Exploring Characteristics of Word Co-occurrence Network in Translated Chinese

Jianyu Zheng*, Kun Ma§†, Xuemei Tang§†, Shichen Liang§†
*Advanced Innovation Center for Future Education, Beijing Normal University, Beijing China
zheng_jianyu@126.com
§Institute of Chinese Information Processing, Beijing Normal University, Beijing China
†UltraPower-BNU Joint Laboratory for Artificial Intelligence, Beijing China
{201821090021, tangxuemei, shichen}@bnu.edu.cn

***Abstract*—The translation activity involves both the source language and the target language. Compared to the standard texts in the two language, translated texts show unique language characteristics. In order to explore them from the perspective of integrality and complexity, we introduce complex network method into the study on translated Chinese. Firstly, selected the experimental texts from The ZJU Corpus of Translational Chinese(ZCTC) and its corresponding six sub-corpora, such as Press reportage and Popular lore. And then removed the punctuation and did word segmentation. Secondly, constructed a word co-occurrence network of translated Chinese. After analyzing and counting the parameters, such as their shortest path lengths, degree distributions and clustering coefficients in these networks, we verify that the word co-occurrence network of translated Chinese has small world effect and scale-free property. Finally, by constructing co-occurrence networks of standard Chinese and calculating their network parameters, we compare and verify the differences between translated Chinese and standard Chinese: "simplification" and the more usage of common words. Our work expands the application of complex network in translation studies, and provides a feasible approach for studying translated Chinese based on complex networks.**

Keywords-translated Chinese; complex network; word co-occurrence; small world effect; scale-free property

I. INTRODUCTION

Translation gets involved in the source language and target language. So the translated text is closely related to these two languages. Since the 1980s, linguists started to pay increasing attention to the linguistic characteristics of translated texts. The researchers[1][2] in early stage described characteristics of translated language by vocabulary and syntax with the help of corpus. Based on that, they revealed the general characteristics of translation as follows:

- "simplification": "translators unconsciously simplify language, information or the both at the same time."[1]. In other words, translated texts are simpler than standard texts in original language, such as the smaller range in word usage and more common words;
- "explicitation": the information implied in the source language or need to be deduced from the context will be expressed directly in the translated texts[3]. This characteristic is mainly reflected in sentence extension, modifier increase and pronoun explicitation;
- "normalization": typical and normative ways in the target language will be used to produce translated texts. It is mainly reflected by the reduction of special words and sentences, the stronger logic.

However, language is a special symbol system. The linguistic units in different levels do not exist independently. They depend on and restrain each other so that the language system has stringency and complexity. If using the traditional statistical approach, the association of different linguistic units will be cut apart to some extent. So it will be impossible to consider characteristics of translated language from the "systematic perspective".

Facing to the complexity and systemization of language, it is more reasonable to describe it through a perspective of network. The network with self-organization, self-similarity, small world effect , scale-free property is called complex network, which studies the complex system from the global perspective. Regardless of how complex and large the network is, it always adopts two basic elements—nodes and edges for study. The complex network has already become a powerful tool in the study of the complex system. At present, the study and application of the complex network are not only involved into mathematics, computer science, physics, biology and engineering technology, but also widely used in society, politics, economics, management and language. Therefore, we attempt to use the complex network method to analyze translated Chinese text. The main contributions of this paper are stated as follows:

- to study translated Chinese from the perspective of the complex network and reveal small world effect and scale-free property;
- to compare the measured results of translated Chinese network with the standard Chinese's with the similar stylistics, and reveal speciality of the translated Chinese network.

II. RELATED WORKS

The translated Chinese is a special kind of modern Chinese taking English as the source language and Chinese as the target language. Even if translated Chinese uses Chinese characters as the carrier, it obviously shows some characteristics different from the standard Chinese. Before the large-scale corpus appears, most of researches on translated Chinese belong to impressive type and comment

type, while the corpus method opens a new world for studying translated language. Yifan Zhu et al.[4], once used Chinese translated news corpus to generate subject-word lists and analyzed the quantified characteristics between translated Chinese and standard Chinese. Xiao and Dai(2014))[5] utilized ZCTC and LCMC corpus, and made a comparison on sentence length of the two from different texts. They found out that the largest difference in average sentence length between the two relied in academic texts. While in the novel texts, their gap was closest. By comparing ZCTC with LCMC corpus, Dai Guangrong(2013)[6] studied the matches of 10 high frequency nouns, and found that matching range of high frequency nouns in translated Chinese was wider than the standard Chinese's. The vocabulary change mode could be more diverse.

With a complicated network structure, language network system shows extreme complexity in terms of words, syntax, semantics. Liu Haitao[7] indicated that the complex network analysis method could reveal the overall characteristics of language as a relational system in the large-scale real corpus. Based on the English-speaking country corpus(BNC) with 10^7 word frequency, Cancho and Sole[8] constructed the co-occurrence network of English words and found out the small world effect and scale-free property. Liu Zhiyuan and Sun Maosong[9][10] constructed co-occurrence network of Chinese words and Chinese dependence syntax network to inspect characteristics from the perspective of the complex network. The experimental results indicated that these complex networks all still hold the similar characteristics like above.

III. THE NETWORK MODEL OF TRANSLATED CHINESE

In this paper, we take four steps to explore the characteristics of translated Chinese based on complex network.

1) Acquire translated Chinese corpus and preprocessing. According to actual needs, we acquire translated Chinese corpora. During preprocessing, it is necessary to remove the punctuation and segment words of the corpus.
2) Build word co-occurrence networks. As to word co-occurrence, it means two words within the distance *n* in a sentence. These two words have a co-word relationship, which is the basis of constructing word co-occurrence network.
3) Calculate network parameters & analyze characteristics. According to the indicators commonly used in the study of complex networks, we calculate the corresponding parameters of the word co-occurrence network. Then we analyze the characteristics of the network according to those results.
4) Compare with other networks. By Comparing with the language networks with the similar stylistics, we study the characteristics of translated Chinese further.

A. Construct word co-occurrence network of translated Chinese

The word co-occurrence network is a kind of network which is used to describe language through co-occurrence relation between words. A word co-occurrence network can be abstracted into an undirected graph *G*, where nodes delegate the corresponding words, and edges delegate the co-occurrence relationship between words. At present, the mainstream methods for constructing co-occurrence networks are n-order Markov co-occurrence model and similarity-based co-occurrence model. In this paper, we adopted n-order Markov co-occurrence model. According to the definition, if there is a co-occurrence relation between the two words within the distance *n*, an edge need to be added between the two nodes. By processing all the sentences in the corpus according to this method, a word co-occurrence network can be constructed. This method can not only fully reflect the contextual information between words, but also control the complexity of the model better[9]. According to Liu's study on word co-occurrence network[9], it is ideal when the distance is 2. The specific rule is: represent the words in the translated Chinese corpus as nodes in the co-occurrence network, and then create an edge between the two nodes in the same sentence.

We choose a sentence from a translated Chinese corpus ZCTC randomly[11]and take it as an example, "他一面开车，一面思忖着在教堂和牧师寓所听到的那些话，更不用说教堂外发生的事了。" ("While driving, he was thinking about the words heard in the church and the pastor's apartment, not to mention the words happened outside the church.").After removing punctuation and segmenting the sentence, the word co-occurrence network is shown as Fig 1.

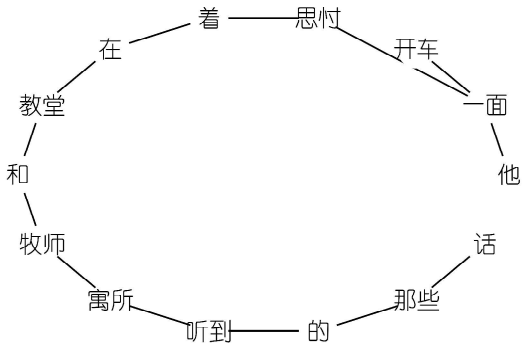

Figure 1. a co-occurrence network of translated Chinese

B. parameters and characteristics

The following parameters are usually used and analyzed when studying a complex network:

- **the number of nodes**: An actual network can be abstracted into a graph consisting of node sets and edge sets. Where, the number of nodes refers to how many nodes in this network, denoted by *N*.
- **the number of edges**: the sum of edges in a network, denoted by *E*.
- **degree**: the number of edges connecting the node *i* in an undirected network, denoted by k_i.
- **average degree**: the average degrees of all nodes in a network, denoted by $<k>$

$$< k >= \frac{\sum_i k_i}{N} \quad (1)$$

- **degree distribution**: select a node randomly from the network, the probability with a degree k among all the nodes, denoted by p(k)

$$p(k) = \sum_{i=k} Pr(i) \quad (2)$$

- **cumulative degree distribution**: the probability of nodes' degrees no less than *k* in a network

$$P(k) = \sum_{i=k}^{\infty} Pr(i) \quad (3)$$

- **power law index**: the index corresponding to the the power law distribution function, if the degree distribution of a network obeys that distribution, demoted by γ. Where,

$$Pr(k) \propto k^{-\gamma} \quad (4)$$

- **shortest path length**: the length of a path with the minimum edges connecting node *i* and node *j* in the network, denoted by d_{ij}
- **average shortest path length**: the average shortest path length of all pairs of nodes in the network, denoted by *L*

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \quad (5)$$

- **diameter**: the maximum of the shortest path lengths among all node pairs in the network, denoted by *D*

$$D = \max_{i,j} d_{ij} \quad (6)$$

- **reference coefficient of average shortest path length**: the average shortest path length in a random network with the same number of nodes and edges, denoted by *Lr*
- **clustering coefficient**: E_i delegates the actual number of edges of node *i* when it is connected by *k* nodes. Then the ratio between the maximum number of edges k(k - 1) and *Ei* is the clustering coefficient, denoted by *Ci*

$$C_i = \frac{2E_i}{k(k-1)} \quad (7)$$

- **average clustering coefficient**: the average clustering coefficients of all nodes in the network, denoted by *C*

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i \quad (8)$$

- **reference coefficient of average clustering coefficient**: the average clustering coefficient in a random network with the same number of nodes and edges, denoted by *Cr*

As to a complex network, it usually holds the following two characteristics:

- **Small world effect**

The average shortest path length and clustering coefficient are important indicators for measuring whether the network has small world effect. The small world effect means that although the network is very large, there is a short path between any two nodes in the network, and the clustering coefficient is much larger than that of a random network, that is $L \approx L_r, C >> C_r$

- **Scale-free property**

If the degree distribution of a network follows power-law distribution, it has the scale-free property.

IV. EXPERIMENT AND ANALYSIS

A. Corpus collection and preprocessing

Our experimental corpus is mainly from The ZJU Corpus of Translational Chinese(ZCTC). ZCTC was created, taking The Lancaster Corpus of Mandarin Chinese(LCMC) as a reference. The range of material involves 15 different types of written texts, such as Press reportage, Religious writing, Popular lore and General fiction. A total of 500 samples with 2000 words each was selected. The corpus size is one million words. The ZCTC corpus is one of the earliest translated Chinese corpus with high quality. In addition, ZCTC takes the balance of category into consideration well, so it can be used to investigate the characteristics of the word co-occurrence network of translated Chinese. About the detailed introduction to the corpus, please refer to the related works of Richard Xiao[11] and Guangrong Dai[6].

In addition, we also explore complex network characteristics of various translated Chinese texts. Because the number of texts in each category is different in ZCTC, we rearranged the 15 types of texts in a descending order according to the number of texts. And then selected out 6 types of texts with the maximum numbers. By doing that, we had sufficient corpora to carry out experiments and ensured the rationality of the experimental results. The six types are: Press reportage, Skill/trade/hobby, Popular lore, Biography and essay, Miscellaneous, and Science-academic prose. Then we randomly selected 30 texts from each type to carry out experiments later.

When preprocessing the corpus, we mainly removed the punctuation in the corpus and separated the sentences. Then with the Language Technology Platform(LTP),we segmented the sentences into words.

B. Characteristics of word co-occurrence networks of translated Chinese

1) Small world effect

In order to study the characteristics of word co-occurrence network of translated Chinese from various perspectives, we collected the whole corpus ZCTC, and the sub-corpora according to the six categories mentioned in section 4.1. As to the 7 corpora, we calculated the corresponding parameters according to section 3.2. The results are shown as Table I.

From Table I, all the 7 word co-occurrence networks have shorter average shortest path L, and $L \approx L_r$. In all networks, the average shortest path of Science-academic prose is the smallest, while Popular lore's is the largest. It indicates that the word usage is more compact in Science-academic prose, but Popular lore's is more sparse; In addition, as to the average clustering coefficient, there exists $C >> C_r$ in all the seven networks, where the clustering coefficient of Miscellaneous-reports and the official document is the largest, which indicates that the inter-connectivity of words in this category is the closest..

It can be seen that there is a significant small world effect both in the whole corpus and the six sub-corpora. Besides, the word usage of translated Chinese is more compact, and the distance between any two words in this network is no more than 11. It means that the word linkage is more closer, and the word grouping is more obvious. The above analysis indicates that although translated

Chinese might be different form standard Chinese in some aspects, it still takes Chinese as language carrier. When using translated Chinese to bear and convey information, it still needs to be complied with expression forms and habits of this language. Therefore, as to small world effect commonly owned by all human languages, translated Chinese also hold this characteristic.

TABLE I. STATISTICAL PARAMETERS OF WORD CO-OCCURRENCE NETWORK OF TRANSLATED CHINESE

Type	Length	N	E	D	<k>	L	L_r	C	C_r
A	60k	8820	30359	10	6.8841	3.1789	4.9264	0.1793	8.7717×10^{-4}
E	60k	7673	28679	9	7.4753	3.1529	4.6751	0.1876	8.0688×10^{-4}
F	60k	9027	29782	11	6.5984	3.2019	5.0425	0.1780	7.2260×10^{-4}
G	60k	8942	30109	10	6.7343	3.1624	4.9952	0.1910	8.5438×10^{-4}
H	60k	5724	22507	8	7.8641	3.0762	4.4305	0.2014	1.4227×10^{-3}
J	60k	6870	27289	10	7.9444	3.0751	4.4979	0.1952	1.1433×10^{-3}
Total	1M	47296	318782	9	13.4803	2.8911	4.4345	0.3372	2.7233×10^{-4}

Note: Considered that the space, we represent every sub-corpus according to their category numbers in corpus ZCTC, they are:
2)A: Press reportage 2)E: Skill/trade/hobby 3)F: Popular lore 4)G: Biography and essay
5)H: Miscellaneous-reports and official document 6)J: Science-academic prose

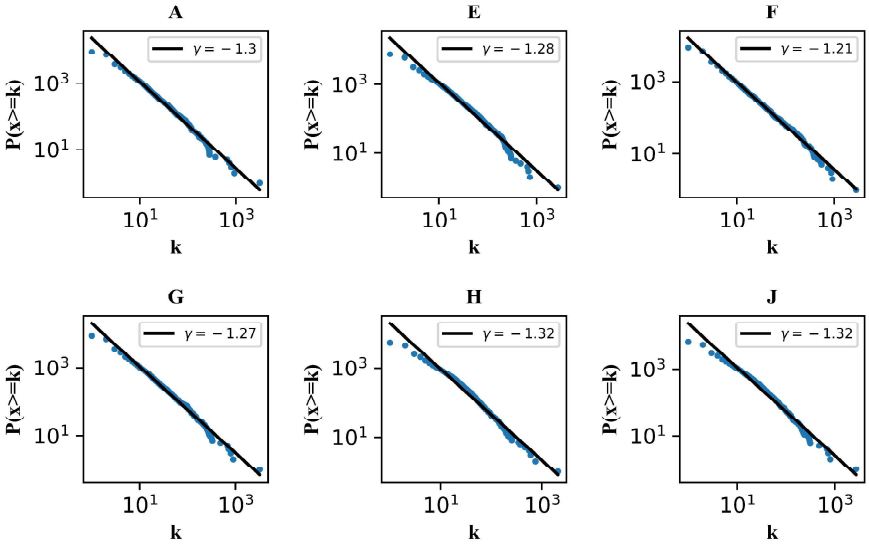

Figure 2. Cumulative degree distribution of word co-occurrence network about 6 various sub-corpus

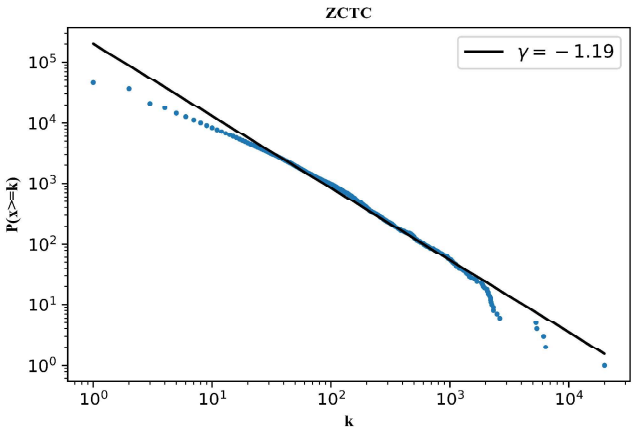

Figure 3. Cumulative degree distribution of word co-occurrence network about ZCTC

TABLE II. STATISTICAL PARAMETERS OF WORD CO-OCCURRENCE NETWORK OF TRANSLATED CHINESE IN THE ASPECT OF SCALE-FREE PROPERTY

Type	γ	R2
A	-1.3002	0.9879
E	-1.2797	0.9794
F	-1.2100	0.9927
G	-1.2701	0.9893
H	-1.3213	0.9803
J	-1.3191	0.9794
Total	-1.1898	0.9832

2) Scale-Free Property

We also measured their cumulative distributions, as shown in Fig. 2 and Fig. 3. Those results also follow the power-law distribution according to the location of points in these figures. It indicates that all the networks hold the the scale-free property.

For further exploration, we fit parameters of these power law distributions. The power law indexes and R-squared are shown in Table II. From this table, the results of R-squared can all pass the test, and these power law indexes all belong to (1, 2]. Wang Lin pointed out[12] that when the power law index belongs to (2,3], the network obeys a mechanism of priority link; While belonging to (1,2], there may be other link mechanisms besides the priority link. We try to explain this phenomenon. In Zhiyuan Liu's study[9], the words with higher degrees are usually function words in Chinese, which are used to ease the sentences. When a node representing a word enters the network system, in addition to following the mechanism of priority link, it must be collocated with some function words, so that the sentences can follow the grammar rules. The particular law is also suitable to translated Chinese, so that we can obtain the parameter fitting results in Table II. In addition, if these function words are removed from sentences, people can still understand the meaning, but the network behind it will become fragmented[13].

C. *Parameter comparison between word co-occurrence networks of translated Chinese and standard Chinese*

In order to reveal the speciality of the word co-occurrence network of translated Chinese better, we constructed a standard Chinese network to compare with it. The corpus was specifically selected from The Lancaster Corpus of Mandarin Chinese(LCMC), which is a blueprint for the construction of ZCTC. The two corpora are comparable both in total size and the proportion of text types. According to the same experimental steps, we measured the word co-occurrence network parameters of the LCMC corpus and the corresponding six sub-corpora, as shown in Table III.

By Comparing Table III to Table I above, between the corresponding corpora, the number of nodes in translated Chinese network is less than the standard Chinese's, but the average degree<k> of translated Chinese network is also larger than standard Chinese's.Those results above are all caused by the characteristic "simplification" in translated Chinese. That is to say, under the same scale of corpus, there are fewer word types in translated Chinese, but the use of vocabulary in standard Chinese is more diverse. In addition, the diameter and average shortest path lengths in these translated Chinese networks are often shorter than the corresponding standard Chinese's, but their clustering coefficients are higher than the standard Chinese's. This is because words are more closely related in translated Chinese, and the usage frequencies of common words are more higher.

TABLE III. STATISTICAL PARAMETERS OF WORD CO-OCCURRENCE NETWORK OF STANDARD CHINESE

Type	Length	N	E	D	<k>	L	Lr	C	Cr
A	60k	10428	30253	12	5.8023	3.3975	5.4563	0.1314	4.9969×10-4
E	60k	9243	27890	10	6.0348	3.3504	5.2787	0.1332	5.9738×10-4
F	60k	9847	29758	11	6.0441	3.3673	5.3164	0.1402	4.8351×10-4
G	60k	10755	29863	11	5.5533	3.4096	5.6029	0.1404	4.6109×10-4
H	60k	6096	22422	10	7.3891	3.3579	4.5870	0.1422	1.1883×10-3
J	60k	6935	26546	13	7.6557	3.1894	4.5791	0.1752	7.8820×10-4
Total	1M	59743	342959	10	11.4811	3.0423	4.7693	0.2714	2.0183×10-4

V. CONCLUSION

In this paper, it starts from the viewpoint of complex network to build the word co-occurrence network with ZCTC and its six sub-corpora, such as Press reportage, Skill/trade/hobby and Popular lore. By counting the corresponding parameters of these networks, it reveals and verifies the small world effect and scale-free property of the translated Chinese network. Then compared with the standard Chinese network, we discover that the translated Chinese network has less nodes, higher average degree and clustering coefficient, which further reveals the characteristics of the translated Chinese: "simplification" and higher use frequency of common words.

Our experiment in this paper has verified that the complex network is effective as a measure of language and translation research. Meanwhile, we also realize that the above indexes are far from revealing the other characteristics of translated Chinese completely. In future, an more scientific and comprehensive index system will be formulated to deeply explore laws and characteristics in the translated Chinese network.

REFERENCES

[1] M. Baker, "Corpus-based Translation Studies: The challenges that Lie Ahead," Benjamins Translation Library, 1996, 175-186.

[2] M. Olohan, "Introducing Corpora in Translation Studies," Routledge, 2004.

[3] Vinay, J. Paul and J. Darbelnet. "Comparative stylistics of French and English: A methodology for translation", John Benjamins Publishier, 1995.

[4] Y. Zhu and X. Li, " A Quantitative Study on Lexical Features of Translated Chinese: Based on the Corpus of E-C Translated News Articles and the Corpus of Chinese News Articles(in Chinese)," Foreign Languages in China, vol. 88, Apr. 2019, pp. 81-90.

[5] R. Xiao and G Dai, " Lexical and grammatical properties of Translational Chinese:Translation universal hypotheses reevaluated from the Chinese perspective," Corpus Linguistics and Linguistic Theory, vol. 10, Jun. 2013, pp.11-55.

[6] G. Dai, "Collocational Features in Translated Chinese: A Case Study of Source Language(SL)(in Chinese)," Contemporary Foreign Languages Studies, vol.229,Jan. 2013, pp.50-55.

[7] H. Liu, "Language Is a Complex Network(in Chinese)," Journal of Shanxi University(Philosophy and Social Science Edition), vol.167, Sep. 2013, pp. 66-69.

[8] R. Cancho and R. Sole, "The Small World of Human Language," Proc. the Royal Society B: Biological Sciences, The Royal Society, July. 2001, pp.2261-2265.

[9] Z. Liu, M. Sun, "Chinese Word Co-occurrence Network:Its Small World Effect and Scale-free Property(in Chinese)," Journal of Chinese Information Processing, vol.96, Nov. 2007,pp.52-58.

[10] Z. Liu,Y. Zheng and M. Sun, "Complex Network Properties of Chinese Syntactic Dependency Network(in Chinese)," Complex Systems and Complexity Science, vol 18.Jun 2008, pp.37-45.

[11] Z. Xiao, "Corpus-Based Studies of Translational Chinese in English-Chinese Translation(In Chinese)," Shanghai,CN: Shanghai Jiao Tong University Press, 2012.

[12] L. Wang and G. Dai, "On Degree Distribution of Complex Network(in Chinese)," Journal of Northwestern Polytechnical University, vol. 24, Aug 2006, pp.405-409.

[13] M. Kurant, P. Thiran and P. Hagmann, "Error and attack tolerance of layered complex networks," Nature, vol. 76, Jan. 2007, pp. 388-394.