Japanese Kanji Characters are Small-World Connected Through Shared Components

Mark Jeronimus*, Sil Westerveld†, Cees van Leeuwen‡, Sandjai Bhulai§ and Daan van den Berg¶

* Airsupplies Nederland BV, The Netherlands, Email: mark.jeronimus@gmail.com

† Nishino, Amsterdam, The Netherlands, Email: research@silwesterveld.com

‡ Laboratory for Perceptual Dynamics, KU Leuven, Leuven, Belgium, Email: Cees.vanLeeuwen@kuleuven.be

‡ Center for Cognitive Science, TU Kaiserslautern, Kaiserslautern, Germany

§ Faculty of Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, Email: s.bhulai@vu.nl

¶ Docentengroep IvI, Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands, Email: d.vandenberg@uva.nl

***Abstract*—We investigate the connectivity within different incremental sets of Japanese Kanji characters. Individual characters constitute the vertices in the network, components shared between them provide their edges. We find the resulting networks to have a high clustering coefficient and a low average path length, characterizing them as *small worlds*. We examine the statistical significance of these findings and the role of the degree distributions. We review the evidence that the small-world topologies of these networks are due to the successive elimination of components in the writing system and discuss the implications of the results for language evolution.**

Keywords–Japanese characters; Kanji; components; radicals; small-world networks; phase transition; Zipf's law; Gelb's hypothesis

I. INTRODUCTION

Small-world networks are sparsely connected networks that have a high cluster coefficient (CC) in combination with a low average path length (APL) [1]. The CC on a vertex A which is adjacent to vertices $B_1, \ldots, B_n$, is the number of edges between nodes $B_1, \ldots, B_n$ divided by the maximum of $n(n-1)/2$. As such, the CC on A expresses A's local interconnectivity; the CC of a network is the average of all its vertices. The APL is defined as the average number of edges in the shortest path between all pairs of vertices in the network, and as such expresses its global connectivity.

In real-life, small-world networks have been found in a broad variety of fields: power grids [1], neuronal networks in nematode worms [2], the primate brain [3] [4], the World Wide Web [5], and networks of social relationships [6]. Some evidence suggests that small-world topologies are an emergent property resulting from self-organization in a population of communicating agents [7]–[11]. Small worlds have also been found in language networks of co-occurring words [12], and even more specifically, in far eastern writing systems. An investigation in Chinese characters sharing 'radicals' [13] appears to be closest to ours. These authors investigated the network topology of modern-day Chinese characters and found small-world properties, as well as a non-Poisson degree distribution. Even though Chinese and Japanese characters differ considerably nowadays, computational results of these authors are comparable to ours and others in the field. On a slightly higher level, various research teams constructed networks of co-occurring characters [14], words [14]–[16] and phrases [17] in Chinese. Like [13], these authors find small-world properties, possibly indicating that the same self-organizing forces shaping logographic languages at character level are also shaping writing systems on a larger scale.

Interestingly enough, a similar word level investigation was conducted in Japanese two-Kanji words as well [18] [19]. Despite the difference in characters and methods, these authors also find small-world networks, affirming consistent sharing of characters between words in logographic languages. But as it turns out, an investigation of network topologies in Japanese at character level is still missing. It is this gap that our investigation hopes to fill, conjoining all aforementioned investigations, and as such interconnecting the field of research on network structures in Japanese and Chinese writing systems at both word and character level.

The structure of the paper is as follows. We discuss the Japanese writing system in Section II. We then proceed to show that Kanji is a small-world network in Section III. In Section IV we state our conclusions, provide a discussion on the results, and discuss possible extensions of our work.

II. THE JAPANESE WRITING SYSTEM

A writing system reflects the history of the civilization in which it emerged, and some writing systems have developed a striking level of complexity. The Japanese language, notably, employs four character sets: Hiragana, a 46-piece syllabic script; Katakana, also 46 characters, is similar to Hiragana though mainly used for foreign words, expressions and emphases; Kanji, a logographic symbol script related to the Chinese characters, and finally Romaji, the Roman alphabet, used mostly for numbers, advertisements and in pop culture. All four character sets are represented in the following sentence:

マークは明日、月曜１０時にあの寺で待っています。

Tomorrow, Monday, at 10 o'clock, Mark
will be waiting near that temple

 ISBN: 978-1-61208-603-3

Table I. THREE TIMES THE CHARACTER FOR 'FUN' OR 'ENTERTAINING'. NOTICE THE DIFFERENCE IN COMPOSITIONAL STRUCTURE, ESPECIALLY REGARDING THE 'THREAD'-COMPONENT (THE 'LITTLE SIDEBURNS' IN THE TRADITIONAL CHARACTER).

乐	simplified (modern day) Chinese
樂	traditional Chinese, Cantonese, Taiwanese
楽	Japanese

The first three characters: マーク, 'Mark', are Katakana; the number 10 is written in Romaji. The characters: は, に, あ, の, で, っ, て, い, ま, and す are Hiragana. The remaining characters: 明, 日, 月, 曜, 時, 寺, and 待 are Kanji. Japanese words are usually comprised of Katakana only (マーク), Hiragana only (あの), Kanji only (月曜, 時, 寺) or a combination of Kanji and Hiragana (待って). In Kanji-only words, combinatorial deployment of characters shows close correspondence to word compounds in other languages. For instance, the single character word for 'gold' (金) and the single character word for 'fish' (魚) are commonly combined into a single two-Kanji word 金魚, meaning 'goldfish'. Fishing (釣) and stick (竿) make 'fishing rod' (釣竿). Estimations for the total number of existing Kanji characters range from 40,000–100,000 and new characters could theoretically still be added today [20], but the vast majority of these characters are rarely used. Although all far eastern logographic languages are thought to stem from the same source, there are considerable differences between Japanese Kanji, Chinese characters, and the writing systems in Taiwan and Hong Kong nowadays. Japan has some unique Kanji and a post-war simplification effort in China resulted in a substantial difference between the sets (see Table I). Japanese, Cantonese (from Hong Kong) and Taiwanese characters did not undergo such simplification, but nonetheless diverged over time, and are different from Japanese Kanji too.

Many complex Kanji characters can be seen as compounds of elementary building blocks. We will call these building blocks *components*, and a clear distinction should be made from a Kanji's *radical*, which is traditionally the Kanji's component used for dictionary indexing. As an example, the single-component character 日 (meaning 'day' or 'sun') and the single-component character 月 (meaning 'moon') are combined into a two-component character 明, which means 'bright'. Only the sun-component, however, is considered to be its radical. Both Japanese and Chinese dictionaries traditionally recognize 214 radicals, but many modern electronic Kanji dictionaries employ a 252-piece component file, from which any and every combination of components can be selected for character lookup. It is this 252-piece set, which has considerable overlap with but is not identical to the traditional 214-piece radical set, that was used for this investigation. The exact specification of the 252-component KRADFILE can be found on [21].

Japanese Kanji is organized into several cumulative sets. The KyouIku ("education") is a 1,006-piece set of commonly used Kanji maintained by the Japanese ministry of education. It covers roughly 90% of the Kanji used in the Japanese corpus and is used to determine which characters should be learned by Japanese children in each year of elementary school. The JouYou (lit.: "commonly used") is a set of 1,945 Kanji characters and has also been maintained by the Japanese

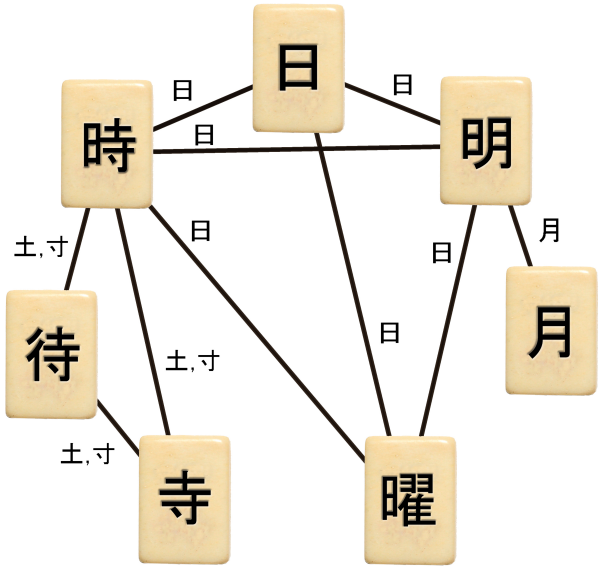

Figure 1. Graph of the Kanji from the example sentence mentioned in the introduction. Vertices represent individual Kanji characters, connected if they share at least one component as identified by the label.

ministry of education, since 1981. It is a superset of KyouIku, extending it by 939 characters learned in secondary school, covering 98.66% of the Kanji used in the Japanese corpus and contains all Kanji allowable in governmental documents. Finally, the JIS X.0208 is a Japanese Industrial Standard defining a 6,355-piece character set, which extends the JouYou by another 4,410 characters, covering 99.98% of all Kanji characters used. Our focus will be on these three character sets, in particular with their intrinsic structures. These structures, as we will show, share characteristic properties with other spontaneously evolved self-organized complex systems.

III. KANJI IS A SMALL-WORLD NETWORK THROUGH SHARED COMPONENTS

We may envisage the cumulative sets of Kanji characters as networks, in which the vertices are characters connected by an edge if they share at least one component (see Figure 1).

An undirected network of n vertices can have a maximum of $n(n-1)/2$ edges; all three networks have a small fraction of this, classifying them as sparse. Omitting disconnected vertices, the KyouIku network has 1,004 nodes and 73,173 edges (density 14.53%), the JouYou network has 1,943 nodes and 292,234 edges (density 15.49%), the JIS.X.0208 has 6,355 nodes and 3,354,225 edges (density 16.61%).

In literature, values for CC and APL in real-life small-world networks have mostly been compared to theoretical values of CC and APL in Erdös-Rényi (ER) random networks [22] and to those of actually randomized networks [1] of similar numbers of vertices and edges. A stricter comparison can be made by randomly cross-wiring pairs of edges, a procedure known as the Maslov-Sneppen (MS) algorithm [23]. This algorithm randomly selects two pairs of connected vertices (v_1, v_2) and (v_3, v_4) such that all four vertices are different and then rewires them to (v_1, v_4) and (v_3, v_2). Repeating this operation

Table II. FROM EACH KANJI NETWORK, WE CONSTRUCTED 1,000 ERDÖS-RÉNYI RANDOM NETWORKS AND 1,000 MASLOV-SNEPPEN RANDOM NETWORKS BY 10^6 ITERATIONS OF THE RESPECTIVE ALGORITHM. BOTH ER-RANDOMIZATIONS AND MS-RANDOMIZATIONS WERE USED AS DIFFERENT NULL HYPOTHESES TO TEST THE STATISTICAL SIGNIFICANCE OF THE VALUES FOR CC AND APL FOUND IN THE KANJI NETWORKS.

		ER-randomized			MS-randomized		
	actual	mean (μ)	std.dev. (σ)	z-score	mean (μ)	std.dev. (σ)	z-score
Kyoulku CC	0.629	0.145	$2.178 \cdot 10^{-4}$	$2.218 \cdot 10^{3}$	0.330	$1.377 \cdot 10^{-3}$	$2.169 \cdot 10^{2}$
Kyoulku APL	1.962	1.855	$3.129 \cdot 10^{-14}$	$3.440 \cdot 10^{12}$	1.900	$7.620 \cdot 10^{-4}$	$8.196 \cdot 10^{1}$
Jouyou CC	0.604	0.155	$6.072 \cdot 10^{-5}$	$7.395 \cdot 10^{3}$	0.321	$5.198 \cdot 10^{-4}$	$5.446 \cdot 10^{2}$
Jouyou APL	1.888	1.845	$0.437 \cdot 10^{-14}$	$9.760 \cdot 10^{11}$	1.858	$2.724 \cdot 10^{-4}$	$1.089 \cdot 10^{2}$
JIS.X.0208 CC	0.582	0.166	$3.203 \cdot 10^{-5}$	$1.298 \cdot 10^{3}$	0.321	$1.200 \cdot 10^{-4}$	$2.174 \cdot 10^{3}$
JIS.X.0208 APL	1.843	1.834	0	∞	1.836	$5.165 \cdot 10^{-5}$	$1.360 \cdot 10^{2}$

for a large number of iterations effectively randomizes the network, but preserves the exact degree distribution.

For assessing the statistical significance of the actual CC and APL of the three Kanji networks, we generated 1,000 ER-randomized networks and 1,000 MS-randomized networks from each of the three Kanji networks by rewiring the original networks 10^6 iterations with the respective methods. The CC of all three Kanji networks is far larger than that of either an ER-random network or a MS-random network making them highly clustered (Table II). It should be noted that the significant difference between CC for MS-random networks and ER-random networks also suggests that the exact degree distribution might play a crucial role in facilitating the high clustering coefficient of these networks (Figure 2). The APL-values of the networks are relatively low, making the networks globally well-coupled despite their low edge density.

The KyouIku network has two vertices with degree zero (面, meaning 'surface' and 飛, meaning 'to fly'). As disconnected vertices lead to undefined path lengths, these two vertices were excluded from the experiments. The network's best-connected Kanji vertex is 速, meaning 'fast', which has 450 edges, connecting almost half the network. The characters in KyouIku contain a total of 219 different components of which 30 are used only once. The most common component is the enclosure component 口 which is shared by 181 characters in the set. The JouYou network, like the KyouIku network, also had two isolated vertices removed from calculations (again 面 and 飛). In this network, the Kanji vertex with the highest degree is 籍, meaning 'one's family register', which has 922 connections, connecting almost half the JouYou network. In this set of characters, there are 237 different components of which 13 are used for one character only. The enclosure component 口 is again the most common component, being present in 345 different characters. The JIS.X.0208-network is a connected network; there is a path from every vertex to every other vertex. Only one Kanji vertex in the network (飛, meaning 'to fly') has exactly degree one (being adjacent only to 飜, meaning 'to turn over'). In this network, the most edges on a single vertex is 3,001, for 檀 (meaning 'cedar'), which is thereby connected to almost half the network. Again, the enclosure 口 is the most occurring component, which is shared by 1,325 Kanji vertices in this network. The rarest components are 鬯 and 鼎 both appearing only once. Plotted vertex degree histograms show irregular distributions for all three networks, meaning that they do not qualify for the characteristic of being scale-free [24] (Figure 2).

IV. DISCUSSION AND FUTURE WORK

The clusters in the Japanese Kanji sets are an immediate result of the distribution of components among characters. Any single component completely interconnects all characters having that component, effectively creating an integrated module, often increasing the network's CC significantly. However, this has possibly not always been the case; the 2nd century A.D. dictionary *Shuōwén Jiězì* ("Explaining Simple and Analyzing Compound Characters"), describes 9,353 characters indexed by 540 radicals and possibly containing an even larger number of components. None of the far eastern logographic languages has nearly that many components nowadays, signifying the number must have dropped dramatically through time.

This considerable reduction in visual complexity might have been symbiotically accompanied by another interesting development. Research has shown that the pronunciation of a modern-day Kanji closely corresponds to the components it contains [25]–[27]. For instance, the single-component Kanji 中, meaning 'middle' is pronounced as *chuu*, but so are very different compound Kanji containing the same component: 忠 ('loyal'), 沖 ('shore'), 仲 ('relation'), 虫 ('insect'), 狆 ('Japanese Spaniel'). It might be somewhat problematic to understand how these five Kanji would actually share a quantity of meaning (in this case the concept of 'middle') but clearly, they do share their pronunciation.

Together, these observations may be interpreted as support for *Gelb's hypothesis*. In his influential study, Ignace Gelb hypothesized that characters in human written languages tend to evolve from picture-based logographic symbols to sound-based alphabetic symbols through time, with the "degree of abstractness serving as the main index of sophistication" [28], [29]. Interestingly enough, the statistical footprint of both Kanji characters and their components seems to provide further support for this theory. Although words in Japanese are distributed very much like words in non-logographic languages (Russian, Arabic, Spanish), following Zipf's law [30]–[33], the frequency distribution of individual Kanji characters more closely resembles those of alphabetic letters in Cyrillic, Arabic, and Roman scripts (Figure 3) but perhaps more surprisingly – so does the set of individual components. So, even if Kanji and Kanji components would be pure logographic symbols and word-like carriers of meaning, their statistical behavior tells a different story, classifying them as letters like б, л, ш, й, m, u, ñ, d, ð, ك, ب, ش. more than anything else.

The composition of Kanji characters is often primarily explained through its combined meanings of their components (e.g., in textbooks [34], [35]) but these explanations might be of mnemonical or folkloristic nature more than historically

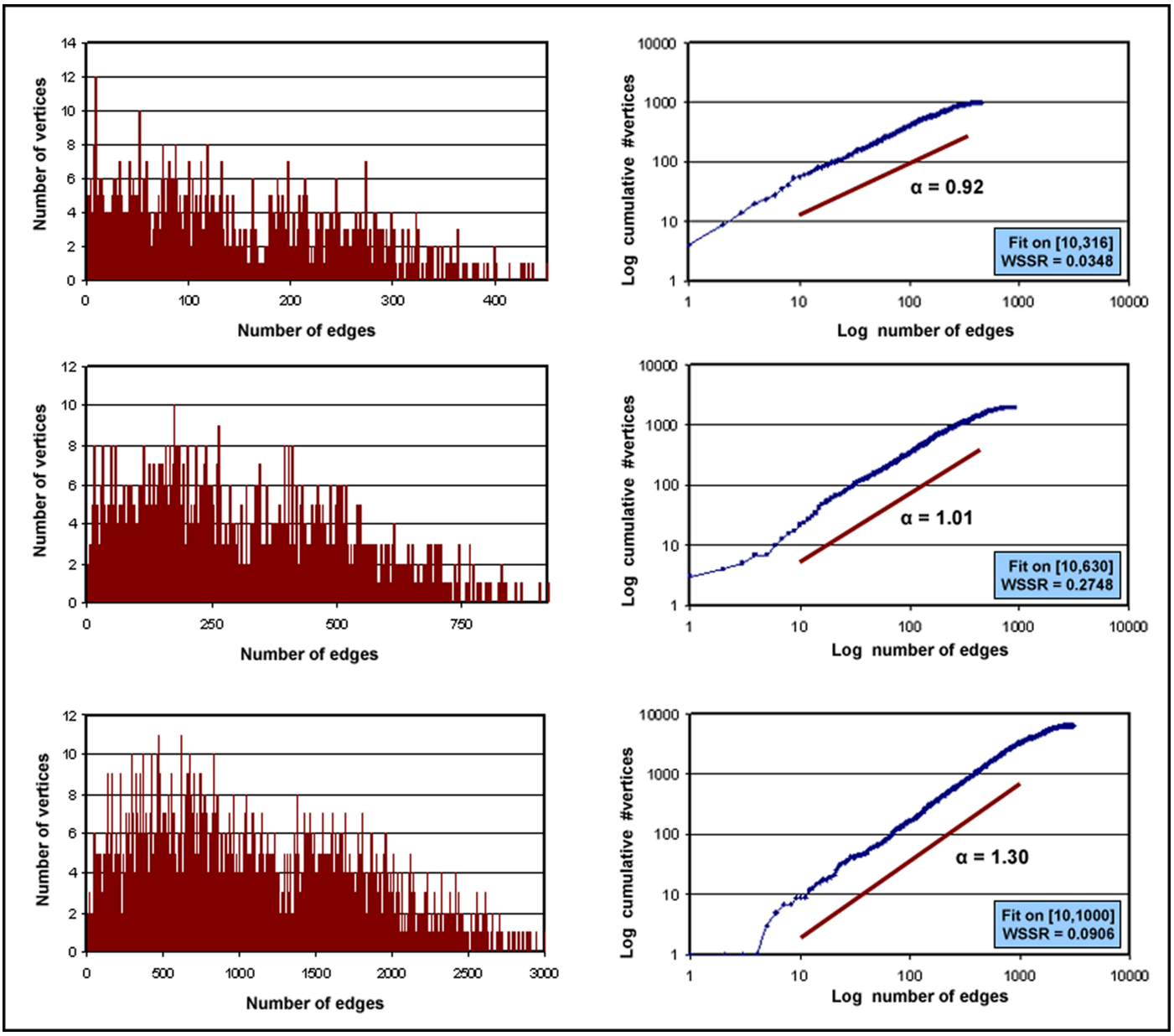

Figure 2. Degree distributions (left) of Kanji networks (top: KyouIku, middle: JouYou, bottom: JIS X.02.08) show apparent featureless patterns. However, log-log plots of the cumulative networks of the same plots (right-hand side) show largely linear tendencies on the central part. The red line corresponds to both the exponent and the interval used for the specific network, the text directly below is the value of its exponent. The values in the inset boxes are fitting intervals and fitting errors. All fits were done with Fityk on Linux by the Levenberg-Marquardt method from random initial conditions.

accurate accounts of Kanji construction. If the interpretation of our findings is correct, modern-day Kanji, at least to some extent, behave more like alphabetic letters than like logographic pictures. From this perspective, clustering (or small-worldness) in a network of Japanese (or Chinese) characters might just be a side effect of a much larger process: that of a language being in the midst of a phase transition from being picture-based to being alphabet-based.

ACKNOWLEDGMENTS

The starting point for this investigation was sparked by JWPCE, a freely available Japanese word processor for Windows OS written by Glenn Rosenthal (http://www.physics.ucla.edu/ grosenth/japanese.html). Sergei Sharoff from Leeds University maintains a broad multitude of freely available language corpora. We are grateful to both Glenn and Sergei for disclosing these resources. Our thanks go out to and to Charles Adamson for many helpful remarks and comments, to anonymous reviewers of both Entropy and NDPLS for their high-quality work and constructive comments. The support from Ramon Ferrer-i-Cancho from Llenguatges i Sistemes Informàtics (LSI) of Universitat Politècnica de Catalunya has been indispensable – thanks, Ramon. Finally, a recognition to Hans Dekkers, Guus Delen, Betty Bijl and Cristine Cabi from IvI/UvA who have freed some of my (Daan) time to do some research, which I have

used to produce this paper.

REFERENCES

[1] D. Watts and S. Strogatz, "Collective dynamics of small-world networks." Nature, vol. 393, no. 6684, 1998, pp. 440–442.

[2] T. Achacoso and W. Yamamoto, AY's Neuroanatomy of C. elegans for Computation. CRC Press, 1992.

[3] K. Stephan et al., "Computational analysis of functional connectivity between areas of primate cerebral cortex," Phil. Trans. R. Soc. B., vol. 355, no. 1393, 2000, pp. 111–126.

[4] O. Sporns and J. Zwi, "The small world of the cerebral cortex," Neuroinformatics, vol. 2, no. 2, 2004, pp. 145–162.

[5] L. Adamic, The Small World Web. Springer, 1999.

[6] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

[7] P. Gong and C. van Leeuwen, "Emergence of scale-free network with chaotic units," Phys. A, vol. 321, no. 3-4, 2003, pp. 679–688.

[8] ——, "Evolution to a small-world network with chaotic units," Europhys. Lett., vol. 67, no. 2, 2004, pp. 328–333.

[9] M. Rubinov, O. Sporns, C. van Leeuwen, and M. Breakspear, "Symbiotic relationship between brain structure and dynamics," BMC Neurosci., vol. 10, 2009, p. 55.

[10] D. Van den Berg and C. van Leeuwen, "Adaptive rewiring in chaotic networks renders small-world connectivity with consistent clusters," Europhys. Lett., vol. 65, no. 4, 2004, pp. 459–464.

[11] D. van den Berg, P. Gong, M. Breakspear, and C. van Leeuwen, "Fragmentation: loss of global coherence or breakdown of modularity in functional brain architecture?" Frontiers in systems neuroscience, vol. 6, 2012.

[12] R. Ferrer-i-Cancho and R. Sole, "The small world of human language," Proc. R. Soc. B., vol. 268, no. 1482, 2001, pp. 2261–2265.

[13] J. Li and J. Zhou, "Chinese character structure analysis based on complex networks," Phys A, vol. 380, 2007, pp. 629–638.

[14] Y. Shi, W. Liang, J. Liu, and C. K. Tse, "Structural equivalence between co-occurrences of characters and words in the chinese language," in International symposium on nonlinear theory and its applications, 2008, pp. 94–97.

[15] Z. Liu and M. Sun, "Chinese word co-occurrence network: its small-world effect and scale-free property," Journal of Chinese Information Processing, vol. 21, no. 6, 2007, pp. 52–58.

[16] S. Zhou, G. Hu, Z. Zhang, and J. Guan, "An empirical study of chinese language networks," Physica A: Statistical Mechanics and its Applications, vol. 387, no. 12, 2008, pp. 3039–3047.

[17] Y. Li, L. Wei, Y. Niu, and J. Yin, "Structural organization and scale-free properties in chinese phrase networks," Chinese Science Bulletin, vol. 50, no. 13, 2005, pp. 1305–1309.

[18] K. Yamamoto and Y. Yamazaki, "A network of two-Chinese-character compound words in the Japanese language," Physica A: Statistical Mechanics and its Applications, vol. 388, no. 12, 2009, pp. 2555–2560.

[19] ——, "Structure and modeling of the network of two-Chinese-character compound words in the Japanese language," Physica A: Statistical Mechanics and its Applications, vol. 412, 2014, pp. 84–91.

[20] Y.-M. Chou, S.-K. Hsieh, and C.-R. Huang, Lecture Notes in Computer Science, State-of-the-Art Survey. Springer-Verlag, 2007, pp. 133–145.

[21] "Kradfile," http://nihongo.monash.edu//kradinf.html, last access date: 31 October, 2017.

[22] A. Fronczak, P. Fronczak, and J. A. Hołyst, "Average path length in random networks," Phys Rev E, vol. 70, no. 5, 2004, pp. 1–4.

[23] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," Science, vol. 296, no. 5569, 2002, pp. 910–913.

[24] A. Barabasi and R. Albert, "Emergence of scaling in random networks," Science, vol. 286, no. 5439, 1999, pp. 509–512.

[25] H. Townsend, "Phonetic components in japanese characters," Master's thesis, San Diego Status University, 2011.

[26] E. Toyoda, A. Firdaus, and C. Kano, "Identifying useful phonetic components of kanji for learners of japanese," Japanse Language and Literature, 2013, pp. 235–272.

[27] K. Tamaoka, "Psycholonguistic nature of the japanese orthography," Studies in Language and Literature, vol. 11, no. 1, 1991, pp. 49–82.

[28] T. Miyamoto, "The evolution of writing systems: against the gelbian hypothesis," New Frontiers in Artificial Intelligence, 2007, pp. 345–356.

[29] I. Gelb, A study of writing: the foundations of grammatology. The University of Chicago, 1952.

[30] G. Zipf, Human Behavior and the Principle of Least Effort. Addison-Wesley Press, 1949.

[31] ——, The Psycho-Biology of Language. MIT Press, 1935.

[32] R. Ferrer-i-Cancho and R. Sole, "Least effort and the origins of scaling in human language," in Proc. of the National Academy of Sciences, vol. 100, no. 3, 2003, pp. 788–791.

[33] R. Ferrer-i Cancho, "Optimization models of natural communication," Journal of Quantitative Linguistics, 2017, pp. 1–31.

[34] J. Heisig, Remembering the Kanji, Volume 1: a complete course on how not to forget the meaning and writing of Japanese characters. University of Hawaii Press, 2007.

[35] Y. Watanabe, Learning Kanji Through Stories. Kurosio Publs., 2008.

 ISBN: 978-1-61208-603-3

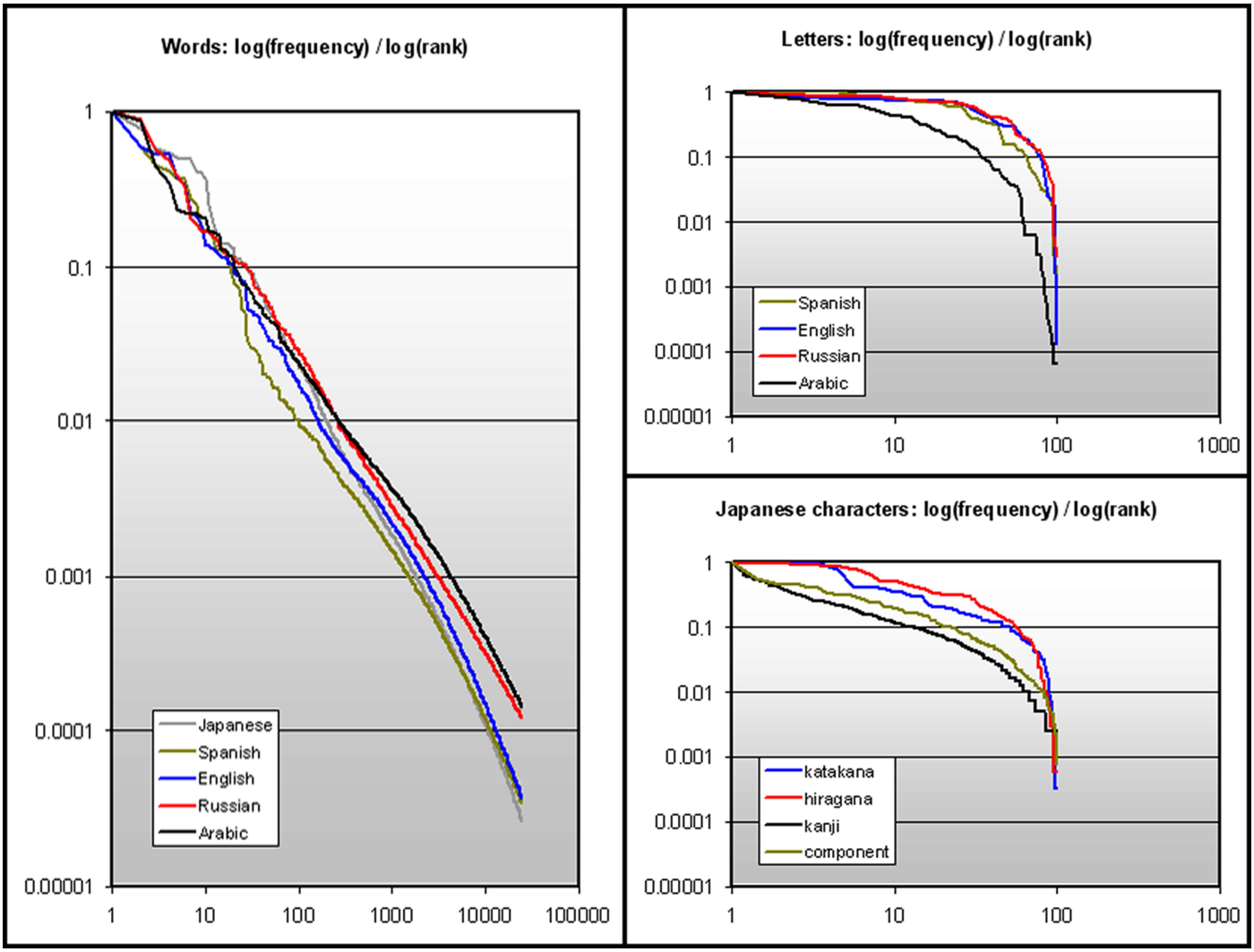

Figure 3. Written words in Japanese, as in most other languages, closely follow a power law distribution known as Zipf's law in linguistics (left). But even though single Kanji characters are often interpreted as carriers of meaning, their statistical behaviour more closely resembles that of letters in non-logographic writing systems (top right), and the same goes for its components.

 ISBN: 978-1-61208-603-3