

The Networks of Syllables and Characters in Chinese[†]

Gang Peng^{1,2,*}, James W. Minett², and William S-Y. Wang²

¹ Department of Electrical Engineering, University of Washington,
Seattle, WA 98195, USA

² Language Engineering Laboratory, Department of Electronic Engineering,
The Chinese University of Hong Kong, Hong Kong SAR, China

Abstract

We develop networks using the syllables (both base syllables and tonal syllables) and characters of Chinese. The nodes (vertices) of the networks represent the syllables of the syllable network and the characters of the character network respectively. The links (edges) are established each time two syllables (or two characters) form part of the same word. We use two dictionaries to perform the analysis: a Putonghua³ dictionary and a Cantonese dictionary. All networks here show low distances and high clustering coefficients compared with ER random networks. The degree distributions all follow a power law, however the exponents for the base syllable, tonal syllable and Chinese character networks differ considerably. These differences may account for the different cognitive processes used when constructing new Chinese words. The networks are compared to the syllabic networks of Portuguese in terms of the magnitude of the power law exponent. The Chinese character network is found to be the most similar to the Portuguese syllabic network ($\gamma \approx 1.4$).

Keywords: Scale-free network; Small-world network; Syllabic network.

Introduction

Innumerable studies reveal that real networks, such as social interactions, biological connections, ecological webs, networks in linguistics, the World Wide Web and Internet, etc. are considerably different from random networks. Random networks have very low clustering coefficient and Poisson-like degree distribution; while many real networks have high clustering coefficient and power-law degree distribution [Barabási & Albert, 2002]. These studies from the complex network perspective provide intuitive and useful insights for investigating complex phenomena in real systems.

Recently, theories of complex networks have been applied to linguistic problems. The

[†] The authors gratefully acknowledge the support of the Research Grants Council of Hong Kong (1224/02H and 1127/04H).

^{*} Address correspondence to: Gang Peng, Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA. E-mail: gpeng@ee.washington.edu (Gang Peng)

³ Putonghua is a generally accepted term for Modern Standard Chinese (MSC).

networks of words in human languages can be linked in several ways: for instance, semantic relations, and word co-occurrence [Dorogovtsev and Mendes, 2001; Ferrer i Cancho and Solé, 2001; Steyvers and Tenenbaum, 2005]. Such networks may imply certain mechanisms by which humans organize concepts while choosing them for communication [Ferrer i Cancho and Solé, 2001, 2003]. A large word database (WordNet) containing 66,025 nouns along with their semantic relationships allows for a study that shows this network to have both scale-free and small-world characteristics [Sigman and G. Cecchi, 2002]. This network exhibits a degree distribution with power law tail, along with a very high clustering coefficient. Networks in which two words are linked if they co-occur no more than two words apart with a frequency higher than a chosen threshold, as well as co-occurrence networks based on the British National Corpus, were found to have a degree distribution with two distinct regimes of power law scaling [Ferrer i Cancho and Solé, 2001].

From a different perspective, Medeiros Soares et al. have built syllabic networks for Portuguese [Medeiros Soares et al., 2005]. The nodes of their networks represent the syllables of the target language; the links are established when two syllables form part of the same word. The authors observe that the degree distribution of the Portuguese syllable network has a power law distribution, and conclude that the evolution of Portuguese follows a rule similar to preferential attachment.

In this work, following a method similar to that used by Medeiros Soares and his colleagues [Medeiros Soares et al., 2005], we develop syllabic and graphemic (in terms of Chinese characters) networks for both Putonghua and Cantonese. In section 2, we introduce the structure of the Chinese language and the procedures for constructing networks. In the third section, we analyze the numerics of the networks: the average distance, the clustering coefficient and the degree distribution. Finally, we compare the Chinese networks with the Portuguese network, and propose possible explanations for the differences between these networks.

Networks of Chinese syllables and characters

The Chinese writing system is quite different from those languages having alphabetic scripts. The basic unit of Chinese writing system is the Chinese character, which is composed of two types of smaller unit called the stroke and the radical¹ [Wang, 1973]. Each character has one or more spellings based on the Latin alphabet. For instance, the Chinese character “市 (city)” is pronounced as “shi4” in the Pinyin system for Putonghua and as “si5” in the Jyutping system for Cantonese, where the numbers, 4 and 5, indicate the corresponding tones in Putonghua and Cantonese, respectively.

Typically, each Chinese character is pronounced as a mono-syllable. However, a Chinese character may have multiple mono-syllabic pronunciations, while one pronunciation can be shared by several characters. Considering both the suprasegmental and segmental

¹ Radical [e.g., Wang, 1973], or “部首”, literally means the “section header” under which a character is listed in the dictionary.

features of a syllable, there are 1,471 so-called 'tonal syllables' in Putonghua [CCDICT, 2000], e.g. “shi4”, and 1,761 in Cantonese [LSHK, 1997], e.g. “si5”. Ignoring the suprasegmental features of the syllable, i.e., the tone, the remaining segmental features correspond to the 'base syllable'. Analysis at the base syllable level may reveal something about the structure of the Chinese lexicon that occurs only at the segmental level. There are about 420 base syllables in Putonghua [CCDICT, 2000], e.g. “shi”, and 625 in Cantonese [LSHK, 1997], e.g. “si”.

Unlike some languages with alphabetic scripts², such as Portuguese and English, there is no space between Chinese words. Moreover, there is no precise definition of what is a Chinese word. Nevertheless, a Chinese word can be mono-syllabic (made up of one Chinese character) or multi-syllabic (comprising words with two or more Chinese characters).

There are three levels of syllable networks in Chinese. At the base syllable level, the nodes represent the base syllables; a link is established between a pair of nodes for which there exists at least one Chinese word that shares the corresponding base syllables. Networks at the tonal syllable and Chinese character levels are established similarly. We illustrate such networks for Chinese using an example comprising seven bi-syllabic words, shown in Table 1.

Words	Pinyin	Jyutping	Meaning
油井	you2 jing3	jau4 zeng2	oil well
汽油	qi4 you2	hei3 jau4	gasoline
火警	huo3 jing3	fo2 ging2	fire alarm
汽車	qi4 che1	hei3 ce1	automobile
火車	huo3 che1	fo2 ce1	train
火災	huo3 zai1	fo2 zoi1	fire accident
貨車	huo4 che1	fo3 ce1	wagon

Table 1. Seven Chinese bi-syllabic words.

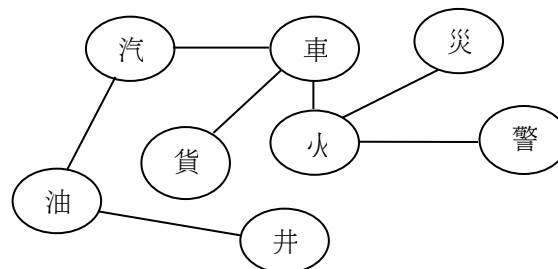


Figure 1. Network of Chinese characters.

² Not every language with alphabetic scripts uses spaces to separate words. And comparatively more languages even do not have a writing system.

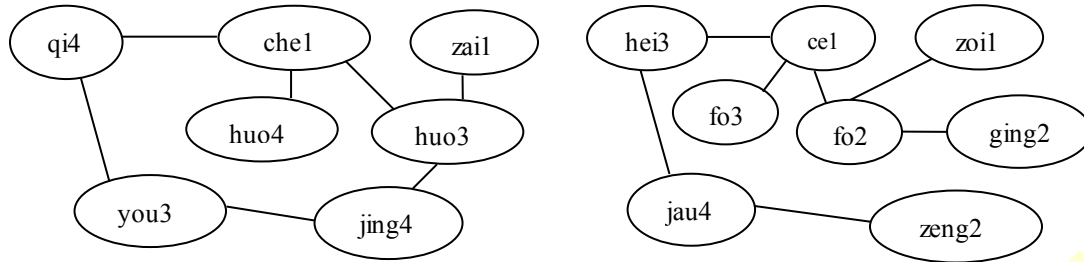


Figure 2. Tonal syllabic networks for Putonghua (left) and Cantonese (right).

For the seven Chinese words shown in Table 1, the corresponding character network, shown in Figure 1, comprises eight nodes, one node for each of the characters (in type). Each link in the network indicates that the two linked Chinese characters form a word. The Chinese characters “井” and “警” share the same tonal syllable in Putonghua, hence there are only seven nodes in the Putonghua tonal syllabic network shown in Figure 2. For the base syllabic network, shown in Figure 3, there are six nodes for Putonghua, because the tonal syllables “huo4” and “huo3” share the same base syllable “huo”. The Cantonese base syllable network consists of seven nodes because the tonal syllables “fo3” and “fo2” share the same base syllable “fo”.

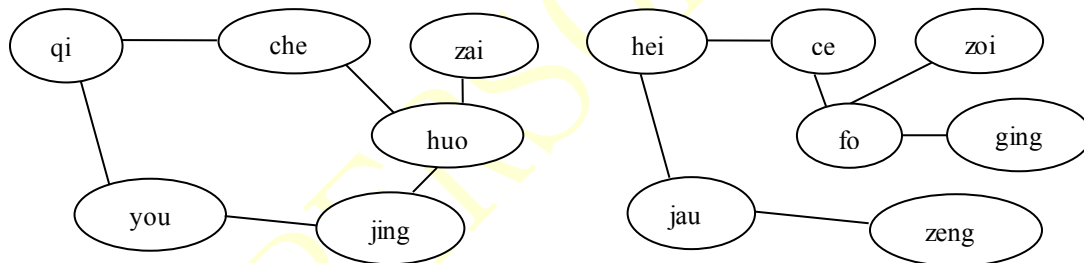


Figure 3. Base syllabic networks for Putonghua (left) and Cantonese (right).

Two distinct databases have been used in our analysis. The first is a Putonghua dictionary, CEDICT, available on the World Wide Web [Denisowski, 2005], that aims to provide a comprehensive Chinese-to-English dictionary with pronunciation in Pinyin. The other is a Cantonese dictionary, CULEX, which was developed for Cantonese speech recognition [Lee et al., 2002]. CEDICT is open source, so anyone can contribute new entries to it. As a result, CEDICT contains numerous uncommonly used Chinese characters, many of which appear as monosyllabic words. The version of CEDICT that we use comprises 21,727 multi-syllabic words and 8,834 mono-syllabic words. CULEX comprises 35,732 multi-syllabic words, and 5,737 monosyllabic words. Because CULEX was built for speech recognition, its lexical items were carefully chosen to include only those in common use. For both databases, monosyllabic words are excluded from our analysis because they do not contribute any links to the networks.

The 21,727 multi-syllabic words in CEDICT comprise 3,773 Chinese characters, 1,240 tonal syllables, and 393 base syllables (all in type). Because each base syllable may have several tones, and each tonal syllable may be shared by many Chinese characters, so the number of tonal syllables is less than that of characters; the number of base syllables is even less. The 35,732 multi-syllabic words in CULEX include 4,942 characters, 1,671 tonal syllables, 614 base syllables. Note that the greater number of characters in Cantonese is mainly due to the size of the CULEX database, which, in terms of the numbers of multi-syllabic words, is about 65% larger than CEDICT.

Numerics of the networks

In this section, the three levels of Chinese networks (base syllable, tonal syllable and character) built from CEDICT and CULEX will be characterized in terms of measures and methods in common use in the study of complex networks: average distance, L , the clustering coefficient, C , as well as the degree distribution. Then, these results will be compared with the ones from corresponding random networks having the same numbers of nodes and links. In the following text, the subscripts CE and CU stand for CEDICT and CUDICT respectively.

The number of connections of each node, i , is denoted by k_i , and is termed the degree of node i . Then the average degree of the network is denoted by,

$$\langle k \rangle = \frac{2K}{N},$$

where the K is the total number of links or connections, and the N is the total number of nodes. The values of N and $\langle k \rangle$ for the three levels of Chinese networks are shown in Table 2.

Networks	N	$\langle k \rangle$	L_{real}	L_{ER}	L_R	D	C_{real}	C_{real}^*	C_{ER}	C_R	γ
Putonghua B.S.	393	104	1.77	1.29	2.39	4	0.61	0.53	0.265	0.74	0.21
Cantonese B.S.	614	109	1.91	1.37	3.31	4	0.54	0.46	0.178	0.74	0.40
Putonghua T.S.	1,240	54.3	2.4	1.78	11.91	5	0.32	0.27	0.044	0.74	0.91
Cantonese T.S.	1,671	60.8	2.34	1.81	14.23	5	0.27	0.20	0.036	0.74	0.97
Putonghua Character	3,773	21.1	3.07	2.71	89.88	8	0.23	0.12	0.006	0.71	1.40
Cantonese Character	4,942	21.2	3.04	2.79	117.0	10	0.19	0.09	0.004	0.71	1.49
Portuguese S_{DIC}	2,285	27.6	2.44	2.33	41.88	6	0.65	N/A	0.012	0.72	1.35
Portuguese S_{MA}	3,188	28.2	2.61	3.40	57.01	8	0.50	N/A	0.009	0.72	1.36

Table 2. The numerics of several networks. B.S. stands for Base Syllable, while T.S. for Tonal Syllable. For each network, we include the number of nodes N , the average degree $\langle k \rangle$, the average distance L , the diameter D , the clustering coefficients C , and the exponent γ of the best-fitting power-law distribution. The numerics of Portuguese networks [Medeiros Soares et al., 2005] are also included for further comparison. L_{real} , C_{real} and C_{real}^* denote values for the observed syllabic

networks; L_{ER} and C_{ER} denote values for the corresponding Erdős-Renyi random networks; L_R and C_R denote values for the corresponding regular networks [Watts, 1999].

The distance between nodes i and j , d_{ij} , is the minimum number of links lying on a path connecting these two nodes. The average distance of node i from all other nodes is

$$d_i = \frac{\sum_j d_{ij}}{N-1} .$$

So the average distance, L , of the network is the average value of d_i over all the nodes:

$$L = \frac{\sum_i d_i}{N} .$$

The diameter is defined as the maximum distance:

$$D = \max_i d_i .$$

The value of L and D for the three levels of Chinese networks are shown in Table 2. For comparison, we also calculate L and D for an Erdős-Renyi random network and a regular network having the same number of nodes and average degree $\langle k \rangle$. This distance can be approximated as $L_{ER} = \frac{\ln(N)}{\ln(\langle k \rangle)}$ [Bollobas, 1985] and $L_R = \frac{N(N + \langle k \rangle - 2)}{2 \langle k \rangle (N - 1)}$ [Watts, 1999], respectively. In all cases, the distance of real networks is close to the distance of the corresponding random network. Therefore, the networks all show the characteristics of small-world effect, popularly known as six degrees of separation [Milgram, 1967].

The clustering coefficient [Watts and Strogatz, 1998], C , is used to measure the interconnectivity of a network. To better understand this measure, let us first consider a selected node i , having k_i links which connect it to k_i other nodes. The ratio between the number E_i of links that actually exist among these k_i neighbors and the total number $\Omega k_i(k_i - 1)$ of possible links among them defines the value of the clustering coefficient of node i :

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

Clearly, $0 \leq C_i \leq 1$ for all $k_i > 1$. If the neighbors of the target node i are all fully connected with each other, the number of actual links will be $k_i(k_i - 1)/2$. Consequently, C_i will have the value 1. Another extreme case is that none of the k_i nodes is connected to any other. Consequently, C_i will be 0. The clustering coefficient of a network is defined as the average clustering coefficient of all the nodes:

$$C = \frac{\sum_i C_i}{N} .$$

Bollobas and Riordan [2003] introduced an alternative clustering coefficient, C^* , which is a weighted sum of the clustering coefficients of each individual node i :

$$C^* = \frac{\sum_i C_i k_i(k_i - 1)/2}{\sum_i k_i(k_i - 1)/2} = \frac{2 \sum_i E_i}{\sum_i k_i(k_i - 1)} .$$

In Table 2, we calculate the clustering coefficients for all three levels of Chinese networks. Meanwhile, we also estimate the clustering coefficient of the ER random network: $C_{ER} = \frac{\langle k \rangle}{N}$ [Bollobas, 1985] and regular network: $C_R = \frac{3(\langle k \rangle - 2)}{4(\langle k \rangle - 1)}$ [Watts, 1999]. In all cases, the clustering coefficients of the real networks are significantly larger than those of the corresponding random networks, and smaller than those of the corresponding regular networks. In the case of the Chinese character networks, the magnitudes of C are tens of times greater than their random counterparts: $C_{CEDICT} \approx 38.3C_{ER}$ ($C^*_{CEDICT} \approx 20C_{ER}$), $C_{CULEX} \approx 47.5C_{ER}$. ($C^*_{CULEX} \approx 22.5C_{ER}$)

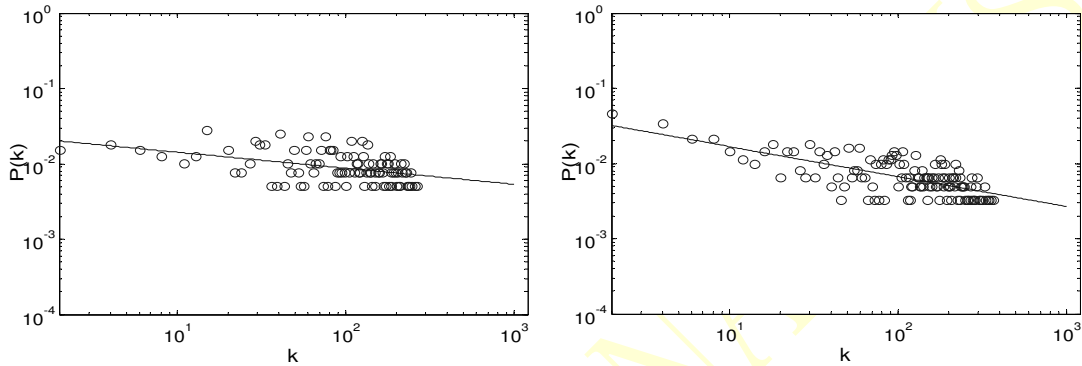


Figure 4. Degree distribution for base syllables of the two datasets, CEDICT (left) and CULEX (right), on log-log scale. Each circle indicates the magnitude of $P(k)$ within the specific range of k . For instance, the left most circle in each figure indicates the magnitude of $P(k)$ for $k \leq 2$; the second left most circle in each figure indicates the magnitude of $P(k)$ for $2 < k \leq 4$. The solid lines indicate the best fitting power-law distributions.

The degree distribution, $P(k)$, measures the proportion of nodes having connectivity k . This measure is a standard method for investigating the structure of various types of complex network. The degree distribution can sometimes provide a clue as to how a network evolved. Therefore, it will be a promising measure for studying the development of the lexicon. The degree distributions or three levels of networks are plotted in Figures 4, 5, and 6 respectively. Each of these plots approaches a power-law. The magnitudes of the best-fitting power-law exponents are shown Table 2.

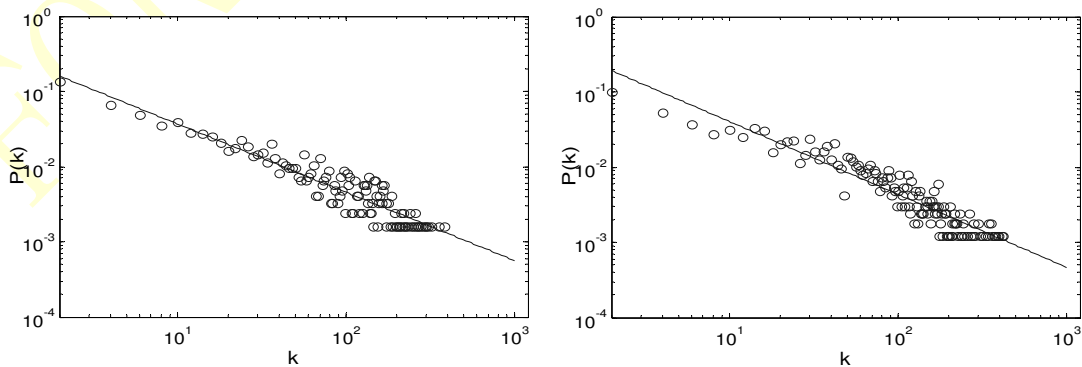


Figure 5. Degree distribution for tonal syllables of the two datasets, CEDICT (left) and CULEX (right), on log-log scale.

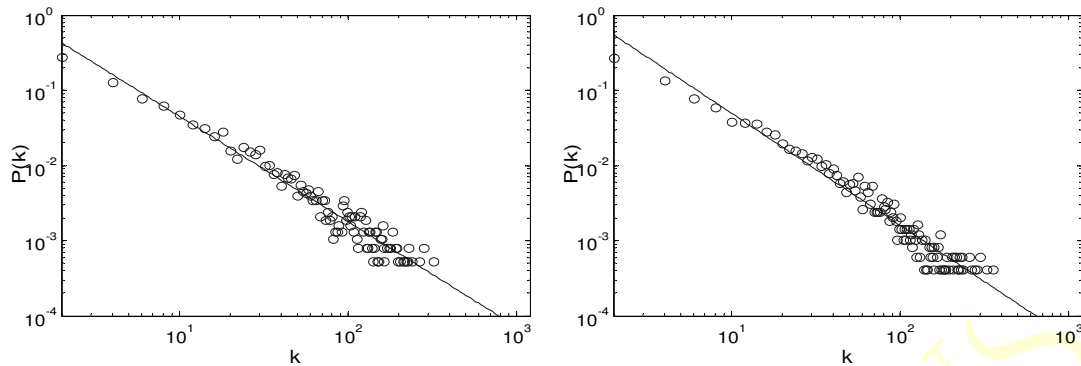


Figure 6. Degree distribution for Chinese characters of the two sets, CEDICT (left) and CULEX (right), in log-log scale.

Discussion

Unlike Erdős-Renyi random networks, which have small clustering coefficients and Poisson-like degree distribution, all three networks for the Putonghua and Cantonese datasets (CEDICT and CULEX) have large clustering coefficients and power-law distribution. Especially at the Chinese character level, the γ values are consistent with the values of γ usually observed in complex networks ($1 < \gamma < 3$) [Albert and Barabási, 2002]. This value is also consistent with variants of preferential attachment network models, indicating that the evolution of the lexicon might follow a rule similar to that of preferential attachment; Medeiros Soares et al. [2005] came to this conclusion regarding the formation of Portuguese syllabic networks.

The “standard” preferential attachment mechanism of Barabási and Albert [1999] operates as follows: For a given initial network, add nodes one-by-one, each time connecting the added node to a fixed number of pre-existing nodes, the nodes to which it connects selected with probability proportional to its degree. In other words, “the rich get richer”. Studying the networks at the base syllable level, the tonal syllable level, and then the character level, we find that the value of the power law exponent, γ , increases, which indicates that the preferential attachment is stronger at the tonal syllable level than at the base syllable level, and stronger still at the character level. This suggests that as new characters were added to the Chinese language over phylogenetic time scales, they tended to combine with existing high-degree characters to form new words. However, the tendency for the corresponding syllable to combine with existing high-degree syllables would have been less pronounced, particularly for base syllables. Thus the preferential attachment is most dominant for the units that most constrain the semantic content, i.e., the characters. Tonal syllables correspond to one or more characters, so constrain the semantics less than do characters. For example, as shown in Table 1, Cantonese “火車” (“train”) is distinct semantically from “貨車” (“wagon”). However, the first character of each word corresponds to the same base syllable “fo”. Thus the characters “火” and “貨” constrain the semantics more precisely than the base syllable “fo”. Also as shown in

Table 1, Putonghua “油井” (“oil well”) and “火警” (“fire alarm”) are semantically distinct. However, the second character of each word corresponds to the same tonal syllable “jing3”. Thus characters also constrain the semantics more precisely than tonal syllables. Base syllables correspond to one or more tonal syllables, so constrain the semantics less precisely than tonal syllables. The pattern of values of the power law exponents therefore reflects, in part, the cognitive processes by which new words might have been constructed in Chinese.

Recall the results for the Portuguese syllabic networks. We find that the degree distributions of the Portuguese syllable networks correspond most closely to those of the Chinese networks at the character level. It is likely that Portuguese syllables, like English syllables, carry semantic information to some extent: for instance, in English, the syllable ‘bi-’ means ‘two’; ‘tri-’ means three, etc. Moreover, the number of commonly used syllables in Portuguese is around 3,000, which is close to the number of commonly used Chinese characters (3,000–4,000).

As the child learning Chinese acquires the language, the cognitive pressure to construct new words may be due to the initial lack of lexical items to precisely describe some newly formed compact ideas, i.e., the semantic content of words comes before the words themselves (although, of course, words can undergo semantic extension later). Then, preferential attachment would occur more often at the character level than at other levels.

As yet, we can provide no better explanation for the observed structure of the Chinese syllable and character networks than that suggested by Medeiros Soares et al. (2006) regarding Portuguese syllable networks: that the power-law degree distribution results from a process of preferential attachment with growth. We should point out, however, that power law degree distributions can be generated by other network growth procedures [Kumar, et al., 2000]. Further research may help to illuminate how the syllable networks of different languages come to have power-law degree distribution.

This work only analyzes two dictionary-like datasets. The use of additional data for various time points throughout the evolution of a language would allow us to obtain a much clearer picture of how the lexicon grows. Achieving this is problematic. First, there is no well-organized database comprising the contents and structure of the lexicon for any particular language over, say, the past thousand years (although the long written history of Chinese might allow such a database to eventually be constructed). Second, spoken language must have been mature before the emergence of written language. Nevertheless, we can study further the evolution of the lexicon from two other aspects. On the one hand, we can investigate the course of lexicon acquisition by children, which also exhibits a process of transition from no lexicon to a mature lexicon. This process may be viewed as a window to language evolution. On the other hand, computing models have recently come to be used extensively to study the evolution and emergence of language [Gong, et al, 2005], and may also be used in the future to study the evolution of the lexicon.

References

- Albert, R. and Barabási, A. L. (2002). Statistical mechanics of complex networks. *Review of Modern Physics*, 74: 47.
- Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286: 509–512.
- Bollobas, B. (1985). *Random Graphs*. Academic Press, London.
- Bollobas, B. and Riordan, Oliver M. (2003). Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks: From the Genome to the Internet*. Bornholdt, S. and Schuster, Heinz G. (eds.). Wiley-VCH, Weinheim, 1-34.
- CCDICT (2000). *Dictionary of Chinese Characters, Version 3.0*
<http://www.chinalanguage.com/CCDICT/>.
- Denisowski, P. (2005). CEDICT: Chinese-English Dictionary. (Online). Available from <http://www.Putonghuatools.com/cedict.html>.
- Dorogovtsev, S. N. and Mendes, J. F. F. (2001). Language as an evolving word web. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1485): 2603–2606.
- Ferrer i Cancho, R. and Solé, R. V. (2001). The small-world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1485): 2261–2266.
- Ferrer i Cancho, R. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *PNAS USA*, 100: 788–791.
- Gong, T., Minett, J. W., Ke, J., Holland, J. H., and Wang, W. S-Y. (2005). Coevolution of lexicon and syntax from a simulation perspective. *Complexity*, 10(6): 50–62.
- Kumar, R., Raghavan, P., Rajagopalan S. and Sivakumar, D. (2000). Stochastic models for the web graph. *FOCS: IEEE Symposium on Foundations of Computer Science. Hierarchical Organization of Modularity in Metabolic Networks*. *Science*, 297:1551–1555.
- Lee, T., Lo, W. K., Ching, P. C., and Meng, H. (2002). Spoken language resources for Cantonese speech processing. *Speech Communication*, 36: 327–342.
- Linguistic Society of Hong Kong (LSHK). (1997). *Hong Kong Jyut Ping Characters Table*. Linguistic Society of Hong Kong Press.
- Milgram, Stanley. (1967). The Small World. *Psychology Today* 2: 60–67.
- Peng, G. and Wang, W. S-Y. (2004). An innovative prosody modeling method for Chinese speech recognition. *International Journal of Speech Technology*, 7: 129–140.
- Medeiros Soares, M., Corso, G., and Lucena, L. S. (2005). The network of syllables in Portuguese. *Physica A*, 335: 678–684.
- Sigman, M. and Cecchi, G. (2002). Global organization of the Wordnet lexicon. *PNAS USA*, 99: 1742–1747.
- Steyvers, M. and Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model for semantic growth. *Cognitive Science*, 29: 41–78.
- Wang, W. S-Y. (1973). The Chinese language. *Scientific American*, 228: 50–60.
- Watts, D. J. (1999). *Small World: The Dynamics of Networks Between Order and Randomness*. Princeton University Press.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of “small world” networks.
Nature, 393: 440–442.

FOR PERSONAL USE