A Corpus Linguistic Study of Bollywood Song Lyrics in the Framework of Complex Network Theory

Aseem Behl
IIIT-Hyderabad
Hyderabad, India 500032
aseem.behl@students.iiit.ac.in

Monojit Choudhury
Microsoft Research India
Bangalore, India 560080
monojitc@microsoft.com

Abstract

Every year the Mumbai-based Hindi movie industry, *Bollywood*, produces several hundred movies and associated thousands of songs that collectively reflect the state and evolution of popular culture and language usage in India over last eight decades. In this paper, we report a corpus linguistic study of Bollywood songs conducted through standard statistical methods as well as the more recent techniques of complex network analysis. Our study reveals several interesting features of Bollywood lyrics, some of which are strikingly similar to and some remarkably different from those of the standard natural language text corpora of Hindi.

1 Introduction

Bollywood[1] is the term popularly used for the Hindi-language film industry based in Mumbai, India. More than 800 movies are produced out of Bollywood every year[2]. Almost all Bollywood movies feature several songs that are very popular not only in India, but across the globe. In fact, Bollywood songs are one of the most searched items on the Web from India[3].

Bollywood songs are generally composed in Hindi, though often the vocabulary and language usage deviate from that of standard Hindi. Extensive use of Urdu and Persian words, compositions in various dialects of Hindi (e.g., *Braj* , *Rajasthani*, *Maithili*) and Punjabi songs have always been a part and parcel of Bollywood lyrics. In recent times, English-Hindi code-mixing is on rise in Bollywood lyrics - a reflection of the ongoing socio-cultural changes in India. Bollywood songs are written for diverse themes such as emotional relationships (of love, betrayal, friendship, parent-child, brother-sister, etc.), social events and functions (e.g., marriage, birthday, anniversary), festivals and rituals, devotion, songs for children and so on.

In this work, for the first time, we present a corpus linguistic analysis of Bollywood song lyrics. We crawled the Web to construct a Bollywood lyrics corpus, computed various corpus statistics, and compared them against those for standard text corpus of Hindi. We constructed network models of Bollywood lyrics and standard text corpus, and further compared the topological properties of these networks to understand how similar or different is the language of Bollywood lyrics from that of standard Hindi text; and if they are different, then what are the factors giving rise to these differences.

Use of Complex Network Theory (CNT) for this work is motivated by the recent developments in the study of linguistic networks (see (Choudhury and Mukherjee, 2009) for a review). CNT has provided a suitable framework for modeling and studying complex adaptive systems occurring in disciplines as diverse as biology, sociology, physics, economics and technology. In corpus linguistics, network models have been used to explore and explain universal properties as well as unique distinctive features of languages. Our adoption of this framework is strongly driven by previous successful studies of word co-occurrence networks (Cancho and Solé, 2001a).

[1] http://en.wikipedia.org/wiki/Bollywood

[2] http://geography.about.com/od/culturalgeography/a/bollywood.htm

[3] http://www.google.com/intl/en/press/zeitgeist2010/regions/in.html

Proceedings of ICON-2011: 9^{th} *International Conference on Natural Language Processing*
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2011

The rest of this paper is organized as follows. Sec. 2 gives a brief overview of previous work on complex network theory as a tool to understand linguistic structure. Sec. 3 describes the lyrics corpus creation and some basic statistics of the corpus, including word frequency distributions. Sec. 4 introduces the network models and their topological properties and corresponding linguistic interpretations. Sec. 5 summarizes the salient facts about Bollywood lyrics revealed through our analysis, provides plausible sociocultural explanations for some of these observations, and lists some interesting research questions for future work.

2 Corpus Linguistics and Complex Networks

Corpus linguistics approaches the study of natural language through corpora and deals with finding patterns associated with the lexical or grammatical features of a language (Bennett, 2010). Statistical properties such as (a) the frequency of linguistic elements such as morphemes, words and phrases occurring in a corpus, (b) the distribution of probability mass across these linguistic elements and (c) the frequencies of co-occurrence of the linguistic elements with each other are typically used to characterize a corpus (Gries, 2010).

Recently, *Complex Networks Theory* has been introduced as a new tool to model and study the statistical properties of a corpus (Mehler, 2008). *Complex network* refers to a system of *nodes* denoting physical or abstract entities and a set of *edges* connecting them, which usually represent certain interactions between the entities. CNT has been successfully used to model the structure and dynamics of a broad sphere of complex adaptive systems in nature and society; popular examples include the cell, the Internet, or a network of computers. For review of complex networks see (Albert and Barabási, 2002; Newman, 2003). In linguistics, network models have been used to study linguistic entities and their interactions (Choudhury and Mukherjee, 2009); besides applications in NLP, such studies have unearthed fascinating linguistic universals and helped us understand their origin and emergence.

Of special interest to us are the studies that model a corpus as a complex network. Cancho and Solé (2001a) introduced the concept of *Word Collocation Networks*, where the nodes represent unique words and edges connect nodes that co-occur in a sentence close to each other. Through empirical studies of the British National Corpus, the authors showed that these networks exhibit small-world property similar to social network of people that is necessary for fast access of words in the mental lexicon; the networks also feature two-regime power-law which they argued was due to the existence of distinct kernel and peripheral lexicon in a language. Since then word collocation networks have been used to model corpora in various languages (see for example (Choudhury et al., 2010)) and various other kinds of datasets, such as query logs (Roy et al., 2011) and Indus valley symbols (Sinha et al., 2009) to establish linguistic nature of apparently non-linguistic datasets. As Mehler (2008) has rightly pointed out, "development of complex text networks provides knowledge about constraints of the "nonartificiality" of text corpora by analogy with Zipf's law in the case of lexical systems". The present study takes inspirations and ideas from this rich and fascinating body of research work.

3 Bollywood Lyrics Corpus

There is no known corpus or database exclusively built for Bollywood song lyrics. However, being a very popular entertainment industry and one of the most frequently searched item on the web, there are several managed or user-contributed websites available for Bollywood lyrics. These websites can be readily crawled to build a Bollywood song corpus. In most of these websites, such as *hindilyrix.com* and *lyricsmasti.com*, the lyrics are archived in a non-standard Romanized form instead of *Devanagari* - the script used for writing Hindi. Non-standard Romanization of Hindi words leads to high degree of spelling variation in the lyrics that are crawled from two different websites. For instance, the song वो पहली बार जब हम मिले (movie: *Pyaar Mein Kabhi Kabhi*) is present in one website as *Woh Pehli Bar Jab Hum Mile* and in another one as *Wo Pahli Baar Jab Ham Mile*. While users can easily read and understand Romanized lyrics, such representations cannot be used for a corpus linguistic analysis of songs, unless the spellings are normalized to a standard Roman form or transliterated back to Devanagari. Both of these approaches, however, would require a non-trivial amount of research and experimental efforts and can induce several errors during the

process. In order to avoid this problem, we restricted ourselves to crawling only those websites where the lyrics have been archived in Devanagari. Since most of the popular websites and consequently the bulk of the song lyrics on the Web are present only in Romanized form, we did miss out a large number of potential songs that could have enriched the corpus and the subsequent analysis.

There are quite a few websites that archive lyrics in Devenagari for songs ranging from a few hundreds to several thousands. We identified three websites, *smriti.com, lyricsindia.net* and *10lyrics.com*, which together cover almost all the songs for which Devenagari lyrics is available on the Web. Note that the same song might be present in more than one website, and sometimes multiple versions of a lyrics might be present even in the same website (e.g., lyrics of the original song, and a remix version of the song, which are often identical). Therefore, we identified duplicate copies of the lyrics by a simple word for word comparison to ensure that only a single version of a song is included in our corpus. This resulted in a corpus with 6529 song lyrics and 0.5 million words in Devanagari script. Table 1 reports some basic statistics of the corpus.

Along with the lyrics, the websites also contain several related facts such as the composer, lyricist, movie-title and year of release of the song. These information has also been crawled, and stored as metadata. Amongst these, the date of release is of special interest because it can help us analyse and understand the patterns of evolution of Bollywood lyrics. Such an analysis is beyond the scope of the present work.

Table 1: Corpus Statistics

Number of songs	6529
Number of distinct movies	2634
Tokens	491K
Types	17.9K
Types after morphological analysis	12.3K

3.1 Morphological Analysis

Hindi morphology is rich and productive. Therefore, a high variation in *types* in a Hindi corpus might result from morphological richness rather than diversity of topics. While song lyrics may not sport very diverse themes and vocabulary, due to the presence of morphological forms of only a few distinct roots, the corpus might still contain a large number of distinct words. In order to understand and separate out such effects, we conducted all our analyses on the original corpus where each distinct surface form is treated as a distinct type, and also on a stemmed version of the same corpus, where only the root forms and not their morphological variations are considered as distinct types. We used Hindi morphological analyser developed by Akshara Bharathi Group to derive the morphological roots of the words (Bharati et al., 1995), which is approximately 95% accurate. As shown in Table 1, for 17.9K distinct words there are 12.3K distinct roots, implying that on an average a root has 1.5 variations.

3.2 Standard Text Corpus of Hindi

In order to understand how, if at all, the corpus characteristics of Bollywood lyrics differ from that of standard Hindi text, one should repeat the corpus analysis studies on standard Hindi text corpora of similar size and diversity. For this purpose we sampled a general and two restricted domain corpora, each containing 0.5 million words from the *HC Corpora*[4]. The HC Corpora contain Hindi language texts collected from publicly accessible sources like Hindi news websites. We utilized the topic labels available in HC Corpora to build our domain restricted corpora for the domains of *Business & economics* and *Religion, spirituality & astrology*. One would anticipate that the characteristics of the domain specific corpora will match that of the Bollywood lyrics corpora more closely than those of the general corpus because the vocabulary of Bollywood lyrics is rather restricted to only a certain kind of emotional and socio-cultural themes.

3.3 Frequency distribution of words

One of the most popular and universally observed characteristics of any natural language corpus is *Zipf's Law*(Zipf, 1949), which says that if the words in the corpus are sorted by their frequency of occurrence, then the rank of a word in this sorted list is inversly proportionate to its frequency. In other words, if f is the frequency of a word and r is its rank, then

$$f = Ar^{-\gamma}$$

[4] http://www.corpora.heliohost.org/

where, γ is close to 1 for all languages, and A is some positive constant of proportionality. There have been several empirical studies on Zipfian distribution of natural language text (see `http://www.nslij-genetics.org/wli/zipf/` for comprehensive bibliography on Zipf's law.), including some for Indian language corpora (Jayaram and Vidya, 2008). Furthermore, there have been various mathematical investigations and modeling studies to understand the origin of Zipfian distribution in natural language text (Cancho and Solé, 2002; Kornai, 2002).

Figure 1 plots the rank-frequency distribution of words and their morphological roots for the Bollywood Lyrics corpus in a doubly logarithmic plot. According to the Zipf's law, this plot should be a straight line with slope close to 1 (usually between 0.9 and 1.1 for all languages), and that is precisely the case for the standard Hindi text corpus and the two domain specific corpora (plots omitted due to paucity of space, but reader can refer to (Jayaram and Vidya, 2008) for similar analysis for Hindi). However, to our surprise the rank-frequency distribution of the words for the Bollywood lyrics corpus can be best approximated by two straight lines in the doubly logarithmic plot instead of one. The best fitting lines for the upper and lower parts of the distributions are also shown in Figure 1, which have been constructed through standard regression analysis. The slopes of these lines, which we shall refer to as γ_1 and γ_2 for the left (upper) and the right (lower) regimes respectively are 0.94 and 1.68 for the words, and 1.04 and 1.62 for morphological roots.

This distribution, which significantly deviates from Zipf's law, is commonly referred to in the literature as a *two-regime power-law*. Note that as predicted by Zipf's law, the exponent for the first regime, γ_1 is very close to 1 for both stemmed and unstemmed word distributions. However, γ_2 is significantly higher than 1. Interestingly, there are only around 700 to 1000 high frequent roots/words in the first regime which follow Zipf's law. The majority of the medium and low frequency roots/words comprising 95% of the vocabulary significantly deviates from the universally observable Zipf's law.

3.4 Why deviation from Zipf's Law?

Two regime-power law for rank-frequency distributions was first reported by Cancho and Solé (2001b). Based on an extensive study conducted on a very large English corpus, the authors observed that when the corpus size grows beyond certain limits, a two-regime power-law starts to emerge instead of a single regime Zipfian distribution. However, the power-law exponents of the two regimes were found to lie between 0.9 and 1.1. Cancho and Solé hypothesized that the two-regime was an outcome of kernel and periphery distinctions in natural language vocabulary, which will be discussed later. Note, however, that the deviation from the Zipf's law is much more pronounced for Lyrics corpus (with $\gamma_2 > 1.5$) than that reported in (Cancho and Solé, 2001b). Furthermore, the size of the lyrics corpus is several orders of magnitude smaller than those studied in (Cancho and Solé, 2001b).

Mathematical investigations of Zipf's law have led to several interesting insights, of which the following is of special significance: It has been observed that theoretically a perfect Zipfian distribution is possible only when the vocabulary of a growing corpus is potentially infinite (Kornai, 2002). In fact, parallel threads of investigation based on the concept of preferential usage of words in the text suggest that if the vocabulary or the pool of words to choose from is finite, then even though the rank-frequency distribution might resemble a Zipfian distribution when the corpus is small, actually it is better approximated by a β-distribution[5]. Thus, as the corpus grows beyond a limit the Zipfian nature of the distribution breaks down and it starts resembling a two-regime power-law when plotted on doubly-logarithmic scale (Peruani et al., 2007).

On the basis of these theoretical results, we argue that the vocabulary of Bollywood lyrics is extremely restricted; this restriction is stronger than even that of texts sampled from highly specialized domains, because the latter does obey Zipf's law. On the other hand, a large number of Bollywood songs are composed every year which is leading to a steep rise in the corpus size without any significant rise in the corpus vocabulary. Note that repetition of lines and stanzas in songs also adds to the apparent lack of diversity in the vocabulary of the corpus. In absence of any similar studies of

[5] A random variable is said to have a β-distribution with parameters $\beta > 0$ and $\beta > 0$ if and only if its probability mass function is given by, $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_{\alpha-1}(1-x)_{\beta-1}$ for $0 < x < 1$ and $f(x) = 0$ otherwise. $\Gamma()$ is the Eulers gamma function.

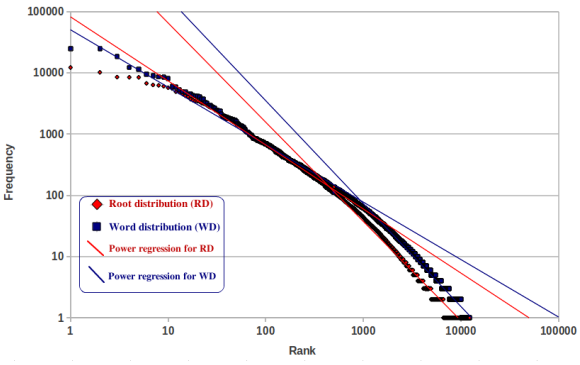

Figure 1: Frequency distribution of words in corpus.

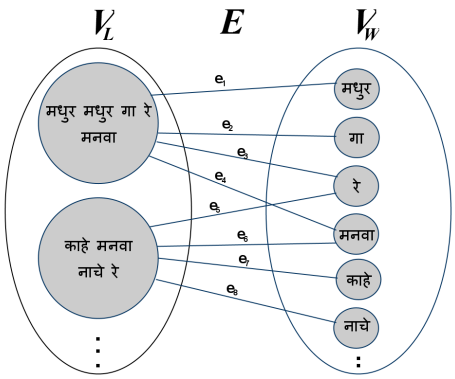

Figure 2: Illustration of nodes and edges of LWNet.

songs in other languages, we are unable to comment whether this property is unique to Bollywood lyrics or it is a property of lyrics in general.

4 Lyrics Networks: Construction and Properties

In this section we will study the properties of Bollywood lyrics by constructing complex network models. We will compute basic topological properties of the networks and compare those against that of complex networks built from standard Hindi text corpora. Following some of the previous studies on network based corpus linguistic analysis, here we will construct two basic network models. Although we conducted all the network analyses on both original and stemmed versions of the corpus, we observe that there are no significant qualitative differences between the two in their network characteristics. Therefore, we shall report the results only for the analyses conducted on the surface forms.

4.1 Network Models

Lyric-Word Network (LWNet): We define LWNet as a bipartite graph $G = \langle V_L, V_W, E \rangle$ where V_L is the set of nodes of type song lyrics and V_W is the set of nodes comprising unique words in the lyrics corpus. E is the set of edges that run between V_L and V_W. There is an edge $e \in E$ between two nodes $v_l \in V_L$ and $v_w \in V_W$ if and only if the word w occurs in the song l. Figure 2 illustrates the nodes and edges of LWNet. The construction of LWNet is driven by similar modeling of various complex phenomena such as Article-author network, where the edges denote which person has authored which articles (Newman, 2001), Movie-actor network,

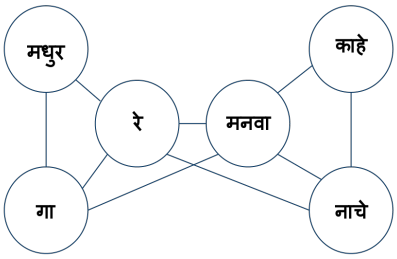

Figure 3: A partial illustration of the nodes and edges in WCNL. The labels of the nodes denote words in the lyrics corpus

where movies and actors comprise the two sets and an edge between them indicates that a particular actor acted in a particular movie (Ramasco et al., 2004), and Phoneme-Language Network where phonemes and languages constitute the two sets and an edge between them indicates that a particular phoneme occurs in a particular language (Mukherjee et al., 2008).

While studying bipartite networks, such as the one defined here, usually the one-mode projection of the network is also studied, which is defined as the network formed by the nodes present in only one of the partitions and the edges running between two nodes if they are connected to at least one common node from the other partition in the bipartite network. In the context of LWNet, the one-mode projection onto the word partition leads to a graph where the set of nodes is V_W, and two word nodes are connected by an edge if they occur in the same song. Since a large number of unique words (52 on an average) are present in a single song lyric, every lyric node will therefore translate into a clique of size 52 in the one-mode projection. Such a one-mode projection will be

extremely dense and not of much interest for complex network analysis. Therefore, along the lines of some other network-based studies in corpus linguistics (Cancho and Solé, 2001a; Choudhury et al., 2010), we study the word co-occurrence network of the lyrics corpora.

Word Co-occurrence Network for Lyrics (WCNL): We define WCNL as a network of words represented by a graph $G = \langle V_W, E \rangle$. An edge $e \in E$ represents co-occurrence of two words in the same song-lyrics, that is, if they are separated by zero or one word in a song lyric (Cancho and Solé, 2001a). Restriction of a network is used to prune the edges which might occur purely by chance (Cancho and Solé, 2001a). In a restricted network, if a pair of words co-occurs less than expected when independence between such words is assumed, the edge between the word nodes is pruned out. Figure 3 partially illustrates the nodes and edges of WCNL.

Note that WCNL is a slightly more restricted definition for the one-mode projection of LWNet; instead of adding edges between two words which "occurs" in the same lyric, we further restrict this occurrence by "co-occurrence within a small window". Saha Roy et al. (2011) has introduced the terms *global* and *local* co-occurrence respectively to denote these two versions of word co-occurrence networks.

4.2 Topological Properties of LWNet

Degree distribution of the nodes in V_L: Degree of a node refers to the number of edges incident on it. Thus degree of a lyric node in LWNet is the number of unique words present in the song. The degree distribution for $LWNet$ of the nodes in V_L is a plot where the x-axis denotes degree k expressed as a fraction of the maximum degree and the y-axis denotes p_k, the fraction of nodes in V_L having degree k. The distribution indicates that the number of unique words present in a Bollywood song follows a β-distribution. The values of α and β obtained using method-of-moments estimates are 5.1 and 18.9 respectively. The distribution peaks at 44, whereas the mean of the distribution for $LWNet$ is 52.

Degree distribution of the nodes in V_W: Figure 4 shows the cumulative degree distribution plot of the nodes in V_W in LWNet. In this figure, x-axis represents the degree k and the y-axis represents P_k, where P_k is the fraction of nodes having degree greater than or equal to k. As is apparent from the plot, the cumulative distribution follows a power-law with exponential cut off. The best-fit power-law exponent for the distribution is 1.02. The cumulative degree distribution is more robust to noise and data sparsity than its non-cumulative counterpart, p_k. However, since p_k is the negative derivative of P_k, we observe the following relationship between k and p_k for partition V_W (B is a positive constant of proportionality).

$$p_k = Bk^{-2.02}$$

Power-law degree distributions are known to emerge in complex networks when the new nodes entering the system connects to the pre-existing nodes through *preferential attachment* (see (Albert and Barabási, 2002) for an introduction). For bipartite networks, Ramasco et al. (2004) proposed a preferential attachment based network growth model that can be rephrased in the context of LWNet as follows: At every time step t, a new lyric node enters the set V_L; this node is connected to $\bar{n}$ nodes from the V_W partition (i.e., $\bar{n}$ is the average number of unique words per song, which is approximately 52 for LWNet); out of $\bar{n}$ nodes or words, $\bar{m}$ words are newly introduced in the V_W set (in other words, every song introduces $\bar{m}$ new words in the song corpus); the remaining $\bar{n} - \bar{m}$ words are chosen from the existing set of words in V_W through a degree based preferential attachment, where the probability of a node to be chosen for connection is directly proportionate to the current degree of the node (i.e., the number of songs in which the word already occurs).

Theoretical analysis of this model shows that the degree distribution, p_k, of the V_W partition in the emergent network will follow a power-law with exponent

$$\gamma = 2 + \frac{\bar{m}}{\bar{n} - \bar{m}}$$

Since, we observe that $\gamma = 2.02$ and know that $\bar{n} = 52$, we can compute $\bar{m}$ from the above equation which turns out to be equal to 1.02. In other words, every new Bollywood song on average introduces only 1.02 new words to the corpus. This again points to the limited diversity in the vocabulary of the Bollywood songs.

4.3 Topological Properties of WCNL

Table 2 lists some of the basic topological characteristics of WCNL as well as for the word co-

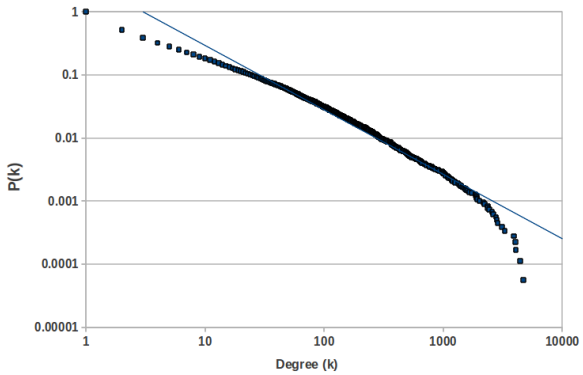

Figure 4: Degree distribution of LWNet for the set V_W in log-log scale.

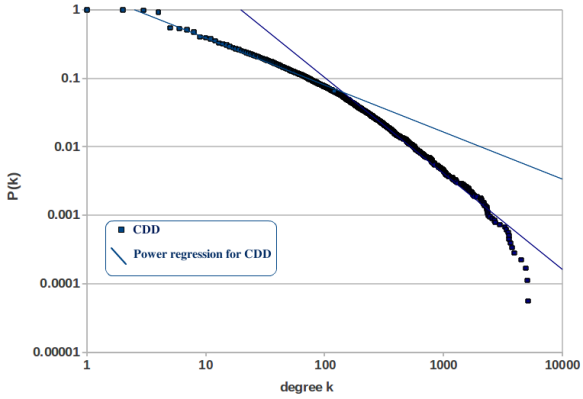

Figure 5: Cumulative degree distribution of WCNL in log-log scale.

occurrence networks constructed in similar fashion for the standard Hindi corpus and the two domain restricted corpora. Note that in order to have a meaningful comparison of the topological properties, we ensured that the number of tokens in the standard text corpora is the same as in the Lyrics corpus. While Table 2 provides a bird's eye view of the four networks and their restricted counterparts, in this subsection we will discuss these topological properties in details and highlight their linguistic significance in the context of the current study.

Connection Density: First, we note that even though all the networks were constructed from similar size corpora, the number of nodes in the networks (N), i.e., the number of unique words in the corpora, is least for Bollywood Lyrics corpus. The number of unique words in the general corpus is almost twice and in the domain specific networks almost 1.5 times that of the lyrics corpus. This further strengthens the fact that vocabulary of Bollywood lyrics is very repetitive. The average degree of a node, $\bar{k}$, gives an indication of the density of a network. Higher values of $\bar{k}$ indicates densely connected networks; $\bar{k}$ close to N indicates that the network is a clique or a complete graph. We observe that for both the restricted and unrestricted versions of the network, the corresponding $\bar{k}$ is highest for lyrics and lowest for the standard corpus, and the values for the domain specific corpora are between these two extremes. High average degree for WCNL indicates that even though the vocabulary for Bollywood lyrics is restricted, the words are used in various diverse contexts. In other words, lyricists do show their creativity by putting these words in various combinations and various orders, even if they are reluctant to use completely new vocabulary or metaphors.

Degree Distribution: Figure 5 shows the cumulative degree distribution of WCNL, which seems to be a two-regime power-law. The power-law fits are also shown for better visualization. The power-law exponents γ_1 and γ_2 for the two-regimes are 0.69 and 1.40, which implies that the corresponding exponents for the non-cumulative degree distribution are 1.69 and 2.40 respectively. This values are comparable to those reported by Cancho and Solé (2001a), which are 1.5 and 2.7 respectively. Interestingly, we do not observe a pronounced two-regime power-law for the corresponding degree distributions for the standard and domain specific corpora. The degree distributions for those can be best approximated by a single power-law with exponential cut-offs. The values of the exponents for the cumulative degree distribution are reported in Table 2, and the corresponding exponents for the non-cumulative distributions lie between 2.18 and 2.27.

Kernel-Peripheral Lexicon: A two-regime power-law is thought to emerge from presence of distinct kernel and peripheral lexicon in a language. While the kernel lexicon consists of function words and other very commonly used vocabulary, the peripheral lexicon consists of domain specific vocabulary that may not be uniformly present across all the documents in a corpus. The absence of a two-regime power-law for the standard Hindi corpora is due to the small size of the corpora which is not sufficient to bring out the kernel-periphery distinction. However, for the lyrics corpus, this distinction is prominent, per-

haps due to the limited vocabulary usage of Bollywood lyrics which makes it possible to observe a word in context a lot more times in the corpus than for the standard text corpus. As can be observed from Figure 1, the kernel lexicon consists of approximately 1000 words and the remaining 17000 words are in the peripheral lexicon giving a kernel-to-periphery ratio of 17 for Bollywood lyrics. For standard English, this ratio is approximately 85 (Cancho and Solé, 2001a). A significantly low kernel-to-periphery ratio again indicates lack of diversity in Bollywood lyrics. By knowing the 1000 words that constitute the kernel of Bollywood lyrics, one should be able to "comprehend" most of the Bollywood songs almost completely, because the peripheral words are much fewer and by definition they occur only sporadically across songs.

Clustering Coefficient: Given a node u, the probability that two nodes v and w, both of which are connected to u, are also connected to each other is known as the *clustering coefficient* of u or $C(u)$. The average of the $C(u)$'s of all nodes in the network is defined as the clustering coefficient of the network or C. As reported in Table 2, all the networks have almost identical values of C, which is around 0.65 for the unrestricted and 0.34 for the restricted versions.

In general we would expect a higher clustering coefficient for a network that has higher connection density. One way to factor out the effect of the connection density on clustering coefficient is to look at C/C_{rand} instead of C, where C_{rand} is the clustering coefficient of a random graph that has the same connection density as the original network. C_{rand} is given by the expression $\bar{k}/N$ (Cancho and Solé, 2001a). This ratio is around 300, 540 and 775 (190, 350 and 485, for restricted versions) respectively for the networks constructed from Bollywood lyrics, domain specific text corpus and standard text corpus. Thus, even though all the networks feature much higher C than their corresponding random networks, WCNL has a relatively lower C to C_{rand} ratio.

Thus, while the connection density of WCNL increases due to repeated use of the same words in various novel contexts, the clustering coefficient remains the same as that for the standard corpus network. This apparent fallacy can be understood as follows: If u (say *man*) co-occurs with v (say *kA*) and v with w (say *dil*), then the co-occurrence of u with w is typically controlled by the syntactic constraints imposed on the language (in this case, it is unlikely that *man* and *dil* will co-occur in any song, no matter how many songs are observed), rather than frequencies of occurrence. To summarize, clustering coefficient of word co-occurrence network is influenced by the syntactic properties, especially the local or intra-phrasal word ordering constraints, of a language and not the vocabulary usage. Since the clustering coefficients of all the networks are same, we can infer that Bollywood lyrics feature the same local syntactic constraints as the standard Hindi language. This is an interesting observation because one would normally expect Bollywood lyrics to feature more unrestricted word ordering, especially because Hindi is a relatively free-word order language, and moreover, lyrical constructions in general exploit and break word ordering restrictions. Nevertheless, our analysis do not support this apparently intuitive fact.

Diameter: Formally, the diameter of a graph is defined as the longest shortest path in the network. However, it is often approximated by the *average shortest path* which is easier to estimate both theoretically and algorithmically than the true *diameter*. Table 2 reports the average shortest path of the networks (d), which all lie between 2.56 and 2.80. Again, we observe that all the networks have very similar diameters that are also close to d_{rand} - the diameter of the corresponding random graphs with same connection density.

Small World Property: Three conditions must be satisfied in order for a network to be *small-world*: Sparseness ($N >> k$), $d \approx d_{rand}$ and $C >> C_{rand}$ (Watts and Strogatz, 1998). All these three properties are satisfied by all the networks, implying that word co-occurrence networks are indeed small-worlds irrespective of the style (stories, news reports or lyrics) and domain (Bollywood, business or religion) of the underlying corpora. Small-world nature of word co-occurrence networks has been associated with faster access of words from the mental lexicon (Cancho and Solé, 2001a). Thus, it is a universal property of natural languages and Bollywood lyrics are no exceptions.

5 Discussion and Conclusion

In this paper, we presented some basic statistical as well as more in-depth complex network analyses of Bollywood song lyrics and standard Hindi

Table 2: Topological characteristics of the Word Co-occurrence Networks of Bollywood Lyrics, standard language corpus and domain specific corpora. * marked values indicate the power-law exponent of the first regime whenever two-regime power law was observed. (Un)Rest: (Un)restricted, Gen: General corpus, Bus: Business corpus, Rel: Religion and Astrology

Network	$\|N\|$ $\times 10^3$	$\|E\|$ $\times 10^4$	$\bar{k}$	γ	γ_2	C	C_{rand} $\times 10^{-3}$	C/C_{rand}	d	d_{rand}
Unrest. Lyrics	17.9	35.2	39.4	$0.69^\star$	1.40	0.66	2.2	300	2.64	2.66
Rest. Lyrics	17.9	28.1	31.4	$0.74^\star$	1.61	0.32	1.7	188	2.77	2.84
Unrest. Gen.	31.1	38.8	24.9	1.20	–	0.62	0.8	775	2.59	3.21
Rest. Gen.	31.1	33.7	21.7	1.26	–	0.34	0.7	486	2.75	3.36
Unrest. Bus.	23.2	32.3	27.8	1.19	–	0.64	1.2	533	2.56	3.02
Rest. Bus.	23.2	27.1	23.4	1.27	–	0.35	1.0	350	2.74	3.19
Unrest. Rel.	21.8	30.7	28.1	1.18	–	0.65	1.2	542	2.63	2.99
Rest. Rel.	21.8	25.6	23.4	1.26	–	0.34	1.0	340	2.80	3.17

text corpora to understand the similar and distinct feature of the language of Bollywood lyrics as compared to that of standard Hindi. Some of the salient facts that came up from our analysis are:

Limited Vocabulary: The vocabulary used for composing Bollywood lyrics is small and limited, as compared to standard and even domain specific corpora. This is evident from the facts that (a) there are only 17000 unique words in 0.5 Million word lyrics corpus, (b) word frequency distribution significantly deviates from Zipf's law, and (c) every new lyric adds, on an average, only 1.02 new words.

Small Kernel-to-periphery ratio: There are around 1000 words that form the kernel of Bollywood lyrics vocabulary. The size of the peripheral lexicon is less than 20 times that of the kernel (this value is close to 100 for standard text).

Creative Word Usage: Very high connection density of WCNL implies that the limited vocabulary is very aptly used in various combinations to yield interesting and novel compositions.

Syntactic Restrictions: Same clustering coefficient across the networks suggest that the syntactic restrictions are no more or no less severe in Bollywood lyrics as compared to standard Hindi. This is contrary to the popular belief that poetry tends to enjoy more liberty from the strict rules of syntax.

Small-world Property: All the networks studied are small-worlds, which seems to be a basic property of every human language irrespective of the style or domain of the corpora.

Thus, most of the differences in the network properties of Bollywood lyrics and standard Hindi text seem to be explainable by one key point that *the vocabulary of Bollywood lyrics is much smaller than one would expect for a corpus of creative compositions of similar size in Hindi.* Why is the vocabulary of Bollywood lyrics so restricted and repetitive? This is a very interesting question whose answer would require a holistic analysis taking into consideration socio-cultural, linguistic and historical factors. This is beyond the scope of the current work, though we are tempted to make certain speculations.

Firstly, we note that the themes of the Bollywood songs are quite repetitive. Secondly, metaphors used also lack in diversity, some of them being derived from literature of the 17th and 18th centuries. For instance, *Radha-Krishna* metaphor is commonly used in romance and devotion themes; similarly, description of body parts involves a few very standard metaphors. Therefore, we hypothesize that lyricist in particular, and Bollywood artists in general have tried to stick to a well tested beaten path instead of exploring glaringly new themes and styles. This tendency might have emerged due to a strong bias or preference (often modeled as preferential attachment in complex network literature) towards reusing artistic or linguistic structures that have been popular in the past, downplaying creativity to minimize risks in the Indian film industry. The alternative, but much less plausible, hypothesis could be that the consumers of this industry themselves do not prefer variation.

Since there are no previous studies on lyrics in other languages to compare our results against, it is hard to comment whether the aforementioned characteristics are unique to Bollywood or in gen-

eral true for poetry or movie songs. Therefore, one starting point for investigating some of these interesting questions would be to conduct similar analysis on other poetry and lyrics corpora. Very recently there has been a wave of change in themes as well as metaphors of Bollywood lyrics. Since our corpus consists of songs mainly till 2005, it is not possible to study these recent trends from the current corpus. Nevertheless, it would be interesting to study the trends and dynamics of word usage in Bollywood lyrics over last 80 years.

Acknowledgement

This work was conceived as a project during the IASNLP-2011 summer school at IIIT Hyderabad. Therefore, we would like to thank the organizers of IASNLP-2011. Also, thanks to Kanika Gupta for her help in crawling and de-duplication of Bollywood lyrics.

References

Albert, R. and Barabási, A.- L. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**, 47–97.

Gena R. Bennett 2010. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. Michigan ELT.

Akshar Bharati, Vineet Chaitanya and Rajeev Sangal 1995. Words and their Analyzer. *Natural Language Processing : A Paninian Perspective*, Prentice Hall of India.

Monojit Choudhury, Diptesh Chatterjee, and Animesh Mukherjee 2010. Global topology of word co-occurrence networks: Beyond the two-regime power-law. *in Proceedings of Coling 2010*.

Monojit Choudhury and Animesh Mukherjee. 2009. The structure and dynamics of linguistic networks. In N. Bellomo,N. Ganguly, A. Deutsch, and A. Mukherjee, editors, *Dynamics On and Of Complex Networks*, Modeling and Simulation in Science, Engineering and Technology, pages 145–166. Birkhäuser Boston.

Ferrer-i-Cancho, R. and Solé, R.V. 2002. Zipf's law and random texts. *Advances in Complex Systems*, Vol. 5, No. 1 (2002) 1-6

Ferrer-i-Cancho, R. and Solé, R.V. 2001a. The small world of human language. *Proceedings of the Royal Society of London B*, 268(1482):2261-2265.

Ferrer-i-Cancho, R. and Solé, R.V. 2001b. Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8:165-173.

Gries, S. T. 2010. In A. Sánchez and M. Almela (Eds.) *A mosaic of corpus linguistics: Selected approaches* (pp. 269-291). Germany: Peter Lang, Frankfurt am Main

Jayaram, B. D., Vidya, M. N. 2008. Zipf's Law for Indian Languages. *Journal of Quantitative Linguistics*, Vol. 15, No. 4., pp. 293-317.

András Kornai. 2002. How many words are there? *Glottometrics 4*, 61-86.

Animesh Mukherjee, Monojit Choudhury, Anupam Basu, and Niloy Ganguly 2008. Modeling the Structure and Dynamics of the Consonant Inventories: A Complex Network Approach. *in Proceedings of Coling 2008*.

Mehler, A. 2008. Large text networks as an object of corpus linguistic studies. *Corpus Linguistics. An International Handbook of the Science of Language and Society* In A. Lüdeling and M. Kytö (Eds.), pp. 328-382. Berlin/New York: De Gruyter.

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

Newman, M. E. J. 2001. Scientific collaboration networks. *Phys. Rev. E* **64**

Fernando Peruani, Monojit Choudhury, Animesh Mukherjee, and Niloy Ganguly. 2007. Emergence of non-scaling degree distribution in bipartite networks: a numerical and analytical study. *Europhysics Letters*, vol. 79, no. 28001.

Ramasco, J. J., Dorogovtsev, S. N. and Pastor-Satorras, R. 2004. Self-organization of collaboration networks. *Physical Review E*, **70**(036106).

Rishiraj Saha Roy, Niloy Ganguly, M. Choudhury and A. Mukherjee 2011. Complex Network Analysis Reveals Kernel-Periphery Structure in Web Search Queries. *SIGIR*.

Sitabhra Sinha, Raj Kumar Pan, Nisha Yadav, Mayank Vahia, Iravatham Mahadevan. 2009. Network analysis reveals structure indicative of syntax in the corpus of undeciphered Indus civilization inscriptions. *in Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, ACL-IJCNLP 2009, pages 5-13

Watts, D. J. and Strogatz, S. H.. 1998. Collective dynamics of 'small-world' networks. *Nature* 393(6684):440-442

George K. Zipf 1949. Human Behavior and the Principle of Least Effort. Addison-Wesley.