Global topology of word co-occurrence networks: Beyond the two-regime power-law

Monojit Choudhury
Microsoft Research Lab India
monojitc@microsoft.com

Diptesh Chatterjee
Indian Institute of Technology Kharagpur
diptesh.chh.1987@gmail.com

Animesh Mukherjee
Complex Systems Lagrange Lab, ISI Foundation
animesh.mukherjee@isi.it

Abstract

Word co-occurrence networks are one of the most common linguistic networks studied in the past and they are known to exhibit several interesting topological characteristics. In this article, we investigate the global topological properties of word co-occurrence networks and, in particular, present a detailed study of their spectrum. Our experiments reveal certain universal trends found across the networks for seven different languages from three different language families, which are neither reported nor explained by any of the previous studies and models of word-cooccurrence networks. We hypothesize that since word co-occurrences are governed by syntactic properties of a language, the network has much constrained topology than that predicted by the previously proposed growth model. A deeper empirical and theoretical investigation into the evolution of these networks further suggests that they have a core-periphery structure, where the core hardly evolves with time and new words are only attached to the periphery of the network. These properties are fundamental to the nature of word co-occurrence across languages.

1 Introduction

In a natural language, words interact among themselves in different ways – some words co-occur with certain words at a very high probability than other words. These co-occurrences are non-trivial, as in their patterns cannot be inferred from the frequency distribution of the individual words. Understanding the structure and the emergence of these patterns can present us with important clues and insights about how we evolved this extremely complex phenomenon, that is language.

In this paper, we present an in-depth study of the word co-occurrence patterns of a language in the framework of complex networks. The choice of this framework is strongly motivated by its success in explaining various properties of word co-occurrences previously (Ferrer-i-Cancho and Solé, 2001; Ferrer-i-Cancho et al, 2007; Kapustin and Jamsen, 2007). Local properties, such as the degree distribution and clustering coefficient of the word co-occurrence networks, have been thoroughly studied for a few languages (Ferrer-i-Cancho and Solé, 2001; Ferrer-i-Cancho et al, 2007; Kapustin and Jamsen, 2007) and many interesting conclusions have been drawn. For instance, it has been found that these networks are small-world in nature and are characterized by a *two regime power-law* degree distribution. Efforts have also been made to explain the emergence of such a two regime degree distribution through network growth models (Dorogovstev and Mendes, 2001). Although it is tempting to believe that a lot is known about word co-occurrences, in order to obtain a deeper insight into how these co-occurrence patterns emerged there are many other interesting properties that need to be investigated. One such property is the *spectrum* of the word co-

Coling 2008: Poster Volume, pages 162–170,
Beijing, August 2010

occurrence network which can provide important information about its global organization. In fact, the application of this powerful mathematical machinery to infer global patterns in linguistic networks is rarely found in the literature (few exceptions are (Belkin and Goldsmith, 2002; Mukherjee et al, 2009)). However, note that spectral analysis has been quite successfully applied in the analysis of biological and social networks (Banerjee and Jost, 2007; Farkas et al, 2001).

The aim of the present work is to investigate the spectral properties of a word co-occurrence network in order to understand its global structure. In particular, we study the properties of seven different languages namely Bangla (Indo-European family), English (Indo-European family), Estonian (Finno-Ugric family), French (Indo-European family), German (Indo-European family), Hindi (Indo-European family) and Tamil (Dravidian family). Quite importantly, as we shall see, the most popular growth model proposed by Dorogovtsev and Mendes (DM) (Dorogovstev and Mendes, 2001) for explaining the degree distribution of such a network is not adequate to reproduce the spectrum of the network. This observation holds for all the seven different languages under investigation. We shall further attempt to identify the precise (linguistic) reasons behind this difference in the spectrum of the empirical network and the one reproduced by the model. Finally, as an additional objective, we shall present a hitherto unreported deeper analysis of this popular model and show how its most important parameter is correlated to the size of the corpus from which the empirical network is constructed.

The rest of the paper is laid out as follows. In section 2, we shall present a brief review of the previous works on word co-occurrence networks. This is followed by a short primer to spectral analysis. In section 4, we outline the construction methodology of the word co-occurrence networks and present the experiments comparing the spectrum of these real networks with those generated by the DM model. Section 5 shows how the most important parameter of the DM model varies with the size of the corpus from which the co-occurrence networks are constructed. Finally, we conclude in section 6 by summarizing our contributions and pointing out some of the implications of the current work.

2 Word Co-occurrence Networks

In this section, we present a short review of the earlier works on word co-occurrence networks, where the nodes are the words and an edge between two words indicate that the words have co-occurred in a language in certain context(s). The most basic and well studied form of word co-occurrence networks are the *word collocation networks*, where two words are linked by an edge if they are neighbors (i.e., they collocate) in a sentence (Ferrer-i-Cancho and Solé, 2001).

In (Ferrer-i-Cancho and Solé, 2001), the authors study the properties of two types of collocation networks for English, namely the *unrestricted* and the *restricted* ones. While in the unrestricted network, all the collocation edges are preserved, in the restricted one only those edges are preserved for which the probability of occurrence of the edge is higher than the case when the two words collocate independently. They found that both the networks exhibit small-world properties; while the average path length between any two nodes in these networks is small (between 2 and 3), the clustering coefficients are high (0.69 for the unrestricted and 0.44 for the restricted networks). Nevertheless, the most striking observation about these networks is that the degree distributions follow a two regime power-law. The degree distribution of the 5000 most connected words (i.e., the kernel lexicon) follow a power-law with an exponent -3.07, which is very close to that predicted by the Barabási-Albert growth model (Barabási and Albert, 1999). These findings led the authors to argue that the word usage of the human languages is preferential in nature, where the frequency of a word defines the comprehensibility and production capability. Thus, higher the usage frequency of a word, higher is the probability that the speakers will be able to produce it easily and the listeners will comprehend it fast. This idea is closely related to the *recency effect* in linguistics (Akmajian, 1995).

Properties of word collocation networks have also been studied for languages other than English (Ferrer-i-Cancho et al, 2007; Kapustin and

Jamsen, 2007). The basic topological characteristics of all these networks (e.g., scale-free, small world, assortative) are similar across languages and thus, point to the fact that like Zipf's law, these are also linguistic universals whose emergence and existence call for a non-trivial psycholinguistic account.

In order to explain the two regime power-law in word collocation networks, Dorogovtsev and Mendes (Dorogovstev and Mendes, 2001) proposed a preferential attachment based growth model (henceforth referred to as the DM model). In this model, at every time step t, a new word (i.e., a node) enters the language (i.e., the network) and connects itself preferentially to one of the pre-existing nodes. Simultaneously, ct (where c is a positive constant and a parameter of the model) new edges are grown between pairs of old nodes that are chosen preferentially. Through mathematical analysis and simulations, the authors successfully establish that this model gives rise to a two regime power-law with exponents very close to those observed in (Ferrer-i-Cancho and Solé, 2001). In fact, for English, the values k_{cross} (i.e., the point where the two power law regimes intersect) and k_{cut} (i.e., the point where the degree distribution cuts the x-axis) obtained from the model are in perfect agreement with those observed for the empirical network.

Although the DM model is capable of explaining the local topological properties of the word collocation network, as we shall see in the forthcoming sections, it is unable to reproduce the global properties (e.g., the *spectrum*) of the network.

3 A Primer to Spectral Analysis

Spectral analysis[1] is a powerful mathematical method capable of revealing the global structural patterns underlying an enormous and complicated environment of interacting entities. Essentially, it refers to the systematic investigation of the eigenvalues and the eigenvectors of the adjacency matrix of the network of these interacting entities. In this section, we shall briefly outline the basic concepts involved in spectral analysis and discuss some of its applications (see (Chung, 1994) for details).

A network consisting of n nodes (labeled as 1 through n) can be represented by an $n \times n$ square matrix $\mathbf{A}$, where the entry a_{ij} represents the weight of the edge from node i to node j. Note that $\mathbf{A}$, which is known as the *adjacency matrix*, is symmetric for an undirected graph and have binary entries for an unweighted graph. λ is an eigenvalue of $\mathbf{A}$ if there is an n-dimensional vector $\mathbf{x}$ such that

$$\mathbf{Ax} = \lambda\mathbf{x}$$

Any real symmetric matrix $\mathbf{A}$ has n (possibly non-distinct) eigenvalues $\lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{n-1}$, and corresponding n eigenvectors that are mutually orthogonal. The *spectrum* of a network is the set of the distinct eigenvalues of the graph and their corresponding multiplicities. It is a distribution usually represented in the form of a plot with the eigenvalues in x-axis and their multiplicities in the y-axis.

The spectrum of real and random networks display several interesting properties. Banerjee and Jost (Banerjee and Jost, 2007) report the spectrum of several biological networks and show that these are significantly different from the spectrum of artificially generated networks. It is worthwhile to mention here that spectral analysis is also closely related to *Principal Component Analysis* and *Multidimensional Scaling*. If the first few (say d) eigenvalues of a matrix are much higher than the rest of the eigenvalues, then one can conclude that the rows of the matrix can be approximately represented as linear combinations of d orthogonal vectors. This further implies that the corresponding graph has a few motifs (subgraphs) that are repeated a large number of time to obtain the global structure of the graph (Banerjee and Jost, 2009).

In the next section, we shall present a thorough study of the spectrum of the word co-occurrence networks across various languages.

4 Experiments and Results

For the purpose of our experiments, we construct word collocation networks for seven different languages namely, Bangla, English, Esto-

[1]The term spectral analysis is also used in the context of signal processing, where it refers to the study of the frequency spectrum of a signal.

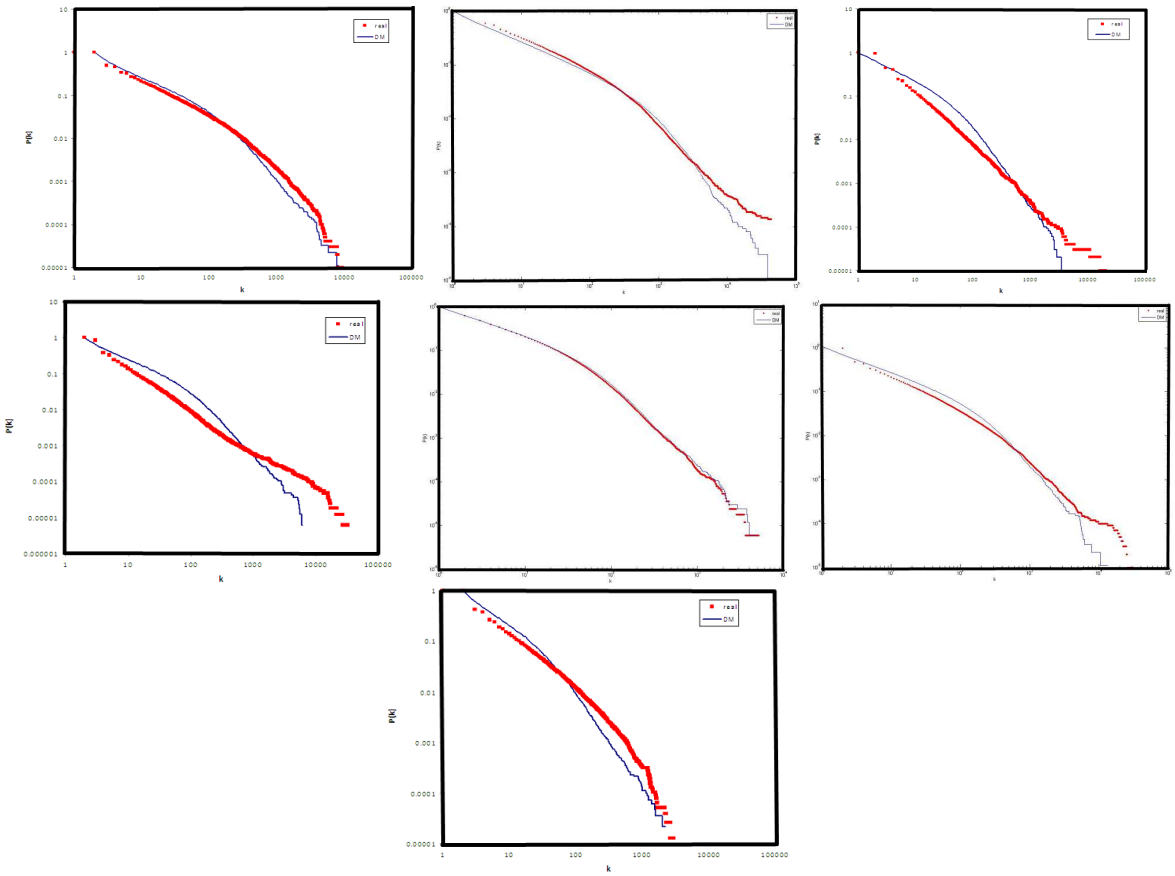

Figure 1: Cumulative degree distributions for Bangla, English, Estonian, French, German, Hindi and Tamil respectively. Each red line signifies the degree distribution for the empirical network while each blue line signifies the one obtained from the DM model.

Lang.	Tokens (Mill.)	Words	KLD	c	Max. Eig. (Real)	Max. Eig. (DM)
English	32.5	97144	0.21	5.0e-4	849.1	756.8
Hindi	20.2	99210	0.32	2.3e-4	472.5	329.5
Bangla	12.7	100000	0.29	2.0e-3	326.2	245.0
German	5.0	159842	0.19	6.3e-5	192.3	110.7
Estonian	4.0	100000	0.25	1.1e-4	158.6	124.0
Tamil	2.3	75929	0.24	9.9e-4	116.4	73.06
French	1.8	100006	0.44	8.0e-5	236.1	170.1

Table 1: Summary of results comparing the structural properties of the empirical networks for the seven languages and the corresponding best fits (in terms of KLD) obtained from the DM model.

nian, French, German, Hindi and Tamil. We used the corpora available in the *Lipezig Corpora Collection* (http://corpora.informatik.uni-leipzig.de/) for English, Estonian, French and German. The Hindi, Bangla and Tamil corpora were collected by crawling some online newspapers. In these networks, each distinct word corresponds to a vertex and two vertices are connected by an edge if the corresponding two words are adjacent in one or more sentences in the corpus. We assume the network to be undirected and unweighted (as in (Ferrer-i-Cancho and Solé, 2001)).

As a following step, we simulate the DM model and reproduce the degree distribution of the collocation networks for the seven languages. We vary the parameter c in order to minimize the KL

divergence (KLD) (Kullback and Leibler, 1951) between the empirical and the synthesized distributions and, thereby, obtain the best match. The results of these experiments are summarized through Figure 1 and Table 1. The results clearly show that the DM model is indeed capable of generating the degree distribution of the collocation networks to a very close approximation for certain values of the parameter c (see Table 1 for the values of c and the corresponding KLD).

Subsequently, for the purpose of spectral analysis, we construct subgraphs induced by the top 5000 nodes for each of the seven empirical networks as well as those generated by the DM model (i.e., those for which the degree distribution fits best in terms of KLD with the real data). We then compute and compare the spectrum of the real and the synthesized networks (see Figure 2 and Table 1). It is quite apparent from these results that the spectra of the empirical networks are significantly different from those obtained using the DM model. In general, the spectral plots indicate that the adjacency matrices for networks obtained from the DM model have a higher rank than those for the empirical networks. Further, in case of the synthesized networks, the first eigenvalue is significantly larger than the second whereas for the empirical networks the top 3 to 4 eigenvalues are found to dominate. Interestingly, this property is observed across all the languages under investigation.

We believe that the difference in the spectra is due to the fact that the ordering of the words in a sentence are strongly governed by the grammar or the syntax of the language. Words belong to a smaller set of lexico-syntactic categories, which are more commonly known as the parts-of-speech (POS). The co-occurrence patterns of the words are influenced, primarily, by its POS category. For instance, *nouns* are typically preceded by *articles* or *adjectives*, whereas *verbs* might be preceded by *auxiliary verbs*, *adverbs* or *nouns*, but never *articles* or *adjectives*. Therefore, the words "car" and "camera" are more likely to be structurally similar in the word co-occurrence network, than "car" and "jumped". In general, the local neighborhoods of the words belonging to a particular POS is expected to be very similar, which means that several rows in the adjacency matrix will be very similar to each other. Thus, the matrix is expected to have low rank.

In fact, this property is not only applicable to syntax, but also semantics. For instance, even though adjectives are typically followed by nouns, semantic constraints make certain adjective-noun co-occurrences (e.g., "green leaves") much more likely than some others (e.g., "green dreams" or "happy leaves"). These notions are at the core of latent semantics and vector space models of semantics (see, for instance, Turney and Pantel (Turney and Pantel, 2010) for a recent study). The DM model, on the other hand, is based on the *recency effect* that says that the words which are produced most recently are easier to remember and therefore, easier to produce in the future. Preferential attachment models the recency effect in word production, which perhaps is sufficient to replicate the degree distribution of the networks. However, the model fails to explain the global properties, precisely because it does not take into account the constraints that govern the distribution of the words.

It is quite well known that the spectrum of a network can be usually obtained by iteratively powering the adjacency matrix of the network (aka power iteration method). Note that if the adjacency matrices of the empirical and the synthesized networks are powered even once (i.e., they are squared)[2], their degree distributions match no longer (see Figure 3). This result further corroborates that although the degree distribution of a word co-occurrence network is quite appropriately reproduced by the DM model, more global structural properties remain unexplained. We believe that word association in human languages is not arbitrary and therefore, a model which accounts for the clustering of words around their POS categories might possibly turn out to present a more accurate explanation of the spectral properties of the co-occurrence networks.

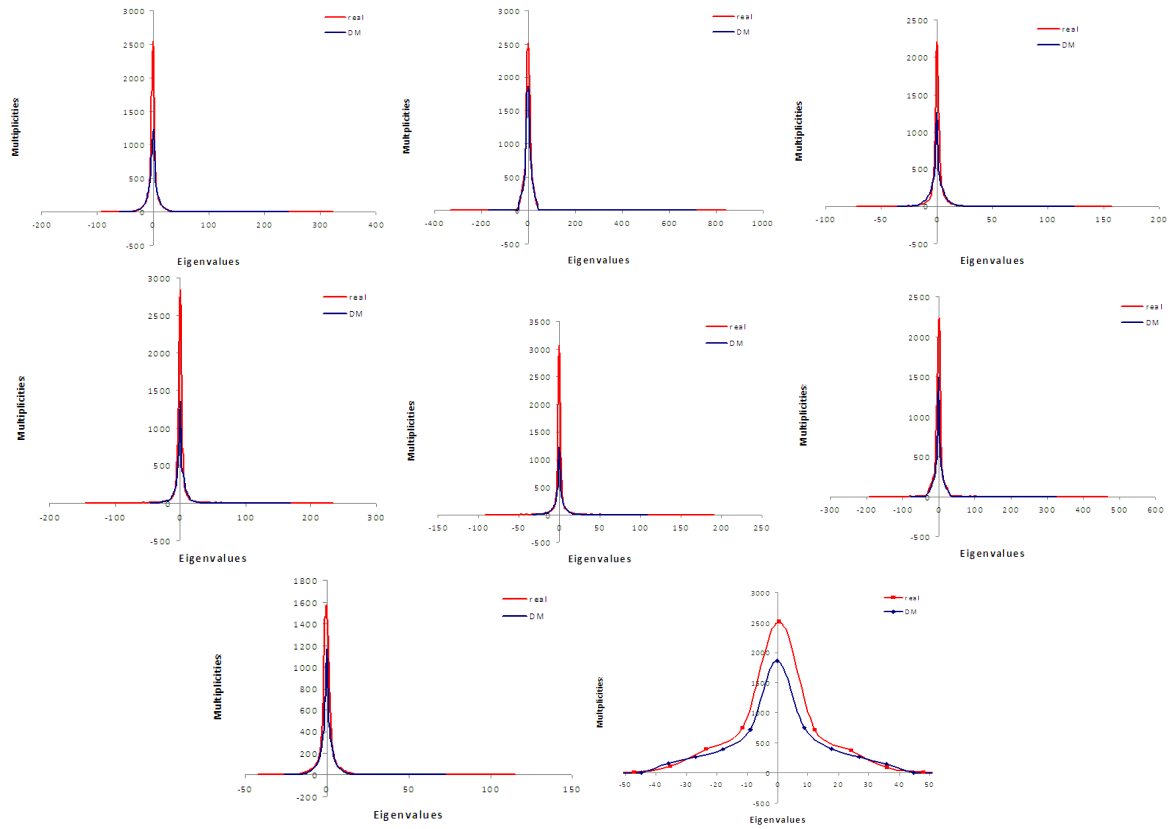

Figure 2: The spectrum for Bangla, English, Estonian, French, German, Hindi and Tamil respectively. The last plot shows a portion of the spectrum for English magnified around 0 for better visualization. All the curves are binned distributions with bin size = 100. The blue line in each case is the spectrum for the network obtained from the DM model while each red line corresponds to the spectrum for the empirical network.

5 Reinvestigating the DM Model

In this section, we shall delve deeper into exploring the properties of the DM model since it is one of the most popular and well accepted models for explaining the emergence of word associations in a language. In particular, we shall investigate the influence of the model parameter c on the emergent results.

If we plot the value of the parameter c (from Table 1) versus the size of the corpora (from Table 1) used to construct the empirical networks for the different languages we find that the two are highly correlated (see Figure 4).

[2]Note that this squared network is weighted in nature. We threshold all edges below the weight 0.07 so that the resultant network is neither too dense nor too sparse. The value of the threshold is chosen based on the inspection of the data.

In order to further check the dependence of c on the corpus size we perform the following experiment. We draw samples of varying corpus size and construct empirical networks from each of them. We then simulate the DM model and attempt to reproduce the degree distribution for each of these empirical networks. In each case, we note the value c for which the KLD between the empirical and the corresponding synthesized network is minimum. Figure 5 shows the result of the above experiment for English. The figure clearly indicates that as the corpus size increases the value of the parameter c decreases. Similar trends are observed for all the other languages.

In general, one can mathematically prove that the parameter c is equal to the rate of change of the average degree of the network with respect to

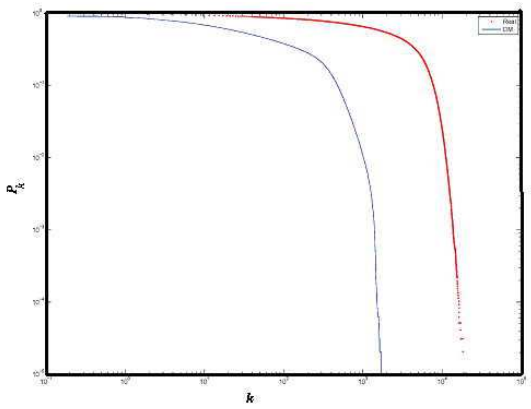

Figure 3: Cumulative degree distribution for the squared version of the networks for English. The red line is the degree distribution for the squared version of the empirical network while the blue line is degree distribution of the squared version of the network obtained from the DM model. The trends are similar for all the other languages.

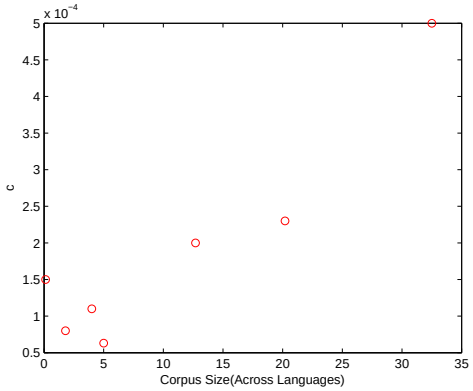

Figure 4: The parameter c versus the corpus size for the seven languages.

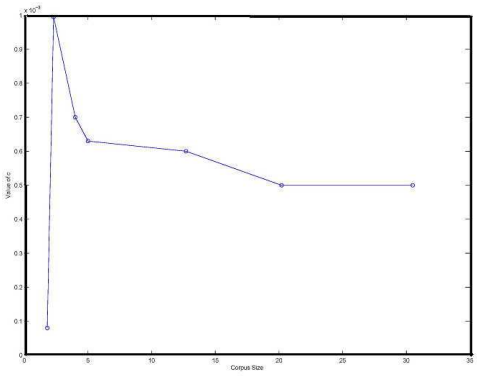

Figure 5: The parameter c versus the corpus size for English.

the time t. The proof is as follows.

At every time step t, the number of new edges formed is (1+ct). Since each edge contributes to a total degree of 2 to the network, the sum of the degrees of all the nodes in the network (k_{tot}) is

$$k_{tot} = 2\sum_{t=1}^{T}(1+ct) = 2T + cT(T+1) \quad (1)$$

At every time step, only one new node is added to the network and therefore the total number of nodes at the end of time T is exactly equal to T. Thus the average degree of the network is

$$\langle k \rangle = \frac{2T + cT(T+1)}{T} = 2 + c(T+1) \quad (2)$$

The rate of change of average degree is

$$\frac{d\langle k \rangle}{dT} = c \quad (3)$$

and this completes the proof.

In fact, it is also possible to make a precise empirical estimate of the value of the parameter c. One can express the average degree of the co-occurrence networks as the ratio of twice the bigram frequency (i.e., twice the number of edges in the network) to the unigram frequency (i.e., the number of nodes or unique words in the network). Therefore, if we can estimate this ratio we can easily estimate the value of c using equation 3. Let us denote the total number of distinct bigrams and unigrams after processing a corpus of size N by $B(N)$ and $W(N)$ respectively. Hence we have

$$\langle k \rangle = \frac{2B(N)}{W(N)} \quad (4)$$

Further, the number of distinct new unigrams after

Language	$B(N)$	$W(N)$	c
English	$29.2N^{.67}$	$59.3N^{.43}$	$.009N^{-.20}$
Hindi	$26.2N^{.66}$	$49.7N^{.46}$	$.009N^{-.26}$
Tamil	$1.9N^{.91}$	$6.4N^{.71}$	$.207N^{-.50}$

Table 2: Summary of expressions for $B(N)$, $W(N)$ and c for English, Hindi and Tamil.

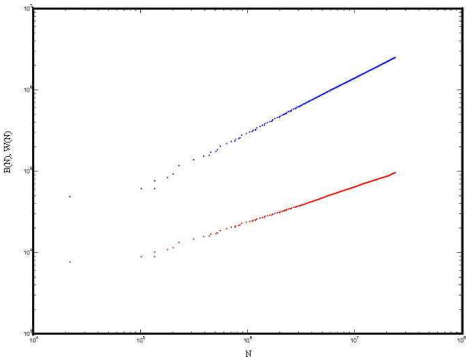

Figure 6: Variation of $B(N)$ and $W(N)$ with N for English (in doubly-logarithmic scale). The blue dots correspond to variation of $B(N)$ while the red dots correspond to the variation of $W(N)$.

processing a corpus of size N is equivalent to T and therefore

$$T = W(N) \tag{5}$$

Sampling experiments across different languages demonstrate that $W(N)$ and $B(N)$ are of the form ηN^{α} ($\alpha < 1$) where η and α are constants. For instance, Figure 6 shows in doubly-logarithmic scale how $B(N)$ and $W(N)$ varies with N for English. The R^2 values obtained as a result of fitting the $B(N)$ versus N and the $W(N)$ versus N plots using equations of the form ηN^{α} for English, Hindi and Tamil are greater than 0.99. This reflects the high accuracy of the fits. Similar trends are observed for all the other languages.

Finally, using equations 3, 4 and 5 we have

$$c = \frac{d\langle k \rangle}{dT} = \frac{d\langle k \rangle}{dN}\frac{dN}{dT} \tag{6}$$

and plugging the values of $B(N)$ and $W(N)$ in equation 6 we find that c has the form $\kappa N^{-\beta}$ ($\beta < 1$) where κ and β are language dependent positive constants. The values of c obtained in this way for three different languages English, Hindi and Tamil are noted in Table 5.

Thus, we find that as $N \rightarrow \infty$, $c \rightarrow 0$. In other words, as the corpus size grows the number of distinct new bigrams goes on decreasing and ultimately reaches (almost) zero for a very large sized corpus. Now, if one plugs in the values of c and T obtained above in the expressions for k_{cross} and k_{cut} in (Dorogovstev and Mendes, 2001), one observes that $\lim_{N\rightarrow\infty} \frac{k_{cross}}{k_{cut}} = 0$. This implies that as the corpus size becomes very large, the two-regime power law (almost) converges to a single regime with an exponent equal to -3 as is exhibited by the Barabási-Albert model (Barabási and Albert, 1999). Therefore, it is reasonable to conclude that although the DM model provides a good explanation of the degree distribution of a word co-occurrence network built from a medium sized corpora, it does not perform well for very small or very large sized corpora.

6 Conclusions

In this paper, we have tried to investigate in detail the co-occurrence properties of words in a language. Some of our important observations are: (a) while the DM model is able to reproduce the degree distributions of the word co-occurrence networks, it is not quite appropriate for explaining the spectrum of these networks; (b) the parameter c in the DM model signifies the rate of change of the average degree of the network with respect to time; and (c) the DM model does not perform well in explaining the degree distribution of a word co-occurrence network when the corpus size is very large.

It is worthwhile to mention here that our analysis of the DM model leads us to a very important observation. As N grows, the value of k_{cut} grows at a much faster rate than the value of k_{cross} and in the limit $N \rightarrow \infty$ the value of k_{cut} is so high as compared to k_{cross} that the ratio $\frac{k_{cross}}{k_{cut}}$ becomes (almost) zero. In other words, the kernel lexicon, formed of the words in the first regime of the two regime power-law and required to "say everything or almost everything" (Ferrer-i-Cancho and Solé, 2001) in a language, grows quite slowly as new words creep into the language. In contrast, the peripheral lexicon making the other part of the two regime grows very fast as new words enter the language. Consequently, it may be argued that since the kernel lexicon remains almost unaffected, the effort to learn and retain a language by its speakers increases only negligibly as new words creep into the language.

References

A. Akmajian. *Linguistics: An introduction to Language and Communication*. MIT Press, Cambridge, MA, 1995.

A. Banerjee and J. Jost. Spectral plots and the representation and interpretation of biological data. *Theory in Biosciences*, 126(1), 15-21, 2007.

A. Banerjee and J. Jost. Graph spectra as a systematic tool in computational biology. *Discrete Applied Mathematics*, 157(10), 2425–2431, 2009.

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

M. Belkin and J. Goldsmith. Using eigenvectors of the bigram graph to infer morpheme identity. In *Proceedings of Morphological and Phonological Learning*, Association for Computational Linguistics, 41-47, 2002.

F. R. K. Chung. *Spectral Graph Theory*. Number 2 in CBMS Regional Conference Series in Mathematics, American Mathematical Society, 1994.

S. N. Dorogovstev and J. F .F. Mendes. Language as an evolving word Web. *Proceedings of the Royal Society of London B*, 268, 2603-2606, 2001.

I. J. Farkas, I. Derényi, A. -L. Barabási and T. Vicsek. Spectra of "real-world" graphs: Beyond the semicircle law, *Physical Review E*, 64, 026704, 2001.

R. Ferrer-i-Cancho and R. V. Solé. The small-world of human language. *Proceedings of the Royal Society of London B*, 268, 2261–2266, 2001.

R. Ferrer-i-Cancho, A. Mehler, O. Pustylnikov and A. Díaz-Guilera. Correlations in the organization of large-scale syntactic dependency networks. In *Proceedings of TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, 65-72, Association for Computational Linguistics, 2007.

V. Kapustin and A. Jamsen. Vertex degree distribution for the graph of word co-occurrences in Russian. In *Proceedings of TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, 89-92, Association for Computational Linguistics, 2007.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics* 22(1), 79-86, 1951.

A. Mukerjee, M. Choudhury and R. Kannan. Discovering global patterns in linguistic networks through spectral analysis: A case study of the consonant inventories. In *Proceedings of EACL*, 585–593, Association for Computational Linguistics, 2009.

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. In *JAIR*, 37, 141-188, 2010.