ferrer i Cancho & Solé 2001; Ferrer i Cancho et al. 2004). Network analysis provides an alternative to

the universal grammar approaches proposed by Chomsky ((Chomsky 1965; Chomsky 2002)), and supported by others (e.g. Pinker 1997). These recent studies have discovered that linguistic networks display both small world (Watts & Strogatz 1998), and scale-free (Albert & Barabási 1999) network properties. They also found that some of these network properties vary between languages, while others seem to remain constant.

In this paper we analyse the syntactic structures formed by 86 texts of 11 well known authors. Our intention here is twofold. First we aim to investigate the topological properties of these linguistic networks and attempt to identify in how far peculiarities can be attributed to individual authors. The second aim is to find out in how far the linguistic networks differ from networks generated for randomized texts. For this purpose, we develop a simple random null model. The model abstracts from all syntactic rules of real language, but still conserves statistical regularities such as Zipf's law (Zipf 1932) for the frequency of word occurrences.

2. Dependency Grammar Networks

Here we have analysed the text of 86 books. The text of these works was obtained from the Gutenberg Project (`http://www.gutenberg.net/`). The texts were preprocessed to remove disclaimer statements, formatting, obscure symbols and standard punctuation marks. Appendix 1 lists the texts used in this study.

The networks that are analysed here have been defined according to the dependency grammar formalism. Dependency grammars are a family of grammatical formalizations, which share the assumption that syntactic structures consist of a lexicon and binary dependencies linking its elements. Such a formalization lends itself easily to a network representation, where the words represent nodes, and the dependencies are depicted as arcs. The networks are constructed in accordance with the procedures defined in (Ferrer i Cancho & Solé 2001).

Essentially each distinct word forms a node within the network. All the links between nodes are directed. An arc is made between two nodes, if a word is within the next two words within a sentence. Figure 1 illustrates the construction of a grammar dependency network for the sentence fragment "*It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...*" from Charles Dickens' A Tale of Two Cities. For a more detailed account of the construction of grammar dependency networks see (Ferrer i Cancho & Solé 2001; Ferrer i Cancho et al. 2004).

3. Word Frequency and Degree Distribution

3.1 Zipf's Law

An understanding of the patterns of language requires an understanding of the regularities that occur in the elements that make up the language. One of the most well known regularities in human language is Zipf's law (Zipf 1932). Zipf's law states that the frequency f of words decays in accordance with a power law of its rank r, i.e. $f \propto r^{-\alpha}$ with an exponent $\alpha \approx 1$. Figure 2 (A), shows the decay of the frequency of the 100 most common words occurring across all of the Gutenberg texts analysed here.

Further, typically the number of words sharing a rank increases with r. Figure 2 (B) illustrates that also the number n of words occurring with the same frequency f obeys a power law $n \propto f^{-\beta}$, with $\beta_h \approx 1.33$ for very high ranks and $\beta_m \approx 1.52$ for intermediate ranks. The first 1000 most frequent words occupy their rank alone, i.e. $\beta_l = 0$. Recasting this, one obtains that the probability p_r to find a word with rank r obeys

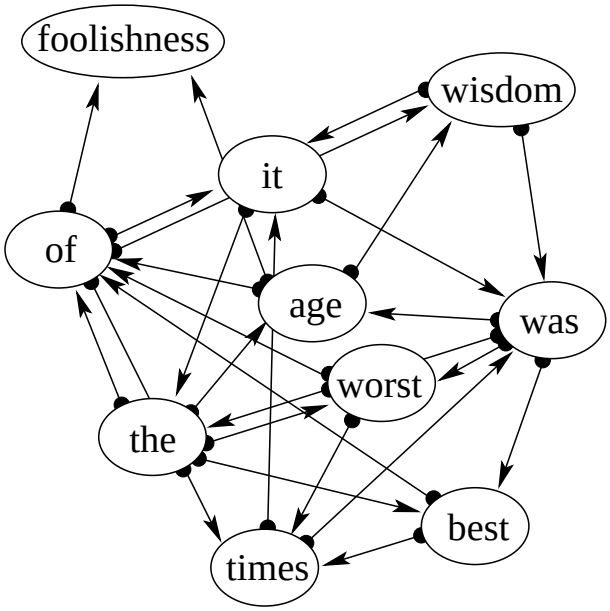

Figure 1. *A sample of a dependency grammar network. This network is generated from the first part of a sentence found in Charles' Dickens story A Tale of Two Cities, "It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ..." Note that the network is highly cliquish and has a relatively short path length between nodes.*

$$p_r \propto \begin{cases} r^{-\alpha} & \text{if } r < 1000 \\ r^{-\alpha-\beta_m} & \text{if } 50 < r < 1000 \\ r^{-\alpha-\beta_h} & \text{otherwise} \end{cases} \quad (1)$$

Various mechanisms have been identified to obtain such distributions. One of more surprising results is that Zipf's law can be obtained by assembling "words" from random sequences of characters (Li 1992). This may be one reason why Zipf's-law-like distributions are so common in many different contexts. An alternative model suggests that the power law observed in the word frequency could be the result of an evolutionary optimisation process (Ferrer i Cancho & Solé 2003).

3.2 Degree Distribution

One of the hallmarks of many complex networks is a degree distribution with a power law tail $P(k) \sim k^{-\gamma}$, with exponents γ typically in the range between 2 and 3 (Albert & Barabási 2002). A previous study (Ferrer i Cancho & Solé 2001) has shown that the syntactic dependency network of the English language also obeys a power law. Distinguishing a kernel lexicon (the most frequently used words), Ferrer i Cancho & Solé (Ferrer i Cancho & Solé 2001) found an exponent $\gamma = -2.7$ holding for words belonging to the kernel. Beyond the kernel, the degrees of less frequently used words are found to obey a power law with exponent $\gamma = -1.5$. The exponent for words belonging to the kernel is similar to exponents found in networks formed by preferential attachment (Albert & Barabási 1999). Accordingly, Ferrer i Cancho & Solé (Ferrer i Cancho & Solé 2001) argue that preferential attachment plays a role in the formation of the network reflecting the core lexicon.

Conversely, the non-kernel part of the lexicon is highly specialised and contains a subset of words not common to all speakers. Figure 3 shows the in-degree distributions for two networks compiled from two of the Gutenberg texts. In comparison to the networks constructed in

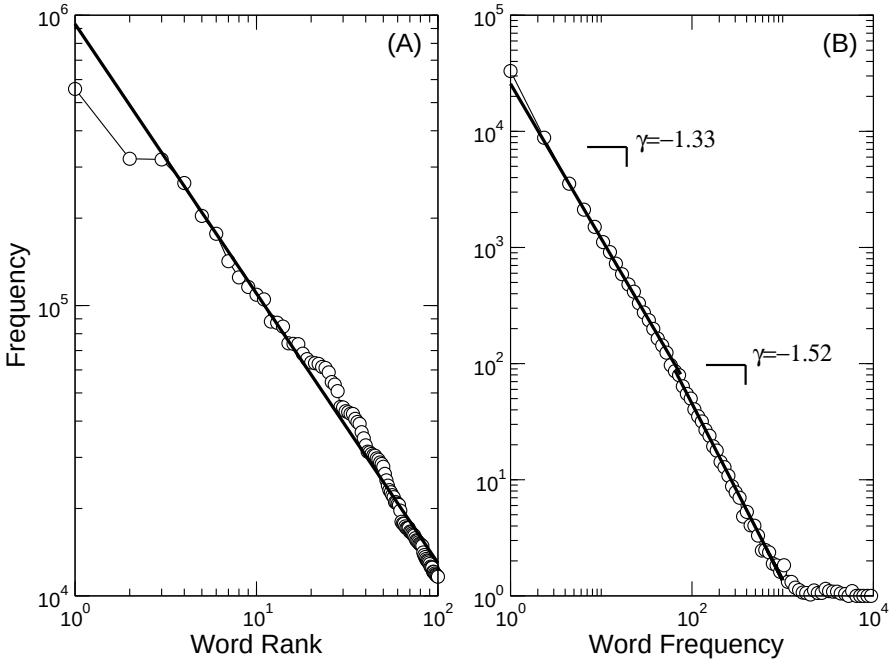

Figure 2. *An Example of Zipf's Law. (A) The decay of the frequency of the 100 most common words occurring across all Gutenberg texts. (B) The decay of the number of words with the same frequency. The decay of the numbers of words sharing the same ranks has several distinct power laws for high, medium and low ranks.*

(Ferrer i Cancho & Solé 2001; Ferrer i Cancho et al. 2004) each network is constructed from a relatively small set of words. Additionally, because of the finite size of the compiled texts, only a subset of the links found in the whole language network are present in these samples. Hence, the compiled networks represent only relatively small subgraphs of the whole language network. Their specific nature is determined by the bias of the text, from which they are compiled. However, the size of the networks—though small in comparison to the whole language network—is still large enough to observe statistically significant patterns.

The data in Fig. 3 show, that the small size of the subgraph does not destroy the overall pattern observed by Ferrer i Cancho and Solé: both networks exhibit a power law degree distribution. However, the exponents $\gamma_{in}^{A} \approx -2.1$ (for Charles Dickens' "A Tale of two Cities") and $\gamma_{in}^{B} \approx -1.76$ (for Jane Austin's "Emma") are different from those observed for the whole lexicon. Further, although both texts are not substantially different in length ($W_{emma} = 158161$ compared to $W_{two\ cities} = 135710$), network size, number of links, and the exponents of the degree distributions differ substantially. These observations seem to confirm the thought that linguistic sub-networks compiled from relatively short texts are not just a representative sample of the whole linguistic network, but define distinct subnetworks, seemingly characteristic for the text that they are compiled from.

We also "scrambled" the texts by randomly selecting pairs of words and exchanging them. Repeated often[1] enough this procedure destroys all correlations between words in the text. The result conserves the frequencies with which individual words appear, but is otherwise an apparently senseless assemblage of words, i.e. a "randomized text".

Figure 3 (overlay panels) shows the in-degree distributions of the networks constructed from

[1]We repeated the procedure 20 times the textlength, which ensures that almost every word has been picked and exchanged with another randomly chosen word at least once.

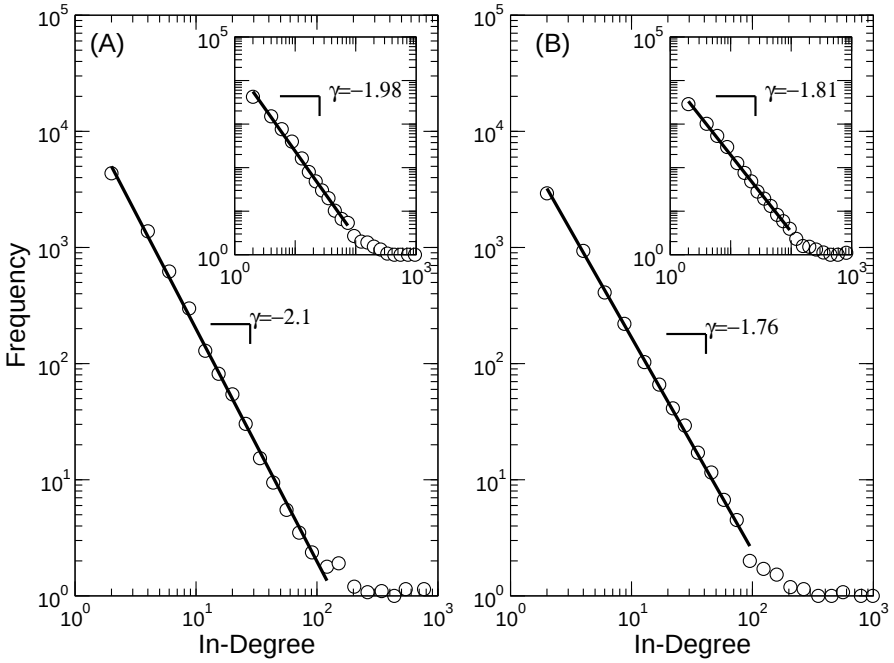

Figure 3. *In-Degree distributions for two linguistic networks compiled from the Gutenberg texts. (A) Charles' Dickens "A tale of two cities". Here* $N = 9815$ *and* $L = 70916$, $\gamma_{in}^{A} \approx -2.1 \pm .05$, $\gamma_{in}^{A_{rand}} \approx -1.98 \pm .07$ *(overlay). (B) Jane Austen's "Emma". The network contains* $N = 7316$ *vertices and* $L = 74093$ *links. We find* $\gamma_{in}^{B} \approx -1.76 \pm .02$ *and* $\gamma_{in}^{B_{rand}} \approx -1.81 \pm .05$ *for the randomized version of "Emma" (overlay).*

the randomized texts. We find that the respective exponents (e.g. $\gamma_{in}^{A} \approx -1.98$ and $\gamma_{in}^{A_{rand}} \approx -1.81$ for "Emma") are within the error bounds of the exponents of the real texts. From this, we can conclude that the exponent of the degree distribution is not specific to the syntactic structure of the network. Assembling a text at random from the same set of words yields almost the same results. Clearly, in the latter procedure almost all syntactic rules are violated.

In the method of network construction explained in §2. one could also assign a "weight" to each link by counting how many times the respective words forming its ends co-occur. It turns out that the distribution of these weights also follows a power law with an exponent $\gamma^{co} \approx -2.3$ (cf. Fig. 4). As for the degree distributions, the differences between the original texts and the scrambled versions are very small.

4. Statistical Properties of Complex Networks

In this section we will briefly introduce some quantities that have previously been used to characterize linguistic and other complex networks. We briefly summarize some of the previous results and then turn to the analysis of text-specific network patterns.

4.1 Small World Properties

Many complex networks display what is known as "small world" properties (Watts & Strogatz 1998). A small world is characterized by the extent to which a network is locally similar to a clique; and how short the average distance is between any pair of nodes within the network.

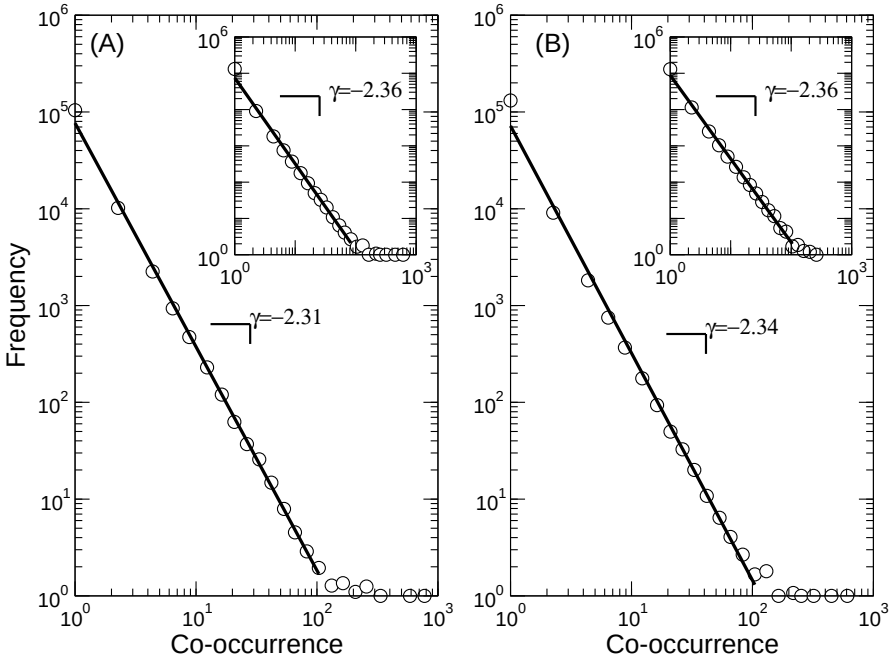

Figure 4. *Word co-occurrence distributions for two texts from the Gutenberg texts. (A) Jane Austen's "Emma", a text containing* $W = 158161$ *words.. We find* $\gamma_{co}^{A} \approx -2.31$ *and* $\gamma_{co}^{A_{rand}} \approx -2.36$ *for the randomized version of "Emma" (overlay). (B) Charles' Dickens "A tale of two cities". Here* $W = 135710$*,* $\gamma_{co}^{B} \approx -2.34$*, and* $\gamma_{co}^{B_{rand}} \approx -2.36$ *(overlay).*

The cliquishness or degree of clustering within a network is measured by the clustering coefficient (Watts & Strogatz 1998). The average clustering coefficient is defined as $\overline{c} = 1/N \sum_i c_i$. Given a node i, with k_i neighbours, E_i is the number of links between the k_i neighbours. We define the clustering coefficient as the ratio between the number of links that actually exist between the neighbours (E_i) and the potential number of links $k_i(k_i - 1)$, i.e.

$$c_i = \frac{E_i}{k_i(k_i - 1)}. \tag{2}$$

The shortest path length, is the minimum number of edges that need to be traversed, in order to move from vertex i to vertex j, and is denoted by $d(i, j)$. The average shortest path length is:

$$\bar{l} = \frac{1}{N(N-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} d(i, j). \tag{3}$$

Another networks property that we measure is the diameter of the network. The diameter of a network is the longest shortest path within the network. More formally

$$l_{max} = \max_{i,j} d(i, j). \tag{4}$$

A network is called a small world if it exhibits a high clustering coefficient while having a low average path length. Hence it combines properties of regular lattices (which are highly cliquish) with the small average path lengths of random graphs (Erdös 1959; Bollobás 1998), the average shortest path length of which only grows logarithmically with the network's size.

It has recently been shown that linguistic networks are small worlds (Ferrer i Cancho & Solé 2001). For instance, analysing the English language, the above authors found a value of

$\bar{c} \approx .687$ while $\bar{l} \approx 2.63$, showing that on average it takes surprisingly few steps to reach any other word from a random starting word.

As noted in (Ferrer i Cancho & Solé 2001) preferential attachment appears to play a role in the formation of the network corresponding to the whole English network. Notably, preferential attachment itself leads to 'ultrasmall' networks for which $\bar{l} \sim \ln \ln N$ (Cohen & Havlin 2003). However, the observed high clustering coefficients can not be attributed to preferential attachment alluding to a different mechanism shaping this aspect of the linguistic networks.

4.2 Assortativness

A network is said to show assortative mixing if nodes of high degree are typically connected to other nodes of high degree. Conversely, in a dissortatively mixed network nodes with many links tend to be adjacent to nodes with few neighbours. Following (Newman 2002) we use a Pearson correlation coefficient Γ to quantify the assortativeness of a network. Newman (Newman 2002) defines this correlation as:

$$\Gamma = \frac{c\sum_i j_i k_i - \left[c\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2}{c\sum_i \frac{1}{2}\left(j_i^2 + k_i^2\right) - \left[c\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2}, \tag{5}$$

where j_i and k_i are the degrees of the veritices at the ends of the i^{th} edge. With m defined as the number of edges we set $c = 1/m$. A network displays assortative mixing when $\Gamma > 0$ and dissortative mixing when $\Gamma < 0$. While some social networks are assortative, many other networks with power law degree distributions are dissortatively mixed (Newman 2003). Values for linguistic networks (Ferrer i Cancho et al. 2004) are found in a range from $\Gamma = -.06$ (Czech) to $\Gamma = -.2$ (Romanian). Thus, linguistic networks show dissortative mixing. Again these values are different from the value ($\Gamma = 0$) of a network which is formed by preferential attachment alone.

5. Analysis of Texts

For all the networks we calculated the connectivity, average shortest pathlength, diameter, clustering coefficient, degree of assortativeness, and the exponents for the total degree, in-degree and out-degree distributions. This information is summarized in Appendix 1. Also, from each of the texts 10 randomized (scrambled) versions were produced. We applied the network generation algorithm to the randomized texts and calculated the average network properties pertaining to them. In the following sections we examine the global properties of these networks; next we attempt to identify author specific characteristics.

5.1 Global Trends

Among the analysed texts we find a substantial variation in text length, ranging from about 3000 to 360,000 words. We included authors (e.g. Charles Dickens) who tend to write very long texts, but also others who wrote shorter texts. The analysis of the networks reveals that their properties change with size. For instance, the connectivity typically decays with size as $p \propto N^{-\delta}$ with $\delta = .6 \pm .1$. As the network size depends on the text length, network properties change with text length. To reduce this size effect we use only network properties relative to the respective property calculated for randomized texts. In other words, for a network property

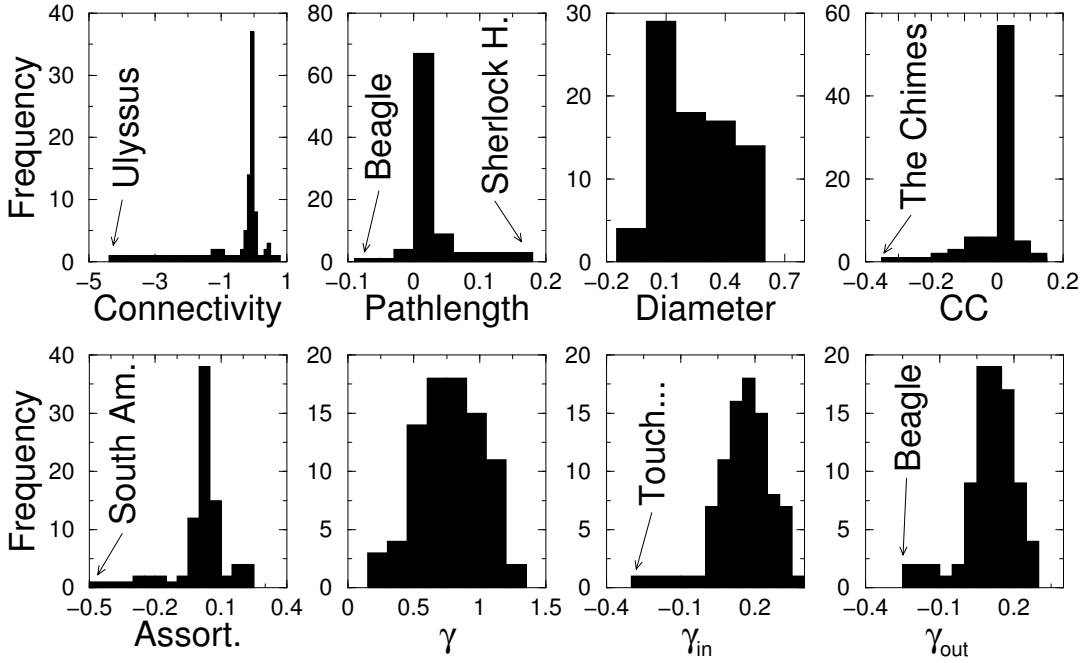

Figure 5. *Relative difference* $(y - y_{rand})/y$ *of the network properties* y *to the properties* y_{random} *of the networks corresponding to the respective randomized texts. Typically, though there are tails of a few outliers, almost all networks are found in a small band, deviating not more than 10% from networks of the randomized texts. More substantial differences are found in the network diameter (which is on average approx. 20% underestimated by the random networks) and the exponent* γ *of the (total) degree distribution.*

y we define a relative property y_{rel} as

$$y_{rel} = \frac{(y - y_{rand})}{y}. \tag{6}$$

The distribution of the relative differences are shown in Fig. 5. Disregarding some outliers, properties of the texts are typically found within 10% deviation from the randomized texts.

In random texts word combinations occur which are forbidden in the syntactic structure of language. Hence, in a random text a word can potentially acquire more links than in a human produced text. So, not surprisingly, we find a trend to lower connectivities in the original networks. Similarly, in the randomized texts, infrequent words typically occur after frequently used words. Accordingly, combinations of infrequent words —which in normal language lead to longer path length— are very scarce in randomized texts. Thus, one also finds a tendency for randomized texts to have slightly higher path length, and considerably larger diameters. The same argument explains a typically steeper decay in the in- and out-degree distributions of the randomized networks. In the randomized texts, highly frequent words have a better chance to acquire more neighbours than in the original texts. Consequently, the randomized networks have more vertices with many neighbours. As a result the exponent characterizing the decay of the in- and out-degree distributions is expected to be lower. For the clustering coefficients and the assortativeness we don't find a systematic trend to higher or lower values in the networks for the randomized texts. The same holds for the exponents γ for the total degree distribution.

Author	p	$\bar{l}$	$\bar{l}_{max}$	$\bar{c}$	Γ	γ	γ_{in}	γ_{out}	average
Alcott	.19	.46	.93	.32	.17	.81	.54	.60	.51
Austen	.48	.58	.93	.89	1.10	1.03	.80	1.20	.87
Conan Doyle	.14	1.48	.63	.60	.49	.50	.53	.67	.63
Darwin	1.48	1.86	.61	1.44	1.16	1.01	1.13	1.58	1.28
Defoe	.29	.63	1.04	.33	.77	1.00	.88	.74	.71
Dickens	.41	.70	.61	1.04	.72	.79	1.04	.91	.78
Joyce	4.26	1.04	1.33	.59	1.60	1.65	.59	.75	1.47
Lawrence	1.74	.82	.49	1.07	1.10	.58	2.23	1.61	1.21
Malthus	.84	1.28	.60	1.51	1.80	1.86	1.45	1.18	1.32
Scott	1.66	.65	.80	.79	1.16	.74	.87	.62	.91
Trollope	.65	.45	1.17	.90	.91	1.29	1.01	1.10	.93
Author avg.	1.05	.90	.83	.86	.99	1.02	1.01	1.00	.97

Table 1. *Table of group consistencies for all analysed authors. The rows give the consistencies as calculated for an individual network property (see Eq. 7). A value less than one means that books of the respective authors are closer to each other than to the average text of all other authors.*

5.2 Author Correlations

In this subsection, we are interested in finding out, whether networks compiled for one author have structural characteristics unique to that author. For this purpose, we define a group consistency of a relative property y, for an author G as

$$c^{(y)}(G) = \frac{|B|(|B|-1)\sum_{\substack{g_1,g_2\in G\\ g_1<g_2}} |y(g_1)-y(g_2)|}{|G|(|G|-1)\sum_{\substack{b_1,b_2\in B\\ b_1<b_2}} |y(b_1)-y(b_2)|}, \tag{7}$$

where B denotes the set all analysed texts, and $|\cdot|$ is the usual notation for the cardinality of a set. This group consistency gives the average distance between any two members of G relative to the average distance of any two elements chosen from all texts. The average relative group consistency of a group G in the space spanned by m properties y_m then is

$$c(G) = 1/m \sum c^{(y_m)}(G). \tag{8}$$

Likewise, the "suitability" $c^{(y_i)}$ of a property y_i to identify groups can be defined as

$$c^{(y_i)} = 1/|G| \sum_{G} c^{(y_i)}(G). \tag{9}$$

For a value $c(G) = 1$ elements of G have the same average distance to each other as any two randomly chosen elements from the whole set B. Subsets whose elements have higher average distance than the whole have $c(G) > 1$, whereas values $c(G) < 1$ indicate the onset of local concentrations of elements.

We calculated the values of $c^{(y_m)}(G)$ and the averages $c(G)$ for all considered network properties and all eleven authors. These results are summarized in table 1. Averaged over all authors, all properties have group consistencies in a range between $c^{(p)} = 1.05$ (for the connectivity) and $c^{(\bar{l}_{max})} = .83$ (for the diameter). Even diameter, the best suited property to identify groupings in our dataset, has a value of the consistency close to one. Averaged over all authors, the data in table 1 show no significant group consistency. Yet, while our measure

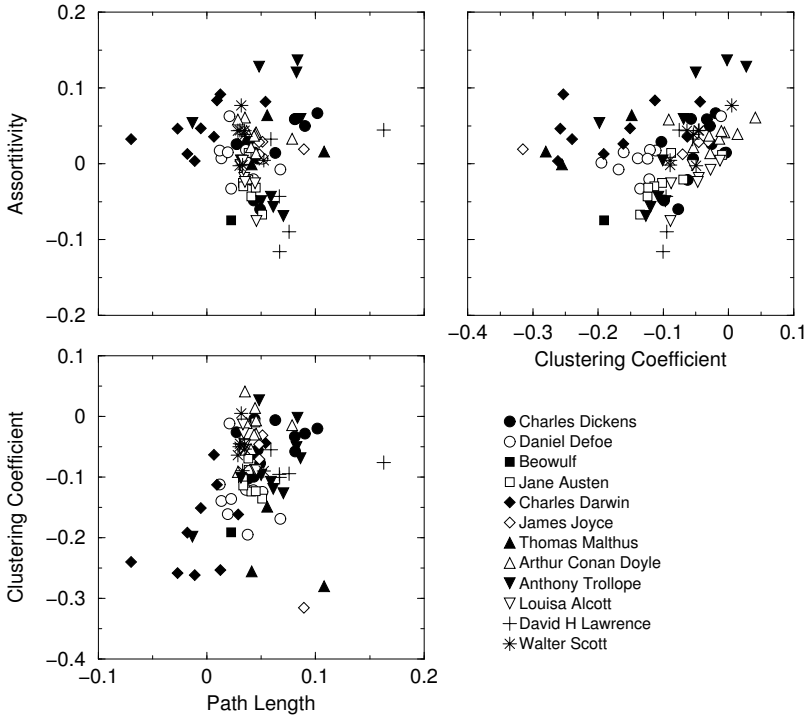

Figure 6. *Relative differences to randomized texts plotted against each other for the three best suited properties to identify groupings, i.e. clustering coefficient, assortativeness, and average shortest pathlength.*

typically reveals no groupings, the works of some authors (e.g. Alcott and Conan-Doyle) form relatively consistent groups.

Judging the results one must bear in mind, that the above pointwise distance-oriented measure for group consistency (Eq. (7)) does not take account of trends and patterns in a group. Changes in an author's writing style or subject would lead one to expect certain patterns of earlier and later works in the property space. From this point of view, it appears not surprising to find more variability in scientific texts (cf. Malthus and Darwin). Constancy, i.e. successive texts having only a small distance to each other, could allude to a relatively unchanged writing style and subject choice, which we find in the works of Alcott or Conan-Doyle.

Figure 6 shows the above identified three best properties [2] plotted against each other. Closer inspection might suggest the formation of groups distinguished by patterns and not by relative distance. Groupings defined in this way could be identified by machine learning techniques (Kennedy et. al 1995), which is, however, out of the scope of the present paper.

6. Networks from Random Texts

The previous sections have underlined that the analysed linguistic networks form small worlds and have power law tails in the degree distribution. Yet the same is observed for networks constructed from randomized texts. These texts lack all the syntactic structure of the originals. Hence it seems unlikely that the small world structure of the linguistic networks is only a consequence of structural rules in language which allow us to form meaningful sentences. However, the randomized texts—though being freed of all grammatical constraints of language—still retain the frequencies with which individual words have been used in the original texts. For

[2] We have chosen to use path length and not diameter, as diameter scales with path length

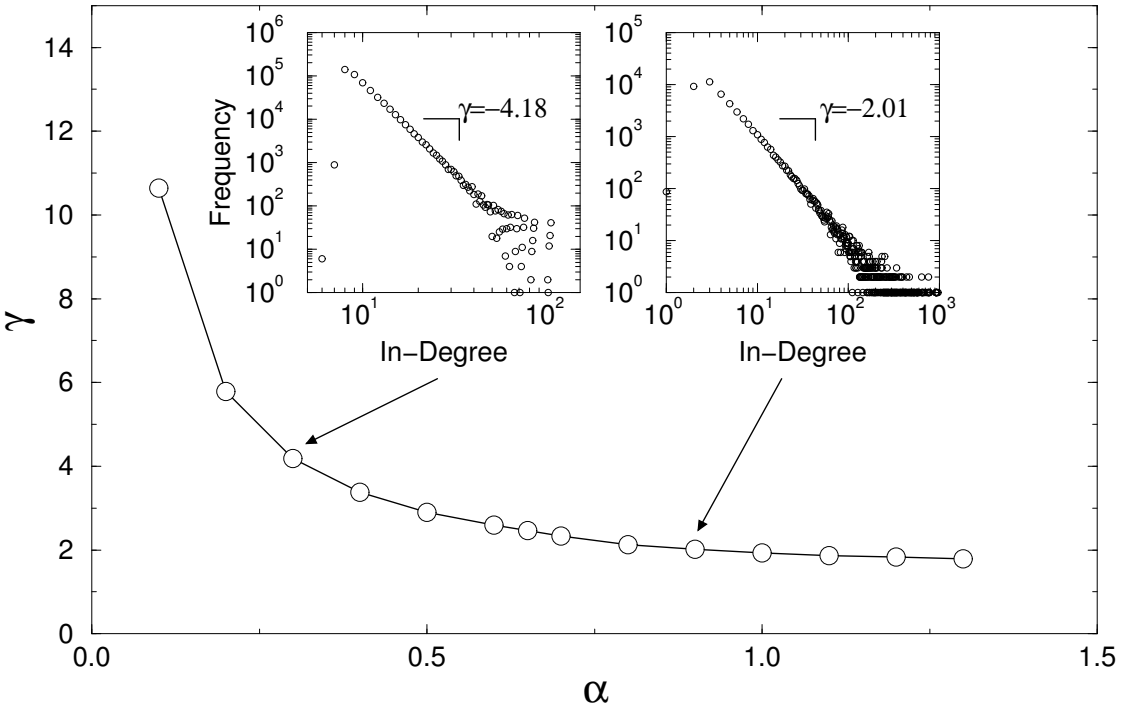

Figure 7. *Dependence of the exponent γ of the degree distribution on the exponent α of the word frequency distribution for networks constructed by the stochastic model described in the text. The data are for networks with $N = 5000$ nodes and a connectivity $p = L/N^2 = 0.0022$. The network was chosen of such a size, that no significant change in γ is observed for further increases of N.*

example, as pointed out earlier in §3.1, the rank ordering of the frequencies follows Zipf's law.

In this section we explore a random construction mechanism for networks. We assume that a sequence of words $W = \{w_r\}_{r=1}^{N}$ with individual frequencies $f(w_r)$ is ordered in descending order $f(w_i) \geq f(w_j)$ for $i \leq j$, starting with the most frequent word w_1. As in the real texts we assume that the frequencies follow a power law $f(r) = Cr^{-\alpha}$, where $C^{-1} = \sum_{i=1}^{N} r^{-\alpha}$. At each iteration $0 < t < T$ we draw a symbol s_t from W with probability $f(r)$ at random. For $t < T$ we connect s_t to $s_t + 1$. The number of iterations T (or the "text length") is chosen such that all the constructed networks have the same connectivity.

Simulation results show that this procedure produces a plethora of networks with power law degree distribution (cf. Fig. 7). A cross check, using non-power law word frequency distributions did not lead to power laws in the degree distributions of the constructed networks. This rules out, that the power laws in the degree distributions are simply a consequence of the linear construction process.

The exponents γ of the identical in- and out-degree distributions are determined by the exponents α of the word frequency distribution $f(r)$. Values of γ are found to decay with α, quickly saturating slightly below $\gamma = -2$ for values of α above 1. The exponents roughly correspond to the exponents found in the analysed linguistic networks (e.g. $\gamma_{in}^{A} = -2.1$ or $\gamma_{in}^{B} = -1.76$ see Fig. 3). These results seem to suggest that the power law tail observed in linguistic networks is essentially a consequence of Zipf's law for the word frequency distribution.

Next we analysed the clustering coefficients, average shortest path length and values for the assortativeness of the above networks (cf. Fig. 8). For comparison, an Erdös-Rényi random graph has a clustering coefficient of $c \approx p = .0022$, assortativeness $\Gamma = 0$ and a path length $l \approx \ln N / \ln(Np) \approx 3.6$. For low exponents $\alpha < .5$ we find clustering coefficients, shortest

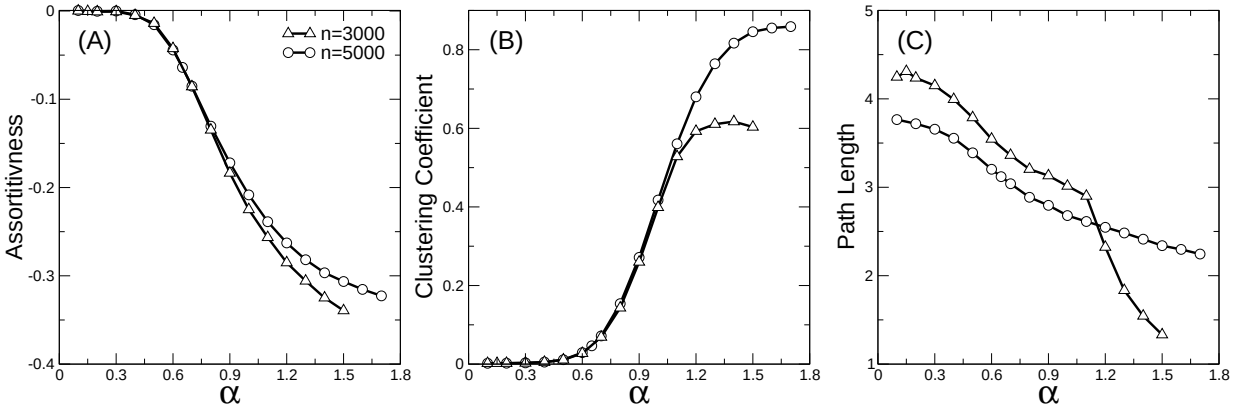

Figure 8. *Simulation results for the assortativeness (A), the clustering coefficient (B), and the average shortest path length (C) depending on the exponent α of the word frequency distribution shown for two different network sizes $N = 3000$ (triangles) and $N = 5000$ (circles). The networks are constructed as described in the text. Each data point represents an average over 10 independent configurations.*

path length and values of assortativeness very close to the values for comparable Erdös-Rényi random graphs. However, from $\alpha \approx .6$ onwards the network properties, change dramatically. Typical networks begin to exhibit high clustering, shortened path length and substantial dissortative mixing. We also note that for $\alpha = 1$ the values of all three quantities are found in a reasonable range of the values obtained for the texts.

The above observation shows that important properties of linguistic networks: A power law tail, small world structure and the high degree of dissortative mixing, can all be obtained from our random null model.

7. Discussion

In this paper we have analysed the syntactic structures of linguistic networks of a number of well known English authors. In agreement with previous studies (Ferrer i Cancho & Solé 2001; Ferrer i Cancho et al. 2004) we have been able to identify a number of characteristics that appear to be universal. The use of a relative-distance based clustering measure did not lead to statistically significant clusters that could be attributed to authors. However our results do suggest that a number of authors do produce "consistent" patterns in their syntactic networks, possibly associated with changes in the author's writing style. It appears an interesting line of research to classify these patterns using machine learning techniques.

Previous studies have demonstrated that dependency grammar networks are highly clustered, have short path lengths, and display dissortative mixing. One of the more surprising results presented here is that many of these network properties deviate only slightly from such calculated for randomly assembled texts. Generating random texts from a word frequency that follows a power law is sufficient to obtain networks that are small worlds and exhibit a degree distribution with a scale-free tail.

Accordingly, in the final section of this paper we analysed a simple null model that examines the relationship between the word frequency distribution and the emergence of the various net-

work properties. This model confirms that network properties similar to those commonly found in dependency grammar networks emerge when the decay of the word frequency distribution is described by a power law with an exponent close to -1.

Understanding the origins and evolution of language requires an understanding of the underlying dynamics. Recent studies have explored this question, through the use of mathematical models of evolutionary dynamics. These studies lend strong support to the explanation that a word frequency distribution following Zipf's law emerges as a result of a trade-off between the effort to send a message and its information content (Ferrer i Cancho & Solé 2003). However, other works (Li 1992) demonstrated that Zipf's-law-like degree distributions is also obtained for random collections of symbols.

The present study indicates, that the small world and scale-free character of linguistic networks merely seems a statistical feature, that in turn is a consequence of a lower level statistical regularity expressed in Zipf's law. However there are small systematic deviation from the original networks. Only very few syntactic rules may allow one to create meaningful language. Combining this result with the above explanation for Zipf's law, two mutually exclusive interpretations appear possible. First, following (Li 1992) the regularities of the complex network formed by our language do not appear special. Indeed, it would also be expected for random sequences of symbols.

More interestingly, following the line of thoughts presented in (Ferrer i Cancho & Solé 2003) one can speculate about an optimization process shaping the network structure of the linguistic network (i.e. need for efficient communication, flexibility in language, etc.). By changing the network structure, also the type of word frequency distribution that can produce such a network is constrained. It appears an interesting prospect for a future study to explore the relation between optimization processes shaping network topology and consequent effects on the word frequency distributions.

References

Albert, R. & Barabási, A.-L. (1999). Emergence of scaling in random networks. Science, **286**, 509–512.

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. Rev. Mod. Phys, **74**, 47–97.

Bollobás, B. (1998). *Modern graph theory*. Graduate Texts in Mathematics Vol. 184, Springer, New York.

Chomsky, N. (1965). *Aspects of the Theory of Syntax Chomsky*, MIT Press, MA.

Chomsky, N. (2002). *On nature and language*, Cambridge University Press, Cambridge.

Erdös, P., (1959). On random graphs. Publ. Math. Debrecen, **6**, 290-297.

Ferrer i Cancho, R. & Solé R. V., (2001). The small world of human language. Proc. Roy. Soc. Lond. B, **268**, 2261-2265.

Ferrer i Cancho, R. & Solé R. V. (2003) Least effort and the origins of scaling in human language. Proc. Nat. Accad. Sci., **100**(3), 788–791.

Ferrer i Cancho, R., Solé R. V. & Köhler R. (2004). Patterns in syntactic depencency networks. Phys. Rev. E, **69**, 051915.

Gell-mann, M. (1994). *The Quark and the Jaguar*. Abacus Books.

Cohen, R. & Havlin, S. (2003). Scale-free networks are ultrasmall. Phys. Rev. Lett. **90**, 058701.

Kennedy, R.L., Lee, Y., Van Roy, B., Reed, C.D., Lippmann, R.P. (1995). *Solving data mining problems through pattern recognition*, Prentice Hall, New Jersey.

Li W. (1992). Random texts exhibit Zipf's-law-like word frequency distributions. IEEE Trans. Info. Theory, **38**(6), 1842–1845.

Newman, M. E. J. (2002). Assortative mixing in networks. Phys. Rev. Lett, **89**, 208701.

Newman, M. E. J. (2003). Mixing patterns in networks. Phys. Rev. E, **67**, 026126.

Pinker, S. (1997) *The language instinct* . Penguin.

Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of 'small world' networks. Nature, **393**, 440–442.

Zipf, G. (1932). Selective studies on the principle of relative frequency in language. Cambridge, MA; MIT Press.

Appendix 1

Author	Title	N	L	N/L^2	$\bar{l}$	l_{max}	$\bar{c}$	Γ	γ	γ_{in}	γ_{out}
Alcott	1	6791	81740	0.0018	2.79	8	0.49	-0.25	-1.75	-2.19	-1.90
	2	6642	82696	0.0019	2.75	7	0.49	-0.23	-1.52	-2.17	-1.97
	3	7741	101967	0.0017	2.78	16	0.51	-0.23	-1.75	-2.10	-2.08
	4	8555	113822	0.0016	2.75	6	0.50	-0.23	-1.70	-1.94	-1.87
	5	7791	109904	0.0018	2.73	7	0.52	-0.24	-1.53	-1.92	-1.81
	6	7409	92786	0.0017	2.78	7	0.50	-0.24	-1.68	-2.23	-2.07
Annon.	7	3608	34863	0.0027	2.87	6	0.39	-0.20	-1.89	-2.27	-1.72
Austen	8	7316	135625	0.0025	2.67	8	0.58	-0.29	-1.61	-1.60	-1.69
	9	2923	28807	0.0034	2.82	5	0.47	-0.28	-1.76	-2.35	-1.85
	10	4214	41877	0.0024	2.83	9	0.47	-0.27	-2.02	-2.31	-1.56
	11	8025	140429	0.0022	2.66	6	0.58	-0.28	-1.44	-1.69	-1.51
	12	6081	82084	0.0022	2.73	6	0.52	-0.27	-1.62	-2.04	-2.04
	13	6380	111175	0.0027	2.66	7	0.56	-0.30	-1.48	-1.69	-1.74
	14	5824	85038	0.0025	2.71	7	0.54	-0.27	-1.53	-1.81	-2.20
	15	6388	109901	0.0027	2.66	5	0.56	-0.30	-1.63	-1.63	-1.67
Conan-Doyle	16	6113	64009	0.0017	2.79	7	0.53	-0.28	-1.86	-2.22	-2.04
	17	8269	88229	0.0013	2.78	8	0.52	-0.24	-1.81	-2.27	-2.20
	18	7986	103460	0.0016	2.74	7	0.54	-0.28	-1.55	-2.09	-1.91
	19	4882	48341	0.0020	2.83	6	0.49	-0.26	-2.22	-1.88	-1.75
	20	5619	64341	0.0020	2.77	6	0.52	-0.27	-1.77	-2.11	-2.05
	21	11970	192302	0.0013	2.64	8	0.55	-0.21	-1.55	-1.81	-1.65
	22	7803	86633	0.0014	2.79	9	0.51	-0.25	-1.84	-2.09	-1.91
	23	8221	87991	0.0013	2.78	7	0.51	-0.24	-1.93	-2.10	-1.99
	24	5026	55140	0.0022	2.77	7	0.54	-0.27	-1.74	-2.24	-2.03

Table 2. *Texts and Network Properties. (1) Eight Cousins, (2) A Garland for Girls, (3) Jack And Jill, (4) Jo's Boys, (5) Little Men, (6) Little Women, (6) Under the Lilacs, (7) Boewulf, (8) Emma, (9) Lady Susan, (10) Love And Friendship, (11) Mansfield Park, (12) Northanger Abbey, (13) Pride And Prejudice, (14) Persuasion, (15) Sense and Sensibility, (16) Uncle Bernac, (17) The Last Galley Impressions and Tales Impressions and Tales,(18) The Adventures of Sherlock Holmes (19) Beyond the City, (20) The Hound Of The Baskervilles, (21) Great Boer War, (22) The Lost World, (23) The Green Flag, (24) The Great Shadow and Other Napoleonic Tales.*

Author	Title	N	L	N/L^2	$\bar{l}$	l_{max}	$\bar{c}$	Γ	γ	γ_{in}	γ_{out}
Darwin	25	11566	175394	0.0013	2.85	8	0.46	-0.20	-1.59	-1.60	-1.62
	26	12192	177192	0.0012	2.90	18	0.46	-0.20	-1.74	-1.76	-1.89
	27	5523	75780	0.0025	2.80	11	0.47	-0.21	-1.64	-2.09	-1.64
	28	3441	48939	0.0041	2.74	10	0.44	-0.21	-1.52	-1.94	-1.88
	29	14532	211770	0.0010	2.43	13	0.49	-0.20	-1.54	-1.71	-1.72
	30	6411	108449	0.0026	2.80	10	0.50	-0.22	-1.40	-1.62	-1.79
	31	6983	122480	0.0025	2.64	7	0.52	-0.24	-1.55	-1.68	-1.70
	32	6797	126069	0.0027	2.79	12	0.49	-0.21	-1.52	-1.71	-1.58
	33	6457	112565	0.0027	2.72	11	0.47	-0.20	-1.41	-1.78	-1.77
	34	12482	195784	0.0013	2.70	11	0.52	-0.21	-1.57	-1.89	-1.57
	35	4717	59339	0.0027	2.76	10	0.44	-0.21	-1.50	-1.78	-1.71
Defoe	36	6563	97124	0.0023	2.67	6	0.55	-0.24	-1.59	-1.76	-1.89
	37	6128	94193	0.0025	2.69	7	0.58	-0.28	-1.51	-2.06	-2.01
	38	2173	15042	0.0032	2.93	11	0.41	-0.24	-2.09	-2.12	-1.95
	39	1754	11589	0.0038	3.00	6	0.37	-0.23	-1.93	-2.31	-2.15
	40	5096	56830	0.0022	2.76	9	0.49	-0.25	-1.81	-2.00	-1.99
	41	5989	84352	0.0024	2.73	10	0.56	-0.26	-1.59	-2.12	-1.79
	42	4216	41584	0.0023	2.80	13	0.50	-0.25	-2.22	-2.04	-1.80
	43	5924	88234	0.0025	2.67	6	0.58	-0.28	-1.66	-1.96	-1.84
	44	6087	97844	0.0026	2.63	6	0.59	-0.29	-1.57	-2.09	-1.96
Dickens	45	9815	135037	0.0014	2.71	6	0.56	-0.25	-1.84	-1.93	-1.88
	46	9328	145254	0.0017	2.63	7	0.56	-0.24	-1.45	-2.00	-1.88
	47	10431	118464	0.0011	2.75	7	0.51	-0.23	-1.89	-2.17	-2.02
	48	4233	38956	0.0022	2.86	7	0.46	-0.24	-1.89	-2.26	-1.98
	49	15166	280918	0.0012	2.67	7	0.59	-0.26	-1.63	-1.85	-1.79
	50	4300	37493	0.0020	2.88	7	0.46	-0.25	-1.96	-1.87	-1.88
	51	14235	268392	0.0013	2.65	7	0.61	-0.27	-1.58	-1.86	-1.65
	52	10911	161667	0.0014	2.71	6	0.59	-0.28	-1.75	-1.83	-1.72
	53	8863	108667	0.0014	2.79	6	0.54	-0.27	-1.71	-1.99	-2.00
	54	13067	228578	0.0013	2.65	6	0.56	-0.24	-1.55	-1.91	-1.84
	55	4259	39768	0.0022	2.87	9	0.47	-0.24	-1.73	-2.14	-2.03
	56	4288	40241	0.0022	2.86	7	0.48	-0.25	-1.83	-2.26	-2.01
Joyce	57	796	4261	0.0067	3.20	7	0.30	-0.17	-1.93		-1.52
	58	7327	77067	0.0014	2.80	7	0.50	-0.25	-1.97	-1.85	-1.91
	59	9029	94639	0.0012	2.79	8	0.51	-0.23	-2.00	-1.86	-1.82
	60	29561	309556	0.0004	2.98	20	0.47	-0.20	-1.78	-1.95	-2.07

Table 3. *Texts and Network Properties (cont.).* *, (25) The Variation Of Animals And Plants Under Domestication – Volume 1, (26) The Variation Of Animals And Plants Under Domestication – Volume 2, (27) Coral Reefs, (28) The Movements And Habits Of Climbing Plants, (29) The Descent Of Man, (30) Insectivorous Plants, (31) On the Origin of Species, (32) The Power Of Movement In Plants, (33) Geological Observations On South America, (34) The Voyage Of The Beagle, (35) Volcanic Islands,(36) Memoirs of a Cavalier, (37) The Life, Adventures & Piracies of the Famous Captain Singleton, (38) Dickory Cronke, (39) Everybody's Business Is Nobody's Business, (40) An Essay Upon Projects, (41) A Journal Of The Plague Year, (42) From London To Land's End, (43) Robinson Crusoe, (44) The Adventures of Robinson Crusoe, (45) A Tale of Two Cities, (46) A Child's History Of England, (47) American Notes, (48) The Battle of Life, (49) Bleak House, (50) A Christmas Carol, (51) David Copperfield, (52) Great Expectations, (53) Hard Times (54)Barnaby Rudge: a tale of the Riots of 'eighty, (55) The Chimes, (56) The Cricket on the Hearth.*

Author	Title	N	L	N/L^2	$\bar{l}$	l_{max}	$\bar{c}$	Γ	γ	γ_{in}	γ_{out}
Lawrence	61	6608	74495	0.0017	2.80	7	0.50	-0.25	-1.94	-1.90	-1.79
	62	6730	67815	0.0015	2.83	9	0.49	-0.22	-1.67	-2.32	-2.15
	63	9504	146016	0.0016	2.73	9	0.56	-0.27	-1.68	-1.98	-1.64
	64	2845	28900	0.0036	2.92	9	0.44	-0.25	-2.16	-1.47	-1.57
	65	11227	167948	0.0013	2.76	8	0.56	-0.25	-1.50	-1.83	-1.71
Malthus	66	1595	12446	0.0049	2.84	8	0.41	-0.25	-1.72	-2.08	-2.19
	67	1761	15358	0.0050	2.94	11	0.42	-0.22	-1.90	-2.42	-2.33
	68	4849	55891	0.0024	2.72	10	0.49	-0.24	-1.76	-1.97	-1.81
Scott	69	13386	187966	0.0010	2.78	14	0.53	-0.23	-1.60	-1.77	-1.88
	70	7632	74848	0.0013	2.81	6	0.48	-0.24	-1.90	-2.40	-2.23
	71	11607	139711	0.0010	2.75	10	0.52	-0.23	-1.78	-2.07	-1.87
	72	10259	105768	0.0010	2.80	13	0.50	-0.23	-1.77	-2.18	-2.15
	73	13825	186556	0.0010	2.75	12	0.52	-0.23	-1.57	-1.86	-1.73
	74	13441	199680	0.0011	2.69	12	0.53	-0.22	-1.68	-1.58	-1.65
	75	15822	235789	0.0009	2.71	10	0.54	-0.23	-1.65	-1.80	-1.67
	76	15014	204921	0.0009	2.73	8	0.53	-0.23	-1.64	-1.92	-1.75
Trollope	77	7828	98287	0.0016	2.79	16	0.56	-0.27	-1.78	-2.06	-2.08
	78	3299	23719	0.0022	2.99	11	0.43	-0.24	-2.12	-2.01	-2.13
	79	1937	16274	0.0043	2.83	6	0.46	-0.27	-1.58	-2.55	-2.17
	80	1807	16272	0.0050	2.84	5	0.44	-0.27	-1.86	-1.85	-1.55
	81	8748	157468	0.0021	2.68	7	0.62	-0.31	-1.64	-1.69	-1.63
	82	1588	13733	0.0054	2.88	6	0.43	-0.27	-1.70	-2.11	-2.17
	83	8510	153907	0.0021	2.67	6	0.61	-0.30	-1.57	-1.70	-1.69
	84	9991	190080	0.0019	2.67	6	0.62	-0.29	-1.55	-1.82	-1.55
	85	2062	16088	0.0038	2.93	6	0.42	-0.25	-1.87	-2.11	-1.80
	86	8754	155735	0.0020	2.65	6	0.61	-0.30	-1.49	-1.87	-1.61

Table 4. *Texts and Network Properties (cont.)* *(69) The Abbot, (70) The Black Dwarf, (71) Bride Of Lammermoor, (72) Chronicles Of The Canongate, (73) Guy Mannering, (74) Ivanhoe, (75) Old Mortality, (76) Rob Roy, (77) Autobiography of Anthony Trollope, (79) Aaron Trow, (80) The Courtship Of Susan Bell, (81) The Duke's Children, (82) Harry Heathcote Of Gangoil, (83) The American Senator, (84) The Eustace Diamonds, (85) An Unprotected Female, (86) John Caldigate.*