Analysis and Synthesis of the Distribution of Consonants over Languages: A Complex Network Approach

Monojit Choudhury and **Animesh Mukherjee** and **Anupam Basu** and **Niloy Ganguly**
Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur
{monojit,animeshm,anupam,niloy}@cse.iitkgp.ernet.in

Abstract

Cross-linguistic similarities are reflected by the speech sound systems of languages all over the world. In this work we try to model such similarities observed in the consonant inventories, through a complex bipartite network. We present a systematic study of some of the appealing features of these inventories with the help of the bipartite network. An important observation is that the occurrence of consonants follows a two regime power law distribution. We find that the consonant inventory size distribution together with the principle of preferential attachment are the main reasons behind the emergence of such a two regime behavior. In order to further support our explanation we present a synthesis model for this network based on the general theory of preferential attachment.

1 Introduction

Sound systems of the world's languages show remarkable regularities. Any arbitrary set of consonants and vowels does not make up the sound system of a particular language. Several lines of research suggest that cross-linguistic similarities get reflected in the consonant and vowel inventories of the languages all over the world (Greenberg, 1966; Pinker, 1994; Ladefoged and Maddieson, 1996). Previously it has been argued that these similarities are the results of certain general principles like *maximal perceptual contrast* (Lindblom and Maddieson, 1988), *feature economy* (Martinet, 1968; Boersma, 1998; Clements, 2004) and *robustness* (Jakobson and Halle, 1956; Chomsky and Halle, 1968). Maximal perceptual contrast between the phonemes of a language is desirable for proper perception in a noisy environment. In fact the organization of the vowel inventories across languages has been satisfactorily explained in terms of the single principle of maximal perceptual contrast (Jakobson, 1941; Wang, 1968).

There have been several attempts to reason the observed patterns in consonant inventories since 1930s (Trubetzkoy, 1969/1939; Lindblom and Maddieson, 1988; Boersma, 1998; Flemming, 2002; Clements, 2004), but unlike the case of vowels, the structure of consonant inventories lacks a complete and holistic explanation (de Boer, 2000). Most of the works are confined to certain individual principles (Abry, 2003; Hinskens and Weijer, 2003) rather than formulating a general theory describing the structural patterns and/or their stability. Thus, the structure of the consonant inventories continues to be a *complex* jigsaw puzzle, though the parts and pieces are known.

In this work we attempt to represent the cross-linguistic similarities that exist in the consonant inventories of the world's languages through a *bipartite network* named **PlaNet** (the **P**honeme **L**anguage **N**etwork). PlaNet has two different sets of nodes, one labeled by the languages while the other labeled by the consonants. Edges run between these two sets depending on whether or not a particular consonant occurs in a particular language. This representation is motivated by similar modeling of certain complex phenomena observed in nature and society, such as,

- Movie-actor network, where movies and actors constitute the two partitions and an edge between them signifies that a particular actor acted in a particular movie (Ramasco et al., 2004).

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 128–135,
Sydney, July 2006.

- Article-author network, where the edges denote which person has authored which articles (Newman, 2001b).

- Metabolic network of organisms, where the corresponding partitions are chemical compounds and metabolic reactions. Edges run between partitions depending on whether a particular compound is a substrate or result of a reaction (Jeong et al., 2000).

Modeling of complex systems as networks has proved to be a comprehensive and emerging way of capturing the underlying generating mechanism of such systems (for a review on complex networks and their generation see (Albert and Barabási, 2002; Newman, 2003)). There have been some attempts as well to model the intricacies of human languages through complex networks. Word networks based on synonymy (Yook et al., 2001b), co-occurrence (Cancho et al., 2001), and phonemic edit-distance (Vitevitch, 2005) are examples of such attempts. The present work also uses the concept of complex networks to develop a platform for a holistic analysis as well as synthesis of the distribution of the consonants across the languages.

In the current work, with the help of PlaNet we provide a systematic study of certain interesting features of the consonant inventories. An important property that we observe is the two regime power law degree distribution[1] of the nodes labeled by the consonants. We try to explain this property in the light of the size of the consonant inventories coupled with the principle of *preferential attachment* (Barabási and Albert, 1999). Next we present a simplified mathematical model explaining the emergence of the two regimes. In order to support our analytical explanations, we also provide a synthesis model for PlaNet.

The rest of the paper is organized into five sections. In section 2 we formally define PlaNet, outline its construction procedure and present some studies on its degree distribution. We dedicate section 3 to state and explain the inferences that can be drawn from the degree distribution studies of PlaNet. In section 4 we provide a simplified theoretical explanation of the analytical results obtained. In section 5 we present a synthesis model for PlaNet to hold up the inferences that we draw in section 3. Finally we conclude in section 6 by summarizing our contributions, pointing out some of the implications of the current work and indicating the possible future directions.

[1]Two regime power law distributions have also been observed in syntactic networks of words (Cancho et al., 2001), network of mathematics collaborators (Grossman et al., 1995), and language diversity over countries (Gomes et al., 1999).

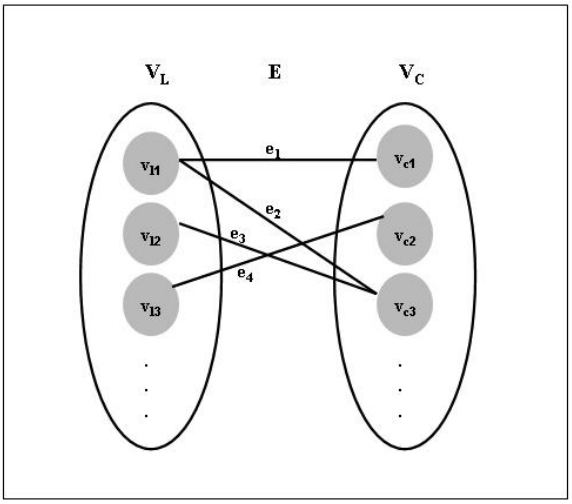

Figure 1: Illustration of the nodes and edges of PlaNet

2 PlaNet: The Phoneme-Language Network

We define the network of consonants and languages, PlaNet, as a *bipartite graph* represented as G = $\langle$V$_L$, V$_C$, E$\rangle$ where V$_L$ is the set of *nodes* labeled by the languages and V$_C$ is the set of nodes labeled by the consonants. E is the set of edges that run between V$_L$ and V$_C$. There is an *edge* $e \in$ E between two nodes $v_l \in$ V$_L$ and $v_c \in$ V$_C$ if and only if the consonant *c* occurs in the language *l*. Figure 1 illustrates the nodes and edges of PlaNet.

2.1 Construction of PlaNet

Many typological studies (Lindblom and Maddieson, 1988; Ladefoged and Maddieson, 1996; Hinskens and Weijer, 2003) of segmental inventories have been carried out in past on the UCLA Phonological Segment Inventory Database (UPSID) (Maddieson, 1984). UPSID initially had 317 languages and was later extended to include 451 languages covering all the major language families of the world. In this work we have used the older version of UPSID comprising of 317 languages and 541 consonants (henceforth UPSID$_{317}$), for constructing PlaNet. Consequently, there are 317 elements (nodes) in the set V$_L$ and 541 elements

(nodes) in the set V_C. The number of elements (edges) in the set E as computed from PlaNet is 7022. At this point it is important to mention that in order to avoid any confusion in the construction of PlaNet we have appropriately filtered out the anomalous and the ambiguous segments (Maddieson, 1984) from it. We have completely ignored the anomalous segments from the data set (since the existence of such segments is doubtful), and included the ambiguous ones as separate segments because there are no descriptive sources explaining how such ambiguities might be resolved. A similar approach has also been described in Pericliev and Valdés-Pérez (2002).

2.2 Degree Distribution of PlaNet

The *degree* of a node u, denoted by k_u is defined as the number of edges connected to u. The term *degree distribution* is used to denote the way degrees (k_u) are distributed over the nodes (u). The degree distribution studies find a lot of importance in understanding the complex topology of any large network, which is very difficult to visualize otherwise. Since PlaNet is bipartite in nature it has two degree distribution curves one corresponding to the nodes in the set V_L and the other corresponding to the nodes in the set V_C.

Degree distribution of the nodes in V_L: Figure 2 shows the degree distribution of the nodes in V_L where the x-axis denotes the degree of each node expressed as a fraction of the maximum degree and the y-axis denotes the number of nodes having a given degree expressed as a fraction of the total number of nodes in V_L .

It is evident from Figure 2 that the number of consonants appearing in different languages follow a β-distribution [2] (see (Bulmer, 1979) for reference). The figure shows an asymmetric right skewed distribution with the values of α and β equal to 7.06 and 47.64 (obtained using maximum likelihood estimation method) respectively. The asymmetry points to the fact that languages usually tend to have smaller consonant inventory size,

[2] A random variable is said to have a β-distribution with parameters $\alpha > 0$ 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. $\Gamma(\cdot)$ is the Euler's gamma function.

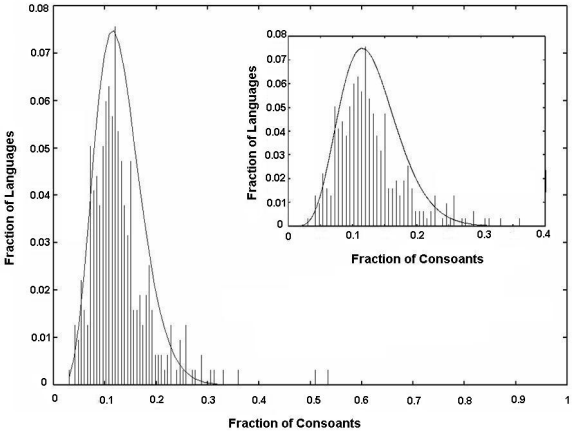

Figure 2: Degree distribution of PlaNet for the set V_L. The figure in the inner box is a magnified version of a portion of the original figure.

the best value being somewhere between 10 and 30. The distribution peaks roughly at 21 indicating that majority of the languages in $UPSID_{317}$ have a consonant inventory size of around 21 consonants.

Degree distribution of the nodes in V_C: Figure 3 illustrates two different types of degree distribution plots for the nodes in V_C; Figure 3(a) corresponding to the rank, i.e., the sorted order of degrees, (x-axis) versus degree (y-axis) and Figure 3(b) corresponding to the degree (k) (x-axis) versus P_k (y-axis) where P_k is the fraction of nodes having degree greater than or equal to k.

Figure 3 clearly shows that both the curves have two distinct regimes and the distribution is scale-free. Regime 1 in Figure 3(a) consists of 21 consonants which have a very high frequency (i.e., the degree k) of occurrence. Regime 2 of Figure 3(b) also correspond to these 21 consonants. On the other hand Regime 2 of Figure 3(a) as well as Regime 1 of Figure 3(b) comprises of the rest of the consonants. The point marked as **x** in both the figures indicates the breakpoint. Each of the regime in both Figure 3(a) and (b) exhibit a power law of the form

$$y = Ax^{-\alpha}$$

In Figure 3(a) y represents the degree k of a node corresponding to its rank x whereas in Figure 3(b) y corresponds to P_k and x, the degree k. The values of the parameters A and α, for Regime 1 and Regime 2 in both the figures, as computed by the least square error method, are shown in Table 1.

Regime	Figure 3(a)	Figure 3(b)
Regime 1	A = 368.70 α = 0.4	A = 1.040 α = 0.71
Regime 2	A = 12456.5 α = 1.54	A = 2326.2 α = 2.36

Table 1: The values of the parameters A and α

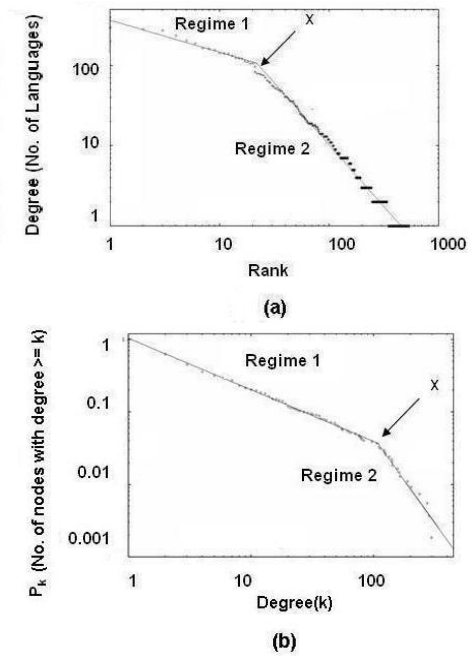

Figure 3: Degree distribution of PlaNet for the set V_C in a log-log scale

It becomes necessary to mention here that such power law distributions, known variously as Zipf's law (Zipf, 1949), are also observed in an extraordinarily diverse range of phenomena including the frequency of the use of words in human language (Zipf, 1949), the number of papers scientists write (Lotka, 1926), the number of hits on web pages (Adamic and Huberman, 2000) and so on. Thus our inferences, detailed out in the next section, mainly centers around this power law behavior.

3 Inferences Drawn from the Analysis of PlaNet

In most of the networked systems like the society, the Internet, the World Wide Web, and many others, power law degree distribution emerges for the phenomenon of preferential attachment, i.e., when "the rich get richer" (Simon, 1955). With reference to PlaNet this preferential attachment can be interpreted as the tendency of a language to choose a consonant that has been already chosen by a large number of other languages. We posit that it is this preferential property of languages that results in the power law degree distributions observed in Figure 3(a) and (b).

Nevertheless there is one question that still remains unanswered. Whereas the power law distribution is well understood, the reason for the two distinct regimes (with a sharp break) still remains unexplored. We hypothesize that,

Hypothesis *The typical distribution of the consonant inventory size over languages coupled with the principle of preferential attachment enforces the two distinct regimes to appear in the power law curves.*

As the average consonant inventory size in UPSID_{317} is 21, so following the principle of preferential attachment, on an average, the first 21 most frequent consonants are much more preferred than the rest. Consequently, the nature of the frequency distribution for the highly frequent consonants is different from the less frequent ones, and hence there is a transition from Regime 1 to Regime 2 in the Figure 3(a) and (b).

Support Experiment: In order to establish that the consonant inventory size plays an important role in giving rise to the two regimes discussed above we present a support experiment in which we try to observe whether the breakpoint **x** shifts as we shift the average consonant inventory size.

Experiment: In order to shift the average consonant inventory size from 21 to 25, 30 and 38 we neglected the contribution of the languages with consonant inventory size less than n where n is 15, 20 and 25 respectively and subsequently recorded the degree distributions obtained each time. We did not carry out our experiments for average consonant inventory size more than 38 because the number of such languages are very rare in UPSID_{317}.

Observations: Figure 4 shows the effect of this shifting of the average consonant inventory size on the rank versus degree distribution curves. Table 2 presents the results observed from these curves with the left column indicating the average inventory size and the right column the breakpoint **x**.

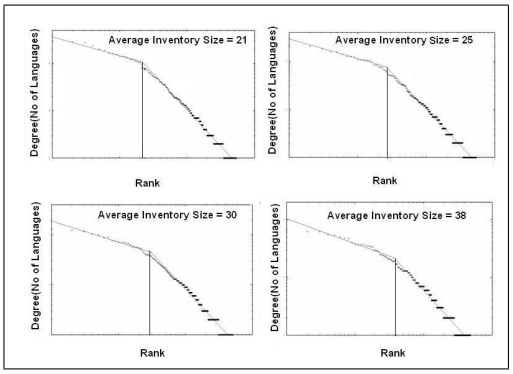

Figure 4: Degree distributions at different average consonant inventory sizes

Avg. consonant inv. size	Transition
25	25
30	30
38	37

Table 2: The transition points for different average consonant inventory size

The table clearly indicates that the transition occurs at values corresponding to the average consonant inventory size in each of the three cases.

Inferences: It is quite evident from our observations that the breakpoint **x** has a strong correlation with the average consonant inventory size, which therefore plays a key role in the emergence of the two regime degree distribution curves.

In the next section we provide a simplistic mathematical model for explaining the two regime power law with a breakpoint corresponding to the average consonant inventory size.

4 Theoretical Explanation for the Two Regimes

Let us assume that the inventory of all the languages comprises of 21 consonants. We further assume that the consonants are arranged in their hierarchy of preference. A language traverses the hierarchy of consonants and at every step decides with a probability p to choose the current consonant. It stops as soon as it has chosen all the 21 consonants. Since languages must traverse through the first 21 consonants regardless of whether the previous consonants are chosen or not, the probability of choosing any one of these 21 consonants must be p. But the case is different for the 22^{nd} consonant, which is chosen by a language if it has previously chosen zero, one, two, or at most 20, but not all of the first 21 consonants. Therefore, the probability of the 22^{nd} consonant being chosen is,

$$P(22) = p\sum_{i=0}^{20} \left(\begin{array}{c} 21 \\ i \end{array} \right) p^i(1-p)^{21-i}$$

where

$$\left(\begin{array}{c} 21 \\ i \end{array} \right) p^i(1-p)^{21-i}$$

denotes the probability of choosing i consonants from the first 21. In general the probability of choosing the n+1^{th} consonant from the hierarchy is given by,

$$P(n+1) = p\sum_{i=0}^{20} \left(\begin{array}{c} n \\ i \end{array} \right) p^i(1-p)^{n-i}$$

Figure 5 shows the plot of the function $P(n)$ for various values of p which are 0.99, 0.95, 0.9, 0.85, 0.75 and 0.7 respectively in log-log scale. All the curves, for different values of p, have a nature similar to that of the degree distribution plot we obtained for PlaNet. This is indicative of the fact that languages choose consonants from the hierarchy with a probability function comparable to $P(n)$.

Owing to the simplified assumption that all the languages have only 21 consonants, the first regime is a straight line; however we believe a more rigorous mathematical model can be built taking into consideration the β-distribution rather than just the mean value of the inventory size that can explain the negative slope of the first regime. We look forward to do the same as a part of our future work. Rather, here we try to investigate the effect of the exact distribution of the language inventory size on the nature of the degree distribution of the consonants through a synthetic approach based on the principle of preferential attachment, which is described in the subsequent section.

5 The Synthesis Model based on Preferential Attachment

Albert and Barabási (1999) observed that a common property of many large networks is that the vertex connectivities follow a scale-free power law distribution. They remarked that two generic mechanisms can be considered to be the cause of this observation: (i) networks expand continuously by the addition of new vertices, and (ii) new vertices attach preferentially to sites (vertices) that are already well connected. They found that

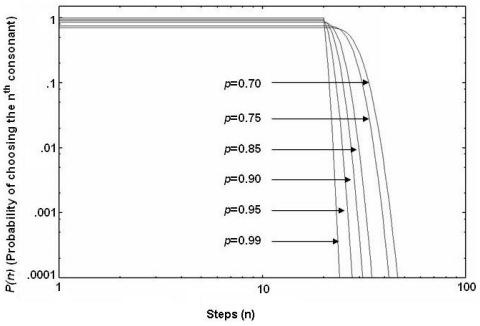

Figure 5: Plot of the function $P(n)$ in log-log scale

a model based on these two ingredients reproduces the observed stationary scale-free distributions, which in turn indicates that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems.

Inspired by their work and the empirical as well as the mathematical analysis presented above, we propose a preferential attachment model for synthesizing PlaNet (PlaNet$_{syn}$ henceforth) in which the degree distribution of the nodes in V_L is known. Hence V_L=$\{\mathrm{L}_1, \mathrm{L}_2, \ldots, \mathrm{L}_{317}\}$ have degrees (consonant inventory size) $\{k_1, k_2, \ldots, k_{317}\}$ respectively. We assume that the nodes in the set V_C are *unlabeled*. At each time step, a node L_j (j = 1 to 317) from V_L tries to attach itself with a new node $i \in \mathrm{V}_C$ to which it is not already connected. The probability $Pr(i)$ with which the node L_j gets attached to i depends on the current degree of i and is given by

$$Pr(i) = \frac{k_i + \epsilon}{\sum_{i' \in V_j}(k_{i'} + \epsilon)}$$

where k_i is the current degree of the node i, V_j is the set of nodes in V_C to which L_j is not already connected and ϵ is the smoothing parameter which is used to reduce bias and favor at least a few attachments with nodes in V_j that do not have a high $Pr(i)$. The above process is repeated until all $\mathrm{L}_j \in \mathrm{V}_L$ get connected to exactly k_j nodes in V_C. The entire idea is summarized in Algorithm 1. Figure 6 shows a partial step of the synthesis process illustrated in Algorithm 1.

Simulation Results: Simulations reveal that for PlaNet$_{syn}$ the degree distribution of the nodes belonging to V_C fit well with the analytical results we obtained earlier in section 2. Good fits emerge

repeat
  for *j = 1 to 317* **do**
    if *there is a node* $L_j \in V_L$ *with at least one or more consonants to be chosen from* V_C **then**
      Compute V_j = V_C-$V(L_j)$, where $V(L_j)$ is the set of nodes in V_C to which L_j is already connected;
    end
    for *each node i* $\in V_j$ **do**
      $$Pr(i) = \frac{k_i + \epsilon}{\sum_{i' \in V_j}(k_{i'} + \epsilon)}$$
      where k_i is the current degree of the node i and ϵ is the model parameter. $Pr(i)$ is the probability of connecting L_j to i.
    end
    Connect L_j to a node $i \in V_j$ following the distribution $Pr(i)$;
  end
until *all languages complete their inventory quota* ;

Algorithm 1: Algorithm for synthesis of PlaNet based on preferential attachment

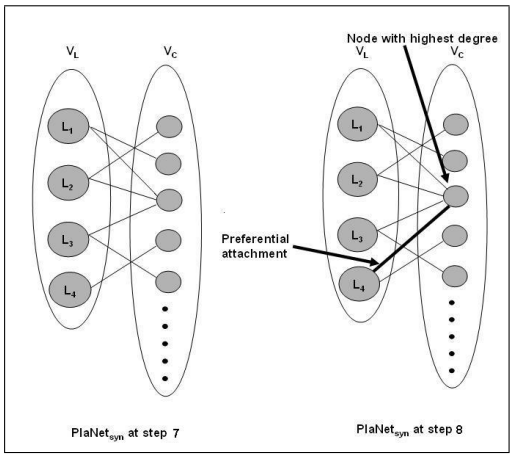

Figure 6: A partial step of the synthesis process. When the language L_4 has to connect itself with one of the nodes in the set V_C it does so with the one having the highest degree (=3) rather than with others in order to achieve preferential attachment which is the working principle of our algorithm

for the range $0.06 \leq \epsilon \leq 0.08$ with the best being at ϵ = 0.0701. Figure 7 shows the degree k versus

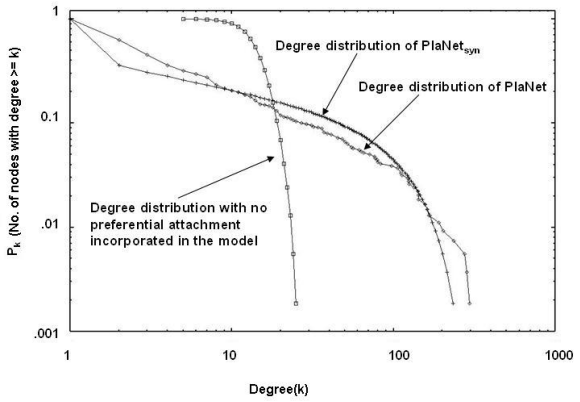

Figure 7: Degree distribution of the nodes in V_C for both PlaNet$_{syn}$, PlaNet, and when the model incorporates no preferential attachment; for PlaNet$_{syn}$, ϵ = 0.0701 and the results are averaged over 100 simulation runs

P_k plots for ϵ = 0.0701 averaged over 100 simulation runs.

The mean error[3] between the degree distribution plots of PlaNet and PlaNet$_{syn}$ is 0.03 which intuitively signifies that on an average the variation in the two curves is 3%. On the contrary, if there were no preferential attachment incorporated in the model (i.e., all connections were equiprobable) then the mean error would have been 0.35 (35% variation on an average).

6 Conclusions, Discussion and Future Work

In this paper, we have analyzed and synthesized the consonant inventories of the world's languages in terms of a complex network. We dedicated the preceding sections essentially to,

- Represent the consonant inventories through a bipartite network called PlaNet,
- Provide a systematic study of certain important properties of the consonant inventories with the help of PlaNet,
- Propose analytical explanations for the two regime power law curves (obtained from PlaNet) on the basis of the distribution of the consonant inventory size over languages together with the principle of preferential attachment,
- Provide a simplified mathematical model to support our analytical explanations, and
- Develop a synthesis model for PlaNet based on preferential attachment where the consonant inventory size distribution is known *a priori*.

We believe that the general explanation provided here for the two regime power law is a fundamental result, and can have a far reaching impact, because two regime behavior is observed in many other networked systems.

Until now we have been mainly dealing with the computational aspects of the distribution of consonants over the languages rather than exploring the real world dynamics that gives rise to such a distribution. An issue that draws immediate attention is that how preferential attachment, which is a general phenomenon associated with network evolution, can play a prime role in shaping the consonant inventories of the world's languages. The answer perhaps is hidden in the fact that language is an evolving system and its present structure is determined by its past evolutionary history. Indeed an explanation based on this evolutionary model, with an initial disparity in the distribution of consonants over languages, can be intuitively verified as follows – let there be a language community of N speakers communicating among themselves by means of only two consonants say /k/ and /g/. If we assume that every speaker has l descendants and language inventories are transmitted with high fidelity, then after i generations it is expected that the community will consist of ml^i /k/ speakers and nl^i /g/ speakers. Now if $m > n$ and $l > 1$, then for sufficiently large i, $ml^i \gg nl^i$. Stated differently, the /k/ speakers by far outnumbers the /g/ speakers even if initially the number of /k/ speakers is only slightly higher than that of the /g/ speakers. This phenomenon is similar to that of preferential attachment where language communities get attached to, i.e., select, consonants that are already highly preferred. Nevertheless, it remains to be seen where from such an initial disparity in the distribution of the consonants over languages might have originated.

In this paper, we mainly dealt with the occurrence principles of the consonants in the inventories of the world's languages. The work can be further extended to identify the co-occurrence likelihood of the consonants in the language inventories

[3]Mean error is defined as the average difference between the ordinate pairs where the abscissas are equal.

and subsequently identify the groups or communities within them. Information about such communities can then help in providing an improved insight about the organizing principles of the consonant inventories.

References

C. Abry. 2003. [b]-[d]-[g] as a universal triangle as acoustically optimal as [i]-[a]-[u]. *15th Int. Congr. Phonetics Sciences ICPhS*, 727–730.

L. A. Adamic and B. A. Huberman. 2000. The nature of markets in the World Wide Web. *Quarterly Journal of Electronic Commerce 1*, 512.

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

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

Bart de Boer. 2000. Self-Organisation in Vowel Systems. *Journal of Phonetics*, Elsevier.

P. Boersma. 1998. *Functional Phonology. (Doctoral thesis, University of Amsterdam)*, The Hague: Holland Academic Graphics.

M. G. Bulmer. 1979. *Principles of Statistics*, Mathematics.

Ferrer i Cancho and R. V. Solé. 2001. Santa Fe working paper 01-03-016.

N. Chomsky and M. Halle. 1968. *The Sound Pattern of English*, New York: Harper and Row.

N. Clements. 2004. Features and Sound Inventories. *Symposium on Phonological Theory: Representations and Architecture*, CUNY.

E. Flemming. 2002. *Auditory Representations in Phonology*, New York and London: Routledge.

M. A. F. Gomes, G. L. Vasconcelos, I. J. Tsang, and I. R. Tsang. 1999. Scaling relations for diversity of languages. Physica A, 271, 489.

J. H. Greenberg. 1966. *Language Universals with Special Reference to Feature Hierarchies*, The Hague Mouton.

J. W. Grossman and P. D. F. Ion. 1995. On a portion of the well-known collaboration graph. *Congressus Numerantium*, 108, 129-131.

F. Hinskens and J. Weijer. 2003. Patterns of segmental modification in consonant inventories: a cross-linguistic study. *Linguistics.*

R. Jakobson. 1941. *Kindersprache, Aphasie und allgemeine Lautgesetze*, Uppsala, Reprinted in *Selected Writings I. Mouton*, The Hague, 1962, pages 328-401.

H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. 2000. The large-scale organization of metabolic networks. Nature, 406:651-654.

R. Jakobson and M. Halle. 1956. *Fundamentals of Language*, The Hague: Mouton and Co.

P. Ladefoged and I. Maddieson. 1996. *Sounds of the Worlds Languages*, Oxford: Blackwell.

B. Lindblom and I. Maddieson. 1988. Phonetic Universals in Consonant Systems. In L.M. Hyman and C.N. Li, eds., *Language, Speech, and Mind*, Routledge, London, 62–78.

A. J. Lotka. 1926. The frequency distribution of scientific production. *J. Wash. Acad. Sci. 16*, 317-323.

I. Maddieson. 1984. *Patterns of Sounds*, Cambridge University Press, Cambridge.

A. Martinet. 1968. Phonetics and linguistic evolution. In Bertil Malmberg (ed.), *Manual of phonetics, revised and extended edition*, Amsterdam: North-Holland Publishing Co. 464–487.

M. E. J. Newman. 2001b. Scientific collaboration networks. *I and II. Phys. Rev.*, E 64.

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

V. Pericliev, R. E. Valdés-Pérez. 2002. Differentiating 451 languages in terms of their segment inventories. *Studia Linguistica*, Blackwell Publishing.

S. Pinker. 1994. *The Language Instinct*, New York: Morrowo.

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

H. A. Simon. 1955. On a class of skew distribution functions. *Biometrika 42*, 425-440.

N. Trubetzkoy. 1969. *Principles of phonology. (English translation of Grundzüge der Phonologie, 1939)*, Berkeley: University of California Press.

M. S. Vitevitch. 2005. Phonological neighbors in a small world: What can graph theory tell us about word learning? *Spring 2005 Talk Series on Networks and Complex Systems*, Indiana University, Bloomington.

William S.-Y. Wang. 1968. The basis of speech, Project on Linguistic Analysis Reports, University of California at Berkeley. Reprinted in *The Learning of Language*, ed. by C. E. Reed, 1971.

S. Yook, H. Jeong and A.-L. Barabási. 2001b. preprint.

G. K. Zipf. 1949. *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Reading, MA.