Key nodes mining of root network and phoneme network of modern Chinese characters

Rong Huang¹, Quanyu Zou², Kaiyu Xu², Mingyang Cai², Wenjun Zhang³, Weibing Deng³ and Bo Liu⁴

¹ National Bases for Basic Scientific Research and Talents Cultivation, College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China
² College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China
³ Key Laboratory of Quark and Lepton Physics (MOE) and Institute of Particle Physics, Central China Normal University, Wuhan 430079, China
⁴ Engineering Training Center, School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan 430070, China

E-mail: wenjun@mails.ccnu.edu.cn and wdeng@mail.ccnu.edu.cn

Abstract. Roots and phonemes are two important ingredients (but not independent ones) when learning the Chinese language. In order to optimize the learning strategy, various key nodes mining algorithms are implemented both in the root network (RN) and phoneme network (PN) of modern Chinese characters. The lists of important roots and phonemes, and eventually the important characters are dug out and suggested. Rather than the commonly used frequency dictionary of characters, one would find that results from the key nodes mining are more systematic. Moreover, roots and phonemes are studied from perspective of the multi-layer network, in the interest of exploring the correlation between them, which are also very useful when learning the Chinese language.

1. Introduction

Considering human languages as the dynamical complex systems, quantitative linguistics, or mathematical linguistics [1, 2, 3] deals with the language statistics, language structure, language evolution, language learning, etc, by using the methods of probability theory [4, 5, 6, 7, 8], machine learning or deep learning [9], statistical physics [3], complex networks [2] and so on. These studies are useful for understanding the origin of human languages [10], the physical and mathematical laws behind the language phenomena [11], the self-adaptive mechanisms of languages [12], the reasons embedded in the evolution of languages [2], the optimization of information transfer [13], etc. They are also applicable for some concrete problems of natural language processing, such as text classification, authorship attribution, information searching, machine translation, etc [14, 15, 16, 17].

Complex networks are very powerful methods for studying human languages [2], relevant works have been rapidly and vastly carried out in the past decades (see the literatures collection website [18] of this direction-Physicists’ papers on natural language from a complex systems viewpoint, 1989-2018). The researches in this area can be roughly divided into three categories:
1) Exploring the multi-level features of human languages: language system is a multi-level complex system, from different levels one can get different kinds of language networks [19, 20, 21]. For example, by using the same set of sentences of Chinese, the following three levels of networks can be constructed: (I) Chinese character co-occurrence network; (II) Chinese syntactic network; (III) Semantic role network. The nodes in (I) are characters, while the links are formed between adjacent characters in a sentence, it can be used to study the formation mechanism of Chinese vocabulary. The nodes in (II) are words, and the links are determined by the syntactic functional relationship. Network III is constructed by a set of sentences that have been tagged with semantic roles, namely the Chinese semantic network. Cancho and Solé [22], Motter and Moura [23], Liu and Cong [19] studied and compared these three kinds of networks, analyzed the different cognitive mechanisms of information expression [24].

2) Using the topological quantities of networks to study the linguistic typology: linguistic typology not only studies the classification of languages, but also pay much attention to find the common features of human languages. Liu et al. constructed the syntactic networks based on two different languages, and found that syntactic networks with the same diameter would have obvious different average degree, the power law exponent of degree distribution, the clustering coefficient, the average shortest path, and the centrality [25]. Later on, they constructed 15 syntactic networks of different languages, and compared the topological parameters of different networks to classify different languages [24]. Amancio et al. extended the methods of linguistic classification and text categorization based on topological parameters of complex networks to the dynasty of literary works, assessment of translation quality, discrimination of language complexity, etc [26].

3) Analyzing the inter-relations among different scales of human languages: the language networks can be studied from three different scale, namely the global scale, local scale and individual scale. Most researches on language networks are based on the global network and focus on the macroscopic global features of language networks. While at the local and individual scales, there are few studies. Therefore, it is an important scientific question to establish the relationship between the macroscopic overall complexity of language networks and the microscopic individual grammatical features. Some recent works have begun to address these issues. For example, Cancho et al. applied spectral analysis to study the community structure of syntactical-dependent networks and found that similar type of words have obvious aggregation behavior [27]. Siew et al. detected the community structure of the speech network, and found that the size of the community has the strong correlation with the length and frequency of words [28].

It is widely recognized that learning Chinese is much more difficult than learning Indo-European languages, especially for people whose native language belongs to the Indo-European language family. While the main difficulty lies in learning to read and write Chinese characters, that is, two most important ingredients—the root and phoneme of Chinese characters. In this work, we dedicated to improve the learning efficiency of Chinese characters via the network approach, i.e., by employing the key nodes mining algorithms, since they are rather mature and commonly applied in various situations. Rather than the traditional frequency dictionary learning strategy, we would like to provide the more complete and reasonable learning lists of roots and phonemes, and explore the interrelationship between roots and phonemes.

This paper is organized as follows. In the next section, the description of the Chinese writing systems and some linguistics concepts are introduced. In section 3, the construction of root network (RN) based on the commonly used Chinese characters is discussed, then some topological properties are presented, such as degree distribution, the hierarchical structure, the community structure, etc. The key nodes mining results are based on various quantities, such as degree, weight, centrality, percentage of network penetration, and so on. Introductions of the Chinese Pinyin and phonemes are given in section 4. In section 5, corresponding results are
discussed and presented for the phoneme network (PN) of modern Chinese characters. In section 6, in order to uncover the interrelations between root and phoneme when people are learning the Chinese characters, the multi-layer network is proposed between roots and phonemes, and some learning strategies are suggested.

2. Chinese Writing System
The writing system [7] is the process or result of recording spoken language using a system of visual marks on a surface. There are mainly two types of writing systems: logographic (Sumerian cuneiforms, Egyptian hieroglyphs, Chinese characters) and phonographic. The latter includes syllabic writing (e.g. Japanese hiragana) and alphabetic writing (English, Russian, German), while the former encodes syllables and phonemes.

The unit of Chinese writing system is the character: a spatially marked pattern of strokes phonologically realized as a single syllable. Generally, each character denotes a morpheme or several different morphemes. The Chinese writing system evolved by emphasizing the concept of the character-morpheme, to some extent blurring the concept of the multi-syllable word.

Chinese characters are logographs primarily used in the writing of Chinese, they represent words of the language using several strategies. A few characters, including some of the most commonly used, were originally pictographs, which depicted the objects denoted, or ideograms, in which meaning was expressed iconically. The vast majority were written using the rebus principle, in which a character for a similarly sounding word was either simply borrowed or (more commonly) extended with a disambiguating semantic marker to form a phono-semantic compound character.

Psycholinguistic research [6, 7] shows that the characters are important cognitive and perceptual units for Chinese writers and readers, e.g. Chinese characters are more directly related to their meanings than English words to their meanings. The explanation of this effect would be that characters (compared to English words) are perceived holistically as a meaning-carrying objects, while English words are yet to be reconstructed from a sequence of their constituents (phonemes and syllables).

3. Root Network of Modern Chinese Characters
The RN is constructed based on the commonly used 7000 modern Chinese characters from the frequency dictionary. Each character is separated into its basic root (or morpheme) according to the roots-combining-rules. For example, the Chinese character “机” can be separated as “木” and “几”. We need to stress here that our current data of roots of modern Chinese characters is achieved by the authors themselves, not by the machine, therefore it is rather accurate, reliable and complete. Then we denote each root as a single node, if two or more roots co-occurrent in the same character, we assign the full links among them. By doing this for all the 7000 modern Chinese characters, the RN is constructed, with 461 nodes and 8299 links.

The degree of a node is the number of nodes connecting with this node. The degree distribution of RN is shown in Fig. 1, with \( P(k) \sim k^{-1.89} \), which means the degree distribution of RN is power law distribution and the RN obeys the scale-free network. The clustering coefficient is calculated, which could represent the closeness of nodes. We study the relation between clustering coefficient and degree of RN and find that they have power law relation, \( C(k) \sim k^{-0.34} \), which is shown in Fig. 2. This means that RN has the underline hierarchical structure.

Then the average shortest path distance and average clustering coefficient of RN was compared with those of the Erdős–Rényi network. Results were shown in Table 1. \( \sigma \), which can be used to judge whether a network has the small-world property [29], can be calculated as \( \sigma = (C_{RN})/(C_{ER}) \), where \( C \) is average clustering coefficient of RN, \( C_{ER} \) is average clustering coefficient of ER network, \( L \) is the average shortest path distance of RN, and the \( L_{ER} \) is the
Figure 1. (Color online) The degree distribution of RN, with \( P(k) \sim k^{-1.89} \). It belongs to the scale free network.

Figure 2. (Color online) The relations between the clustering coefficient and degree, \( C(k) \sim k^{-0.34} \). This means that the RN has the underline hierarchical structure.

Table 1. The average shortest path distance and average clustering coefficient of RN and ER network.

| Network     | Average path distance | Average clustering coefficient |
|-------------|-----------------------|-------------------------------|
| RN          | 2.24                  | 0.39                          |
| ER network  | 6.18                  | 0.052                         |

average shortest path distance of ER network. We can get the \( \sigma = 20.52 > 1 \), this shows that the RN has the small-world property.

Then the centrality of RN was studied, such as the degree centrality \( k_i = \sum_{j=1}^{N} a_{ij} \), eigenvector
Table 2. The top 15 Degree Centrality, Eigenvector Centrality, Node Weight, and Root Frequency of RN. Full list of rankings would be found via this link [39].

| Character | Degree | Character | Eigenvector | Character | Weight | Character | Frequency |
|-----------|--------|-----------|-------------|-----------|--------|-----------|-----------|
| 口        | 204    | 口        | 1           | 口        | 0.018  | 口        | 0.056     |
| 氵        | 167    | 氵        | 0.926       | 氵        | 0.016  | 一        | 0.042     |
| 木        | 153    | 木        | 0.914       | 木        | 0.015  | 人        | 0.026     |
| 氵        | 148    | 氵        | 0.908       | 氵        | 0.015  | 人        | 0.024     |
| 月        | 130    | 月        | 0.824       | 月        | 0.014  | 月        | 0.023     |
| 名        | 114    | 名        | 0.79        | 名        | 0.013  | 名        | 0.022     |
| 土        | 110    | 土        | 0.786       | 土        | 0.012  | 土        | 0.02      |
| 木        | 100    | 木        | 0.782       | 木        | 0.012  | 木        | 0.019     |
| 氵        | 92     | 氵        | 0.743       | 氵        | 0.011  | 月        | 0.016     |
| 一        | 92     | 一        | 0.74        | 一        | 0.011  | 一        | 0.015     |
| 女        | 90     | 女        | 0.698       | 女        | 0.01   | 女        | 0.014     |
| 人        | 79     | 人        | 0.693       | 人        | 0.01   | 人        | 0.012     |
| 木        | 76     | 木        | 0.663       | 木        | 0.009  | 木        | 0.012     |
| 人        | 76     | 人        | 0.659       | 人        | 0.009  | 人        | 0.012     |

The results were shown in Table 2. The average degree centrality and average eigenvector centrality of RN are 18.78 and 0.166 respectively. We also compare these results with the node weight of RN, where node weight is defined as the total weight of the edges connected with this node. One would find that the rankings of those centrality are quite similar, and the top 200 nodes occupy 81.38% weight of the network. We also compare them with the statistic result of frequency of roots in the frequency dictionary. The top 15 most frequent roots are also similar with the above results. While the top 200 most frequent roots occupy 92.1% of the network. This means that we just need to learn the top 200 roots in most cases, then going to the characters the next step.

The community structure shows the different functioning parts of the network, and the nodes in same community connect more densely. This concept was proposed by Newman [30], and the corresponding algorithm was also proposed. The popular algorithms are infomap [31], modularity maximization algorithm [32], and the Label Propagation Algorithm with compression of Flow (LPAF) [33] nowadays. The modularity maximization algorithm was chosen to explore RN, since it has the highest modularity, which could be used to judge the quality of the community structure detection. The community detection result is shown in Fig 3. One could find that the RN has very clear community structures, while the core nodes in each community are also the important nodes in the network. One would pay due attention to these nodes, and learn the roots in the communities one by one, so as to get a more systematic studying. Full lists of roots in each community would be also found via this link [39].

4. Chinese Pinyin versus Phonemes

Language can be viewed as a hierarchic construction: phoneme, syllable, morpheme, word... Each of these objects expresses meaning or participates in its formation, and consists of elements of the previous level, i.e. syllable consists of phonemes. The lowest hierarchic level is phoneme [8], which is defined to be a representative for a group of sounds that are not distinguishable with respect to their meaning-formation function in a concrete language.

The meaning is crucial for the definition of the phoneme, although a single phoneme does not express a separate meaning. The next hierarchic level (syllable) indirectly participates in the
Figure 3. (Color online) The community structure of RN, nodes with the same color belong to one community, and the size of nodes implies the degree of the node.

definition of the phoneme, since the syllable bounds phonemes, i.e. there cannot be a phoneme which belongs to two different syllables; e.g. diphthongs belong to the same syllable.

Chinese language is a tonal language. The pronunciation and spelling of Chinese characters are generally given in terms of initials and finals, which represent the segmental phonemic portion of the language, rather than letter by letter. Initials are initial consonants, while finals are all possible combinations of medials (semivowels coming before the vowel), the nucleus vowel, and coda (final vowel or consonant).

5. Phoneme Network of Modern Chinese Characters
The major processes of constructing the PN are: 1) Translate the Chinese character into its corresponding Pinyin (with totally 1214 different pronunciations ); 2) Split the initial consonant and simple or compound vowel as different phonemes (nodes); 3) Assign the full links among the phonemes if they co-occurrent in the Pinyin of a character. There are 161 nodes and 1350 edges in the PN.

We deal with the PN in the same way as we did with RN. The results of top 15 degree, eigenvector centrality, and node weight are shown in Table 3. We also study the community structure of PN, as is shown in Fig 4. Here we also dedicated to create the frequency dictionary of pronunciations base on the 7000 modern Chinese character (see Table 3). These results (both the key nodes mining and frequency dictionary of pronunciations) are very useful for studying the pronunciations of Chinese characters.
Table 3. The top 15 Degree centrality, Eigenvector centrality, Nodes weight, and Frequency dictionary of PN. Full list of rankings would be found via this link [39].

| Nodes | Degree | Nodes | Eigenvector | Nodes | Weight | Pronunciation | Frequency |
|-------|--------|-------|-------------|-------|--------|---------------|-----------|
| y     | 49     | t     | 1           | l     | 74     | de            | 9080142   |
| l     | 48     | ch    | 0.986       | t     | 65     | o4            | 7034705   |
| x     | 46     | l     | 0.976       | ch    | 63     | an4           | 6337512   |
| zh    | 44     | d     | 0.965       | d     | 60     | e4            | 6297696   |
| j     | 43     | zh    | 0.95        | h     | 59     | da            | 5004554   |
| d     | 41     | sh    | 0.946       | y     | 57     | a1            | 4747652   |
| h     | 41     | h     | 0.934       | m     | 57     | e2            | 4490919   |
| q     | 37     | y     | 0.901       | g     | 57     | ao4           | 4468126   |
| sh    | 37     | g     | 0.9         | b     | 57     | shi4         | 4308697   |
| t     | 34     | n     | 0.821       | zh    | 57     | ai4          | 4140804   |
| g     | 33     | c     | 0.818       | p     | 56     | zhe          | 4065755   |
| m     | 33     | b     | 0.816       | n     | 55     | an1          | 3942001   |
| ch    | 32     | z     | 0.812       | sh    | 55     | ang4         | 3459766   |

Figure 4. (Color online) The community structure of PN, in which the different colors represent different communities and the nodes size represents the degree of the node.
6. Multilayer Network of Roots and Phonemes of Modern Chinese Characters

Multilayer network, a new method for studying the relations of complex systems, has attracted more and more attention [34]. There are some common types of multilayer networks, such as interdependent networks [35], in which the characteristics of nodes in two or more monoplex networks are adjacent to each other via edges that are called dependency edges; networks of networks [36, 37], in which characteristics of each node are represented in different networks; Multiplex networks [38], a multilayer network whose layers are a sequence of graphs, different layers contain the same nodes.

We need to emphasize that roots and phonemes are not independent ingredients of Chinese characters. Therefore, the interdependent networks was employed and constructed for studying the relations of roots and phonemes of modern Chinese characters. The specific method is to set the roots and phonemes as two layers, the interconnection represents that the roots have that typical pronunciations, and the layer is same as the RN and PN. Several examples of the interdependent networks were shown in Fig. 5. It is indeed found that several roots preferentially connects to certain phonemes. Thus, this provide a potential way to study the roots and phonemes more efficiently and systematically.

![Multilayer network of li4.](image1)

(a) The multilayer network of li4.

![Multilayer network of qi2.](image2)

(b) The multilayer network of qi2.

![Multilayer network of yi4.](image3)

(c) The multilayer network of yi4.

![Multilayer network of e4.](image4)

(d) The multilayer network of e4.

Figure 5. (Color online) The four examples of the interdependent networks of roots and phonemes of modern Chinese characters.

7. Conclusions

The modern Chinese characters are studied by the network approach via two basic linguistics units: roots and phonemes, in the interest of improving the learning strategy of Chinese language.
Key nodes mining algorithms are employed to dig out the important roots and phonemes, such as degree centrality, eigenvector centrality, weight and percentage of network occupation, etc. The full list of frequency dictionary of pronunciations of Chinese characters is also obtained. It is also the first time that tries to draw the relations between roots and phonemes.

References

[1] Altmann, E. G., and Gerlach, M. (2016). Statistical Laws in Linguistics. Creativity and Universality in Language. Springer International Publishing.
[2] Cong, J., and Liu, H. (2014). Approaching human language with complex networks. Physics of Life Reviews, 11(4), 598.
[3] Castellano, C., Fortunato, S., and Loreto, V. (2009). Statistical physics of social dynamics. Review of Modern Physics, 81(2), 591-646.
[4] Hu C. K., Kuo W. C. (2005). Universality and Scaling in the Statistical Data of Literary Works (POLA Forever), pp. 115–139.
[5] Chi T. T., Wang S. C., Hu C. K. (2015). Word population analysis and other evidences indicate that Shiji was amended by Liu Xiang. Physical Review E Statistical Nonlinear & Soft Matter Physics, 88(6), 062804-062804.
[6] Allahverdyan, A. E., Deng, W., and Wang, Q. A. (2013). Explaining zipf's law via a mental lexicon. Physical Review E Statistical Nonlinear & Soft Matter Physics, 88(6), 062804-062804.
[7] Deng, W., Allahverdyan, A. E., Li, B., and Wang, Q. A. (2014). Rank-frequency relation for chinese characters. European Physical Journal B, 87(2), 1-20.
[8] Deng, W., Allahverdyan, A. E., (2016). Stochastic Model for Phonemes Uncovers an Author-Dependency of Their Usage. PLoS ONE, 11(4) e0152561.
[9] Lecun, Yann, Bengio, and Hinton. "Deep learning." Nature 521.7553(2015):436.
[10] Stephens, G. J., and Bialek, W. (2010). Statistical mechanics of letters in words. Physical Review E Statistical Nonlinear & Soft Matter Physics, 81(2), 066119.
[11] Bakalis, E., and Galani, A. (2012). Modeling language evolution: aromanian, an endangered language in greece. Physica A Statistical Mechanics & Its Applications, 391(20), 4963-4969.
[12] Solé, R. V. (2003). Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences of the United States of America, 100(3), 788-91.
[13] Cancho, R. F. I. (2005). Zipf’s law from a communicative phase transition. The European Physical Journal B - Condensed Matter and Complex Systems, 47(3), 449-457.
[14] Amancio, D. R., Nunes, M. G. V., Jr, O. N. O., Pardo, T. A. S., Antúnez, L., and Costa, L. D. F. (2011). Using metrics from complex networks to evaluate machine translation. Physica A: Statistical Mechanics & Its Applications, 390(1), 131-142.
[15] Régo, H. H., Braunstein, L. A., D’Agostino, G., Stanley, H. E., and Miyazima, S. (2014). When a text is translated does the complexity of its vocabulary change? translations and target readerships. Plos One, 9(10), e110213.
[16] Serra J., Kartz H., Serra X. and Andrzejak R. G., Predictability of music descriptor time series and its application to cover song detection. IEEE Transactions on Audio, Speech, and Language Processing 20, 2 (2011).
[17] Basile, C., Benedetto, D., Caglioti, E., and Esposti, M. D. (2008). An example of mathematical authorship attribution. Journal of Mathematical Physics, 49(12), 237.
[18] Eduardo A A, Martin G. Physicists’ papers on natural language from a complex systems viewpoint. http://www.maths.usyd.edu.au/u/ega/physicist-language/1989-2018.
[19] Liu, H. T., and Jin, C. (2013). Language clustering with word co-occurrence networks based on parallel texts. Chinese Science Bulletin, 58(10), 1139-1144.
[20] Liu, Z., Zheng, Y., and Sun, M. (2008). Complex network properties of chinese syntactic dependency network. Complex Systems & Complexity Science.
[21] Liang, W., Shi, Y., and Huang, Q. (2014). Modeling the chinese language as an evolving network. Physica A Statistical Mechanics & Its Applications, 393(1), 268-276.
[22] Cancho, R. F. I., Solé, R. V., and Kübler, R. (2004). Patterns in syntactic dependency networks. Phys Rev E Stat Nonlin Soft Matter Phys, 69(1), 051915.
[23] Motter, A. E., de Moura, A. P., Lai, Y. C., and Dasgupta, P. (2002). Topology of the conceptual network of language. Physical Review E Statistical Nonlinear & Soft Matter Physics, 65(2), 065102.
[24] Lieberman, E., Michel, J. B., Jackson, J., Tang, T., and Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. Nature, 449(7163), 713-6.
[25] HaiTao Liu, and WenWen. (2010). Language clusters based on linguistic complex networks. Chinese Science Bulletin, 55(30), 3458-3465.
[26] Amancio, D. R., Aluisio, S. M., Jr, O. N. O., and Costa, L. D. F. (2013). Complex networks analysis of language complexity. Epl, 100(100), 58002.
[27] Ramon Ferrer i Cancho, Andrea Capocci, and Guido Caldarelli. (2007). Spectral methods cluster words of the same class in a syntactic dependency network. International Journal of Bifurcation & Chaos, 17(07), 2453-2463.
[28] Siew, C. S. (2013). Community structure in the phonological network. Frontiers in Psychology, 4(4), 553.
[29] Humphries M D, Gurney K. Network ‘small-world-ness’: a quantitative method for determining canonical network equivalence. Plos One, 2008, 3(4): e0002051.
[30] Girvan M, Newman M E J. Community structure in social and biological networks. Proc. Nat. Acad. Sci., 2002, 99(12):7821-7826.
[31] Rosvall M, Bergstrom C T. Maps of Random Walks on Complex Networks Reveal Community Structure. Proc. Nat. Acad. Sci., 2007, 105(4): 1118-1123.
[32] Newman M E. Fast algorithm for detecting community structure in networks. Phys. Rev. E, 2004, 69: 066133.
[33] Han J, Li W, Su Z, et al. Community detection by label propagation with compression of flow. European Physical Journal B, 2016, 89(12):272.
[34] Domenico M D, Biamonte J. Spectral entropies as information-theoretic tools for complex network comparison. Physical Review X, 2016, 6(4).
[35] Kiveli M, Arenas A, Barthelemy M, et al. Multilayer networks. Ssrn Electronic Journal, 2013, 2(3): 261-268.
[36] Gao J, Buldyrev S V, Stanley H E, et al. Networks formed from interdependent networks. Nature Physics, 2012, 8(1): 40-48.
[37] Ferreira, M. A. M. (2014). Networks of Networks: The Last Frontier of Complexity-A Book Review. International Journal of Latest Trends in Finance and Economic Sciences, 4(1), 708.
[38] Battiston, F., Nicosia, V., & Latora, V. (2013). Metrics for the analysis of multiplex networks. Physical Review E, 89, 032804.
[39] https://github.com/cwj31415926/Chinese-language-data