An Examination of the Disparities Between Traditional and Simplified Chinese Using Complex Network Analysis of Word Co-Occurrence Networks

Gabriel
University of Paris, Paris, France

Abstract: At the moment, the majority of research comparing simplified and traditional Chinese exclusively looks at character or lexical distinctions, ignoring general disparities. This study suggests using complex network analysis of word co-occurrence networks, which have been effectively used in language analysis research and can address global characters and investigate the distinctions between simplified and traditional Chinese, to address this issue. Specifically, we used a few chosen news corpora to build a word co-occurrence network for both simplified and traditional Chinese. To obtain a comprehensive knowledge of these networks, advanced network analysis techniques were then carried out, such as motif analysis, kernel lexicon comparison, and network statistics analysis. Following that, the networks were contrasted using the acquired attributes. Three intriguing outcomes can be obtained by comparison: First, simplified Chinese and traditional Chinese co-occurrence networks are scale-free and small-world networks. Third, motif analysis reveals no difference between the simplified Chinese network and the corresponding traditional Chinese network, indicating semantic consistency between simplified and traditional Chinese. Nevertheless, given the same corpus size, the co-occurrence networks of traditional Chinese tend to have more nodes, which may be due to a large number of one-to-many character/word mappings from simplified Chinese to traditional Chinese; second, the traditional Chinese kernel lexicons have more entries than the simplified Chinese kernel lexicons because traditional Chinese retains more ancient Chinese words and uses fewer weak verbs.

1. Introduction

All Chinese is typically written in two ways: traditional Chinese, which is primarily used in Hong Kong, Macao, and Taiwan, and simplified Chinese, which is primarily used in Singapore and Mainland China. Despite being descended from traditional Chinese, simplified Chinese differs greatly from it on a number of levels, including character set, encoding technique, orthography, vocabulary, and semantics. These differences hinder communication between diverse Chinese-speaking regions. The independent growth of these two homologous systems over the last fifty years is the cause of this language phenomena, and they will continue to change in their different cultural contexts. However, the issue of converting simplified Chinese to traditional Chinese and comparing the differences between the two has drawn the attention of an increasing number of researchers in recent decades due to the growth in exchange activities between the four cross-strait regions [1–4]. To put it briefly, the comparison between There is significant reference value in both simplified and traditional Chinese for the study of language evolution. To date, character or lexical levels have been the main focus of studies analyzing the distinctions between these two versions of Chinese [1, 3, 5]. For instance, Li [7] conducted a thorough analysis of the reasons behind the differences in the form of simplified and traditional Chinese characters from the perspectives of politics, history, and culture as well as the principles of character selection; Liu [8] carried out a thorough analysis primarily from the perspective of eliminating the differences in form; Jiang [9] primarily compared and analyzed simplified and traditional Chinese vocabulary from two aspects: homographs with different meanings and different forms with synonymous meanings; Li and Qiu [10] talked about the types, causes, and processing techniques of dictionaries across the Taiwan Strait.

Complex networks-based approaches, on the other hand, are a valuable methodology for linguistic research because they reveal the global features of language, such as lexical [11–13], word co-occurrence [14–18], syntax [19–21], and semantic [22–24], which have been successfully applied to analyze languages at different levels. The reason for this is that language is a typical hierarchical system with a very complex network structure, and the benefit of using complex network analysis techniques is that they can disclose the rules of language in its entirety. Therefore, in order to investigate the distinctions between simplified and traditional Chinese character systems from a comprehensive standpoint, we employ sophisticated network analysis techniques in this research. In particular, this paper suggested building simplified Chinese and traditional Chinese word co-occurrence networks with varying numbers of nodes and corpus sizes in accordance with the word co-occurrence network construction method. Then, it conducted corresponding research on the intricate features of these networks. We investigated the distinctions between the two languages using the simplified and traditional Chinese core dictionary that was obtained. Furthermore, this work suggested analyzing the semantic distinctions between simplified and traditional languages using primitives that describe language semantics.

1,947 characters, twenty-four percent of similar glyphs (1,170 characters), and thirty-five percent of distinct glyphs (1,669 characters). Traditional and simplified Chinese are descended from the same ancient Chinese and share an ancestor. Thus, from the standpoint of the language as a whole, which examines the distinctions between the two written forms of Chinese development status and law, the disparities between simplified and traditional Chinese must be contrasted and examined methodically and thoroughly. However, additional language levels (such semantics and syntax) have not been included in the current comparative study of simplified and traditional Chinese characters, which has only made notable progress at the character and word levels.

At all levels (phonetics, morphology, syntax, and semantics), language has a very complicated network structure, as is typical of hierarchical systems [25]. Many studies on the intricate features of language networks at various levels have been conducted recently. These include syntactic networks, also known as semantic networks, lexical or vocabulary networks, and word or character co-occurrence networks. These investigations are crucial for determining and comprehending language's topological structure. Among them, the Chinese network research investigations mostly consist of the following: In the morphology or vocabulary network, Li et al. [13] used Chinese characters as nodes based on the idea that two Chinese characters can form words. They then created a Chinese phrase network and examined its dynamic characteristics. In the syntactic network, Liu [20] connected words with syntactic relations using a syntactic labeling tree bank, ultimately establishing the Chinese syntactic dependency network and investigating its complex network characteristics. In the semantic network (current research studies on Chinese semantic networks are still relatively small), Liu et al. [24] built a small semantic network to investigate the complex features of the Chinese semantic network and introduces the relevant work. A brief overview of some fundamental ideas in complex network analysis is provided in Section 3. Then, in Section 4, we built networks using various text scales and conducted corresponding research on the properties of complex networks, such as kernel lexicon, clustering coefficient, cumulative degree distribution, and motif analysis. The paper is finally concluded in Section 5.

2. Associated Work

Character shapes or words are currently the primary focus of comparison and analysis of the differences between simplified and traditional Chinese. The primary cause of readers' difficulties while reading unfamiliar written materials in conventional or simplified characters is the disparity in glyphs. According to studies, there are 4,786 characters in total that can be compared between the traditional and simplified Chinese character lists [6]. The corpus used to create an English word co-occurrence network revealed that the English language network has a small world and scale-free properties, with 41% of the simplified and traditional characters used in Taiwan and mainland China sharing the same glyph. Liu and Sun [15] created a simplified Chinese word co-occurrence network using the same building technique. The experiment demonstrated that the English word co-occurrence network and the simplified Chinese word co-occurrence network have sophisticated network properties. Based on various themes of Chinese and English (prose, novels, popular science articles, and news reports) corpora, other publications [12, 26, 27] employed various construction methodologies to create a Chinese word, word co-occurrence network, and English word co-occurrence network.

3. Bases

This section introduces some fundamental ideas. The fundamental definitions of the complex network are explained in Section 3.1. Small-world networks and scale-free networks are then covered in Section 3.2. Lastly, a succinct overview of motif analysis is provided in Section 3.3.

In other cases, a network is said to as scale-free if it falls between 2 and 3 and follows the power law effectively [29]. 3.3 Analysis of Motifs. The academic field of biology was the first to adopt motifs, which are subgraphs made up of a few edges and vertices [30]. A motif is a subnetwork of a complex network that has a limited number of nodes and edges. In order to investigate the differences in semantic properties between text produced by an N-gram language model and text produced naturally, Biemann et al. [31] initially used motif analysis in linguistic networks and semantic attributes. Additionally, motif analysis uses a network's intermediate level, which means counting the motif created by n nodes in order to compare networks. Typically, n in undirected co-occurrence networks is at least 3. When determining the clustering coefficient, a 3-node motif is entirely triple-contained. In order to compare the semantic differences of co-occurrence networks, we employ 4-node motif analysis. Figure 1 displays all six types of undirected 4-node motifs.

4. Comparing Experiments

Based on techniques from complex network science, this section discusses the experimental comparisons between simplified and traditional Chinese. The dataset used and the creation of the word co-comparisons on scale-free and small-world kernel lexicons and motif analysis, respectively, are described in Section 4.1.

4.1. Building a dataset and network. The experimental materials for this experiment are texts from Chinese GigaWord Third Edition (LDC2007T38) https://catalog.ldc.upenn.edu/LDC2007T38, of which the traditional Chinese texts are from "Central News Agency" (henceforth referred to as CNA) and the simplified Chinese texts are from "Xinhua News Agency" (henceforth referred to as XIN).

Word co-occurrence networks are constructed using the approach suggested by [32] based on the datasets. In specifics, words in the texts are considered nodes in the networks, and any two nodes are linked if the distance between the corresponding words is less than two.

Once the networks are built, their statistical characteristics are examined and contrasted. Please take note that the influence of the text scales is avoided by comparing only networks constructed from similar text scales. Three text scales are employed in this experiment, and Table 1 displays the statistics for each network. We created three sets of experiments for the co-occurrence network of traditional and simplified Chinese terms on the same corpus scale. From the original 7 million words to 10 million words and finally 15 million words, the corpus scales utilized in these three sets grew.

Table 1: Statistics of the built word co-occurrence networks. XIN_1, XIN_2, and XIN_3 are from different parts of the XIN dataset; CNA_1, CNA_2, and CNA_3 are from different parts of the CNA dataset.

	Theme (name)	Text scales (# of words) (M)	Sources	# of nodes
Group 1	XIN_1	55.9	XIN (Jan., 2006–May., 2006)	$1.06 * 10^5$
	CNA_1	55.3	CNA (Jan., 2006–Mar., 2006)	$1.14 * 10^5$
Group 2	XIN_2	79.8	XIN (Jan., 2006–Jun., 2006)	$1.26 * 10^5$
	CNA_2	79	CNA (Jan., 2006–Apr., 2006)	$1.38 * 10^5$
Group 3	XIN_3	115	XIN (Jan., 2006–Sep., 2006)	$1.52 * 10^5$
	CNA_3	114	CNA (Jan., 2006–May., 2006)	$1.69 * 10^5$

4.2. Scare-free and small-world. We compute the statistical features of the constructed networks using Pajek2, a sophisticated network analysis tool. The results are displayed in Table 2.

Table 2: Properties of the built networks. N: number of nodes; E: number of edges; k: average degree of nodes; C: clustering coefficient; d: average path length among reachable pairs of nodes; C_{ER}: clustering coefficient of an ER network with same numbers of nodes and edges; d_{ER}: average path length among reachable pairs of nodes in an ER network with same numbers of nodes and edges; and c: power-law exponent in equation (5).

Metric	Dataset theme					
	XIN_1	CNA_1	XIN_2	CNA_2	XIN_3	CNA_3
N	$1.06 * 10^5$	$1.14 * 10^5$	$1.26 * 10^5$	$1.38 * 10^5$	$1.52 * 10^5$	$1.69 * 10^5$
E	$0.27 * 10^7$	$0.32 * 10^7$	$0.35 * 10^7$	$0.41 * 10^7$	$0.45 * 10^7$	$0.53 * 10^7$
k	50.01	55.08	54.45	59.39		
C	0.68	0.68	0.69	0.70	0.72	0.73
d	2.69	2.72	2.69	2.73	2.70	2.74
C_{ER}	$4.69 * 10^{-4}$	$4.80 * 10^{-4}$	$4.28 * 10^{-4}$	$4.30 * 10^{-4}$	$3.90 * 10^{-4}$	$3.70 * 10^{-4}$
d_{ER}	3.24	3.21	3.26	3.20	3.25	3.20
c	2.17	2.18	2.16	2.17	2.15	2.15

Table 2 shows that every network satisfies d ≈ dER and C≫CER, indicating that every network is a small-world network. The average degrees of conventional networks, on the other hand, are roughly five points greater than those of the corresponding simplified networks. This could be because traditional Chinese and simplified Chinese have many-to-one mappings, meaning that multiple words in traditional Chinese have the same forms.

Additionally, we present each network's cumulative degree distributions and fitting curves in Figure 2. It is evident that the power law fits both conventional and simplified Chinese networks effectively. Furthermore, every network's power-law exponent falls between 2 and 3, suggesting that every network is scale-free.

We further analyze the part-of-speech tags and word lengths in the kernel lexicons to determine the potential causes. Tables 4 and 5 present the findings, respectively.

Table 4 shows that a significant percentage of entity words (nouns and verbs) with about equal ordering are present in both versions of Chinese. In simplified Chinese, verb weakening is a significant development process, as the proportion of verbs in traditional Chinese is typically higher than in simplified Chinese.
Table 5 revealed that compared to the simplified Chinese corpora, the kernel lexicons taken from the traditional Chinese corpora had a higher number of 1-character words. This suggests that, in contrast to simplified Chinese, traditional Chinese retains certain aspects of classical Chinese.

4.3. Verification of the motif. We conducted the motif analysis on each of the networks built in Section 4.1 in accordance with [31].Table 6 displays the findings. With the exception of the fact that traditional Chinese networks typically have more motifs than simplified Chinese networks because of their greater number of nodes and edges, there is no difference between simplified Chinese networks and their corresponding traditional Chinese networks. This demonstrates the semantic consistency between simplified and traditional Chinese.

4.4. Comparative Example. As indicated in Table 4, we discovered that the parts of speech of these various terms are primarily represented in nouns, verbs, time words, gerunds, adverbs, numerals, and ground nouns. Nouns, verbs, gerunds, and adverbs are among those that change depending on the corpus.

In conclusion, there is some flexibility in the core dictionaries of the traditional and simplified Chinese character systems. However, regional usage patterns, the environment, politics, and the creation of new terms have all contributed to certain disparities in the language development process. Furthermore, some aspects of classical Chinese are still present in the written language of the traditional Chinese character system.

5. Conclusion

In order to investigate the distinctions between simplified and traditional Chinese, we suggested a complicated network in this paper. To the best of our knowledge, this is the first study to compare the differences between simplified and traditional Chinese using sophisticated network-based techniques. Three intriguing outcomes are obtained from the comparisons. First off, co-occurrence networks are small-world and scale-free for both simplified and traditional Chinese.

However, the co-occurrence networks for traditional Chinese tend to contain more nodes given the same corpus scale. This could be because there are many one-to-many character/word translations from simplified Chinese to traditional Chinese. Second, compared to simplified Chinese, traditional Chinese kernel lexicons contain more entries. This could be because traditional Chinese retains more old Chinese terms while using fewer weak verbs. Thirdly, the motif analysis reveals no distinctions between the corresponding traditional Chinese networks and the simplified Chinese networks. To put it another way, there is semantic consistency between simplified and traditional Chinese.

References

[1] L. Wang, X. Wang, and J. Wu, “The correspondence sim- plified characters and traditional characters and the mutual conversion,” *Journal of Chinese Information Processing*, vol. 4, 2013.

[2] P. Zhenjun and Y. Tianfang, “Chinese characters conversion system based on lookup table and statistical methods,” *Computer Engineering and Applications*, vol. 51, no. 4, p. 24, 2015.

[3] H. Dai, “Linguistic analysis of the intelligent conversion system of simplified and traditional Chinese characters text,” *Liaoning Normal University (Social Science Edition)*, vol. 39, no. 2, pp. 115–120, 2016.

[4] L. Wang, “Review of and reflections on the hot topics in the application of contemporary Chinese charactersl Chinese characters text,” *Applied Linguistics*, no. 2, 2020.

[5] M.-H. Li, S.-H. Wu, Yi-C. Zeng, P.-C. Yang, and T. Ku, “Chinese characters conversion system based on lookup table and language model,” *Computational Linguistics and Chinese Language Processing*, vol. 15, no. 1, pp. 19–36, 2010.

[6] J. Fei, “Comparative analysis of current Chinese characters across the Taiwan straits,” *Language Application*, vol. 1993, no. 1, pp. 37–48, 1993.

[7] L. Li, “An analysis of the reasons for the differences in the forms of Chinese characters on both sides of the Taiwan straits,”

Journal of Guangxi University, vol. 20, no. 1, pp. 98–101, 1998.

[8] X. Liu, "*Study on the unification of Chinese characters across the Taiwan straits*," M.S. thesis, Northwest University, Kirkland, WA, USA, 2007.

[9] Y. Jiang, "Differences in Chinese vocabulary between the two sides of the taiwan straits and their reasons," *Jimei University Journal*, vol. 9, no. 3, pp. 31–37, 2006.

[10] X. Li and Z. Qiu, "Definement and treatment of difference words in cross-strait dictionaries-new problems in cross-strait co-edited Chinese dictionaries," *Language Application*, vol. 2012, no. 4, pp. 74–81, 2012.

[11] A. E. Motter, A. P. S. De Moura, Y.-C. Lai, and P. Dasgupta, "Topology of the conceptual network of language," *Physical Review E*, vol. 65, no. 6, Article ID 065102, 2002.

[12] Y. Li, L. Wei, W. Li, Y. Niu, and S. Luo, "Small-world patterns in Chinese phrase networks," *Chinese Science Bulletin*, vol. 50, no. 3, pp. 287–289, 2005.

[13] J. Li, J. Zhou, X. Luo, and Z. Yang, "Chinese lexical networks: the structure, function and formation," *Physica A: Statistical Mechanics and Its Applications*, vol. 391, no. 21, pp. 5254– 5263, 2012.

[14] R. F. I. Cancho and R. V. Sole´, "The small world of human language," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 268, no. 1482, pp. 2261–2265, 2001.

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

[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, pp. 3039–3047, 2008.

[17] W. Liang, Y. Shi, C. K. Tse, J. Liu, Y. Wang, and X. Cui, "Comparison of co-occurrence networks of the Chinese and English languages," *Physica A: Statistical Mechanics and Its Applications*, vol. 388, no. 23, pp. 4901–4909, 2009.

[18] H. Liu and W. Li, "Language clusters based on linguistic complex networks," *Chinese Science Bulletin*, vol. 55, no. 30,
pp. 3458–3465, 2010.

[19] R. F. I. Cancho, R. V. Sole´, and R. Ko¨hler, "Patterns in syntactic dependency networks," *Physical Review E*, vol. 69, no. 5, Article ID 051915, 2004.

[20] H. Liu, "The complexity of Chinese syntactic dependency networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 387, no. 12, pp. 3048–3058, 2008.

[21] Z.-Y. Liu, Y.-b. Zheng, and M.-S. Sun, "Complex network properties of Chinese syntactic dependency network," *Complex Systems and Complexity Science*, vol. 2, 2008.

[22] M. Steyvers and J. B. Tenenbaum, "The large-scale structure of semantic networks: statistical analyses and a model of semantic growth," *Cognitive Science*, vol. 29, no. 1, pp. 41–78, 2005.

[23] X. F. Wang, Li Xiang, and G. R. Chen, *Theory of Complex Networks and its Application*, Tsinghua University, Beijing, China, 2006.

[24] H. Liu, "Statistical properties of Chinese semantic networks,"
Science Bulletin, vol. 54, no. 16, pp. 2781–2785, 2009.

[25] R. V. Sole´, B. Corominas-Murtra, S. Valverde, and L. Steels, "Language networks: their structure, function, and evolution," *Complexity*, vol. 15, no. 6, pp. 20–26, 2010.

[26] M. Sigman and G. A. Cecchi, "Global organization of the wordnet lexicon," *Proceedings of the National Academy of Sciences*, vol. 99, no. 3, pp. 1742–1747, 2002.

[27] 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, pp. 1305–1309, 2005.

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

[29] A.-L. Baraba´si and R. Albert, "Emergence of scaling in ran- dom networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[30] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature Genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[31] C. Biemann, S. Roos, and K. Weihe, "Quantifying semantics using complex network analysis," *Proceedings of Coling 2012*, pp. 263–278, 2012.

[32] R. F. I. Cancho and R. V. Sole´, "Two regimes in the frequency of words and the origins of complex lexicons: zipf's law revisited," *Journal of Quantitative Linguistics*, vol. 8, no. 3,
pp. 165–173, 2001.

[33] Z. M. Griffin and K. Bock, "Constraint, word frequency, and the relationship between lexical processing levels in spoken word production," *Journal of Memory and Language*, vol. 38, no. 3, pp. 313–338, 1998.

[34] S. N. Dorogovtsev and J. F. F. Mendes, "Language as an evolving word web," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 268, no. 1485, pp. 2603–2606, 2001.