Translation for professionals: corpus-based study of translation universals in computing

Baorong Huang
Institute of Corpus Studies and Applications,
Shanghai International Studies University
huangbaorong2000@163.com

Abstract

Translation universals have been studied extensively in literature, politics, and business but rarely touched in technical field. This paper presents the investigation of translation universals in a new corpus consisting of articles translated by Chinese professionals in computing from English journals and magazines by using statistic findings, principal component analysis and visualization tools to unveil the general patterns, i.e., the translation universals and the individual characteristics in the specific domain of computing. Insights on the underlying motivations, guided by the entropy theory, the Principle of Least Effort, and the principle of relevance, are blended to explain the universals in technical translation in computing. To my knowledge, it is the first attempt that studies the translation universals in technical translation with both the translation corpus and the reference corpus collected from the same magazine.

1 Introduction

The juxtaposition of the language features and the comparison of them are widely used in the studies of Corpus Linguistics. Douglas Biber (1998) explained the power of language features spanning from lexicography to grammar in the analysis of the discourses, registers and varieties. Mona Baker (1995), in the study of Translational English Corpus (TEC), proposed several features to measure the deviations of the translational English text from the original English text, including TTR, mean sentence length, reporting structures and so on.

Language features are inexhaustive, and linguists, with their ingenious minds, often suggested new variables in the corpus-based studies of the translational Chinese text and the verification of translation universals, including n-gram analysis, connectives analysis, passive construction or specific constructions (Richard Xiao and Xianyao Hu, 2015), reporting structure and direct speech/free indirect speech (Libo Huang, 2015), BEI construction (Kaibao Hu, 2016), specific verbs like 'jinxin' (Guangrong Dai, 2016).

In order to cope with the challenge of the increasing number of features, multivariate analysis (Hermann Moisl, 2006), cluster analysis (Hermann Moisl, 2015) and triangulation (Paul Baker and Jesse Egbert, 2016) are proposed to handle the feature explosion and reveal the relations among these features and the patterns determined jointly by such features.

This paper applied principle component analysis (PCA) and visualization tools to unveil both the general trend and idiosyncratic features in the technical Chinese translations, and made a tentative unification of the information theory in communication, the Principle of Least Effort and the principle of relevance in pragmatics to elaborate on the reasons why the professionals deliberately chose the wordings and syntactic patterns that demonstrated normalization, simplification, and explication in the translation of technical Chinese in computing.

2 Related Literature

Common patterns in translation behaviors, called translation universals, after they have proposed by the scholars (Gideon Toury, 2004; Mona Baker et al., 1993; Alet Kruger, 2004; Kirsten Malmkjær, 2011), have been studied extensively in Chinese-English translation. Translation universals in a wide range of fields have been investigated, including politics (Kaibao Hu et al., 2015a), literature (Kaibao Hu, 2015b), and businesses (Feng Haoda et al., 2018), but few researchers in China had ever touched on the study of translation universals in technical Chinese translation.

3 Methodology

The purpose of the corpus study presented here used a new type of corpus, the Communications of CCF (China Computer Federation) corpus, to verify whether the translational universals, including normalization, explicitation and simplification, exist in technical Chinese translations produced by the professionals, with the reference materials contributed by the same group of professionals, and to reveal both the general trend in the Chinese translations and the characteristics of individual articles using the general variables in corpus linguistics (TTR, STTR, mean sentence length and lexical-to-function word ratio) and the pronoun ratio, one special indicator in English-Chinese corpus translation studies, assisted with the visualization tools such as scatterplots and PCA, a dimension reduction tool.

The paper presented the findings on that corpus, and attempted to reveal the motivations for the universals with the information theory, the Principle of Least Effort, and the principle of relevance in pragmatics. The particularities of the corpus and the specific analyses carried out for this are detailed below.

4 Corpus

The corpus data contains the articles in *Communications of CCF* in the period from Feb, 2017 to December, 2020. *Communications of CCF* (CCCF) briefs the advances in computer science to the members of the China Computer Federation (CCF), mostly the engineers and researchers working in computing. The articles in that magazines are contributed by the professionals in computing and edited by a special translation committee. In each volume of CCCF, there is a translation column that introduces the papers from the renowned publications, such as Science and Communications of the ACM, translated by the professors and PhD candidates in computing，whose names, together with their research interests, are clearly indicated in the translated articles. The translation quality in CCCF is further ensured by the mechanism that a special editorial committee of four professionals is established for that translation column, led by the senior member in CCF who is also an IEEE Senior Member.

The corpus data are divided into two parts. The reference part consists of the original Chinese articles from CCCF, contributed by professionals in computing, and the translation text consists of the translated Chinese articles contributed by the same group of professionals in that magazine. As both the Chinese articles and the translated articles are contributed by the same group of elite professionals instead of the ordinary translators, the corpus offers a strong and sound platform to contrast the stylistic difference between the translations and the original texts in technical Chinese.

In this corpus, for accurate comparison with the translations, only the technical Chinese articles in the special columns, special topics and viewpoints in CCCF are extracted. Typically, an article in CCCF consists of a title, a summary/abstract, several key words, a body, foot notes and references. For simplification, this corpus only contains the body of the article, and the section title, footnotes, citation numbers, figures and tables in the body are removed. After preprocessing, 547 articles are selected, of which 442 articles are of original Chinese and 105 articles are of translated Chinese.

4.1 Annotations and Statistical Results

The descriptive language study depends on the statistics collected through various tools developed by researchers in natural language processing. Generally, to gain insights from the raw data, manual/automatic annotation must be performed, which mainly include sentence segmentation, tokenization, POS tagging, and so on. In this paper, Stanza was used as it was a “language-agnostic fully neural pipeline for text analysis, including tokenization, multiword token expansion, lemmatization, part of speech and morphological

feature tagging, dependency parsing, and named entity recognition" (Peng Qi et al., 2020), supporting multiple languages such as English, German, Arabic, Russian and Chinese.

The annotation pipeline in this paper consists of sentence segmentation, tokenization, and pos-tagging. General variables in corpus linguistics, including TTR or STTR, mean sentence length and lexical-to-function word ratio, are used in this paper.

Table 1 lists the general variables in original Chinese Text (OCT) and translated Chinese text (TCT) side by side for comparison. Obviously, there are considerable gaps in STTR, mean sentence length and lexical to function word ratio, 2.7 for STTR, 11.66 for mean sentence length and 0.14 for lexical-to-function word ratio respectively. The gap in STTR conforms to the finds of Wang and Qin (2009) that "the STTR of non-literary translated Chinese is higher that of non-translated text".

	OCT	TCT
No. of articles	442	105
No. of tokens	1427654	283461
Types	50245	15935
Type-token ratio/100 words	**3.5**	**5.6**
Standardized type-token ratio (1000 words)	39.40	41.61
Sentences (total)	42904	10280
Mean sentence length (in characters)	**57.32**	**44.38**
Lexical-to-function word ratio	**2.24**	**1.87**

Table 1: General distribution in OCT and TCT

In the discussion of pronouns used in translated Chinese text, Richard Xiao and Xianyao Hu (2015, p. 101-103) pointed out that the pronouns are likely to appear in the translated Chinese text due to the influence of the existing pronoun in the source English text. Therefore, we decided to add the pronoun ratio as another variable in the statistical analysis. The results are shown in Table 2, which clearly indicates that the ratio of pronouns in TCT is significantly higher than that in OCT.

No. of pronouns	OCT	TCT
	16724	8034
Pronoun ratio	**1.17%**	**2.83%**

Table 2: Pronoun Distribution in OCT and TCT

4.2 Data Visualization and Exploration

The general corpus variables presented above show the general trends of the translational Chinese text against the original Chinese text in the corpus, with marked difference in TTR, mean sentence length and pronoun ratio. However, the mean values or the general trend somehow obscure the observations on the individual cases because some outliers in the population can elevate/downgrade the mean value by a certain degree. This worry, during the development of the corpus linguistics, is recognized in the research group, but not explored adequately and handled properly. Which factor contributes more to human's perception of the text as translation? Do all the translations in the corpus deviate similarly from the original text? These are the questions waiting for the researchers to unravel.

It's difficult, if not possible, to manually check all the statistics of each article in the corpus, but the indicators, or features, used in the analysis may not be exhaustive. Tukey (1977) proposed the exploratory data analysis and supporting tools to visualize and obtain the patterns of the data, in which scatterplots are a good way to reveal both the general trend and the individual situations. We developed a pair-wise scatterplot with Seaborn, a python visualization library, to illustrate the relevance and the significance of each statistical variable on CCCF corpus.

In Figure 1, the samples from TCT (orange dots) mingle with the samples from OCT (blue dots) and it is difficult to find a clear-cut boundary between the two types of text. Generally, the orange dots are within the range of the blue dots, especially for mean sentence length and STTR. For the lexical-to-functional ratio and pronoun ratio, the orange dots lie around the boundary of the blue dots, with some of them deviates considerably to the cluster of the blue dots. It is obvious that the lexical-to-functional ratio and the pronoun ratio play a much more important role in separating OCT from TCT. However, is it safe to conclude that TCT and OCT differs from each other? What if we combine these four variables and verify again?

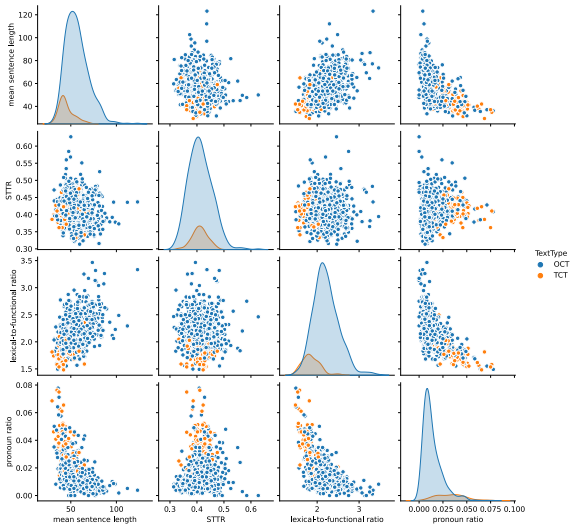

Figure 1: Pair-wise Comparison of the Variables in TCT and OCT

4.3 Principle Component Analysis

As shown in the pair-wise scatterplots in Figure 1, it's difficult to distinguish individual translated articles and translated articles linearly. Assume that there exists a non-linear plan across the high-dimensional space, whose dimensions are these features such as TTR, mean sentence length, lexical-to-function word ratio, and the ratio of pronouns, the paper used the principal component analysis to reduce the dimensions to 2D and visualize them to check whether the styles of OCT and TCT are truly different.

According to George Dunteman (1989, p. 7), the principal component analysis serves to "both simplify and impose some structure on the research domain… in reducing the number of variables from p to a much smaller set of k derived variables that retain the most of the information in the original p variables" through "linearly transform an original set of variables into a substantially smaller set of uncorrelated variables that represents most of the information of the original set of variables."

In this paper, the author decided to reduce the original four dimensions in the corpus analysis to two dimensions for 2D visualization in order to gain an intuitive understanding on the statistical distribution of the variables for individual articles and gain insights into the stylistic difference of OCT and TCT.

PCA in the scikit-learn (Pedregosa et al., 2011), an open-source Python machine learning module that offers a wide range of machine learning algorithms, is used for the principal component analysis, and Matplotlib, an open-source visualization library is used to visualize the results, as shown in Figure 2.

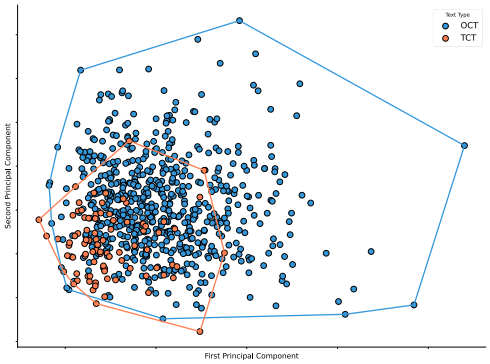

Figure 2: PCA Analysis of the Variables in OCT and TCT

The visualization in Figure 2 shows that there may not be a clear line to separate the translated Chinese text (in red) and the original Chinese text (in blue), despite the fact that the distribution of the translation Chinese articles is near the left-bottom corner and the distribution of the original Chinese text are much more varied. It seems that the majority of the translated articles falls within the scope of the original Chinese articles.

5 Translation Universals in technical Chinese translation: Normalization, Simplification and Explicitation

Normalization is the tendency to conform to the patterns of the target language (Baker, 1996). As shown in Figure 1 and Figure 2, the overlapped distribution clearly indicates that professional translators had tried to deliver the technical information in the patterns as close as to those in the target language, despite several outliers that occurred due to the wide discrepancy between the source language and the target language that required unreasonable efforts from the translators to bridge the gap.

In the translation of technical English text, normalization is often accompanied with simplification and explicitation.

Simplification is evidenced by the gap (11.66 characters) in the mean sentence length, which indicates that professional translators had split one original sentence in English into several Chinese sentences in their translations.

(1) *Source: Table 1 summarizes recent U.S. motor vehicle death rates, which translate to a fatality rate of 5 × 10-7/hour per vehicle, assuming an overall average speed of 40 MPH.* (Lala Jaynarayan et. al, 2020)
Translation: 表 1 总结了美国近期的机动车事故死亡率。假设车辆的平均速度为每小时 40 英里，则每辆车的致死概率为 5×10-7/小时。 (He Yukai and Gening, 2020)

In the above example, the translation of one source sentence contains two target sentences that are separated by "。", the period in Chinese, which contributes to the shorter sentence length in the statistical results. In addition, "assuming (假设)…" is brought to the start position in the second sentence in the translation, which conforms to the ordinary syntactic pattern in China.

Explicitation is reflected in the lower lexical-to-functional ratio (1.87) in TCT, which means that translators had explicitly shown the syntactic clues with more functional words.

(2) *Source: Self-certification of the Boeing 737 MAX led to the MCAS system, at the center of the two crashes, being declared non-safety-critical.* ((Lala Jaynarayan et. al, 2020)
Translation: 例如，波音 737 MAX 的机动特性增强系统就采用了"自认证"方式，后来引发了两起空难。事实证明，"自认证"难以保证系统安全。 (He Yukai and Gening, 2020)

In the above example, apart from the sentence splitting, the translators added "例如 (for example, VERB), 就(ADV)，了(PARTICLE)，后来 (later, NOUN) that established sign posts for readers to untangle the logical and temporal relations among these clauses, explicitly revealing the relations buried in the original English sentence.

Explicitation is also evidenced by the higher pronoun ratio (2.83%) in the translated Chinese text, which explicitly reveal the subject of the clause.

(3) *Source: As shown in the accompanying figure, we identify nine potential biases.* (Silva Selena and Kenney Martin, 2019)
Translation: 如图1 所示，我们识别了9 个潜在的偏见。(Jiang Ting et al., 2020)

In the above example, the translation achieved two types of explicitation. First, it concretized "the accompanying figure" to "图 1 (Figure 1)", an explicit sign post for readers to easily retrieve that figure in the article. Secondly, it uses a subject "我们", which is transferred from the original subject "we".

However, the explicitation of the pronoun in the translation contradicts the hypothesis of normalization to a certain degree, as the normalized translation should have a similar pronoun ratio to that of OCT. In addition, the higher TTR and STTR in the TCT somehow contradict the hypothesis of simplification, as a simplified version would produce a lower STTR and similar pronoun ratio in the translation. The motivations underlying the contradictions will be tentatively explained in the following section.

6 Entropy, least effort, relevance and their implications

6.1 Entropy, normalization and explicitation

Shannon (1948) proposed the encoder-decoder communication model and suggested that the entropy is the function over the possibilities of the variables, as shown in Equation 1:

$$H = -K\sum_{i=1}^{n} p_i \log p_i \quad (1)$$

where K is a positive constant. This equation means that the information load in a message is determined by the possibilities of its constituents. In order words, the information load in a sentence is determined by the possible choices of its constituents syntactically and paradigmatically, and the best choice is determined statistically based on the fact that "anyone speaking a language possesses, implicitly, an enormous knowledge of the statistics of the language" (Shannon, 1951).

In the translation process, translators, as bilingual speakers, implicitly processes the knowledge of the statistics of both the source language and the target language. When they make their lexical, semantic and syntactic choices, normally they encounter the information with higher uncertainties, because they can either render the translated text with some exotic elements, departing from the general patterns or norms in the target environments, surprising the audience and increasing the entropy in the text with unexpected items, or normalize the text and wipe out all the incongruities, increasing the readability, both of which, termed as literal translation and liberal translation, are totally legible and conducted frequently in translation practices. In other words, the entropy for the text produced by the translators is higher than the entropy for the original text in the target language, and it's the choices of translators that can increase/decrease the processing efforts of target readers.

Suppose that a translator encounters a sentence like "we incorporate users because their actions affect outcomes" (Silva Selena and Kenney Martin, 2019), a machine translation produced by Google is "我们纳入用户是因为他们的行为会影响结果", which is comprehensible. But, in Chinese, reasons often precede results, and it is illogical for 我们(we), the human beings, to incorporate another group of human beings. Thus, in translating this simple sentence, the end results "因为(because)用户的行为会影响结果，我们在模型中包含了用户(Jiang Ting et al., 2020)" not only normalize the original English sentence, but explicitly add "模型(model)" to show the relationship between the entities in the sentence.

In other scenarios, translators may choose a more radical strategy to follow the normal pattern, or statistical tendencies, in the target language. For example, the sentence below also contains the same pronoun "we".

(4) *Source: What Can We Learn from the Aviation Example With Respect to Autonomous Vehicle Dependability Requirements?* (Lala Jaynarayan et. al, 2020)
Translation: 自动驾驶汽车如何从航空案例中借鉴系统可靠性经验？ (He Yukai and Gening, 2020)

The translation summarized the original English sentence by dropping "we" and changing "Requirements(需求)" to "experience(经验)" because the collocation of "learn(借鉴)" with "Requirements(需求)", if not forbidden, sounds awkward and foreign to the common Chinese people. In this translation, the translators performed an ingenious maneuver that lowers the lexical surprises, thus reducing the processing efforts for the readers.

In other word, the translators for CCCF, a magazine in the technical communication that aimed to populate the concepts, viewpoints and trends in computing, worked hard to introduce the new ideas in the way that sounded most naturally to the Chinese audience, untangling the syntactic complexities of the English language and lowering their understanding efforts in such patterns as normalization and explicitation, by following the norms in the Chinese technical translation that emphasized on both accuracy and fluency (Liang Qichao, 1879; Li Wei, 1986), with recommended skills such as deletion and summarization (Yan Fu, 1898).

The results of such deletion and summarization actually contributed to the shorter sentence length and higher STTR/TTR ratio, because new information and new items from the external world implied that the chance of repetition was low, and the deletion and summarization drove the repetition even lower. The long sentence was split into shorter ones to reduce the entropy, improving the readability.

6.2 Balanced least effort, relevance and higher pronoun ratio

George Kingsley Zipf (1949), in the study of human behaviors, established the Principle of Least Effort, which means that one will strive to minimize the total work in solving the problems. Later Dan Sperber and Deirdre Wilson (1986) proposed the communicative principle of relevance, which requires that the information sent by the senders should not contain any ostensive behaviors that may invite different interpretations.

In the translation of articles in CCCF, the translators are the speaker in Zipf's speaker-auditor model, whose economy is maximized if word-for-word translation is adopted as this requires little efforts in manipulating the word orders to conform to the norms of the target language. On the other hand, the readers, the auditors Zipf's model, maximize the economy if the translation contains no exotic elements that deviate noticeably from the ordinary language patterns.

In the CCCF corpus, as the translators are professionals in this specific field, it can be safely concluded that they have the ability to produce the translation in the way just like they produce a Chinese text. In the translation of *The Internationale*, the pronoun in the original text is dropped in the Chinese translation to highlight the theme (Wang Hailong, 2021). Why do not they normalize the use of pronouns in the technical translation?

The motivations here reflect translators' tradeoff between their economy and the economy of the readers. The explicitation of pronouns under the impact of English texts, named 'Source Language Shine Through' in English-Chinese translation (Dai Guangrong and Xiao Richard, 2010), mirrored the syntactic construction of the original English text, because pronoun construction is a totally legible lexical and syntactic construction in Chinese, which does not violate the norms. Therefore, to achieve the economy of the translators and reduce the processing efforts in translation, the translators naturally determined to use this strategy, despite a minor increase in the pronoun ratio.

However, the readers' economy cannot be totally ignored and the professional translators must strike a balance. If the translators had made considerable efforts in rephrasing the entire sentence that broke the strict correspondence both lexically and syntactically, as shown in Example 4, they dropped the pronoun "we (我们)" in the translation, as shown in Example 4. It's the dynamic balance in translators' economy and readers' economy that modulated translators' handling of pronoun.

7 Conclusion

This paper presents a tentative corpus-based analysis of the technical translations in CCCF, a popular magazine among Chinese professionals in computing, and applies the scatterplots and the principal component analysis, a powerful tool in dimension reduction to check and visualize the distribution of individual articles, to validate the assumptions of translation universals and improve the traditional analyses that were mostly based on the mean or overall trends in the data. In the analysis, the ideas in information theory, the Principle of Least Effort, the principle of relevance and traditional norms in Chinese technical translation are combined to explain the choices made by the professionals in their translation. However, it seems that more in-depth corpus-based analyses related to the variables used in the technical English should be carried out to investigate and reveal more deciding dimensions, and the advanced tools need to be developed in order to reduce the analysis efforts. In addition, the entropy and the balanced effort hypothesis need to be verified with more research on translators' psychological activities.

Acknowledgments

The author appreciates the valuable feedback from the anonymous reviewers.

References

Alet Kruger. 2004. Corpus-based translation research: its development and implications for general, literary and Bible translation. Acta Theologica Supplementum 2.

Baker, Mona. 1995. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. Target. 7. 223-243. 10.1075/target.7.2.03bak.

Baker Mona. 1996. Corpus-based Translation Studies: The Challenges that Lie Ahead. In Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager Edited by Harold Somers John Benjamins Publishing Company, Amsterdam.

Baker, Mona, G. Francis, E. Tognini-Bonelli. 1993. Corpus linguistics and translation studies:

Implications and applications in Text and Technology: In Honor of John Sinclair (pp. 233-250). John Benjamins, Amsterdam.

Claude E. Shannon. 1948. A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379–423.

Claude E. Shannon. 1951. Prediction and Entropy of Printed English. Bell System Technical Journal, 30, 50-64.

Dai, G., & Xiao, R. 2010. 'SL shining through' in translational language: A corpus-based study of Chinese translation of English. In R. Xiao (Ed.), Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies 2010 Conference (UCCTS2010) Lancaster University. http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2010Proceedings/papers/Dai_Xiao.pdf

Dan Sperber, Deirdre Wilson. 1986. Relevance: Communication and Cognition. Oxford University Press, Oxford, UK.

Douglas Biber, Susan Conrad, Randi Reppen. 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press, Cambridge, UK.

Feng Haoda, Crezee Ineke, Grant Lynn. 2018. Form and meaning in collocations: a corpus-driven study on translation universals in Chinese-to-English business translation. Perspectives. 26. 1-14. 10.1080/0907676X.2018.1424222.

George H. Dunteman. 1989. Principal Components Analysis. SAGE Publications, Newbury Park, CA.

George Kingsley Zipf. 1949. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley Press, Cambridge, Mass.

Guangrong Dai. 2016. Hybridity in Translated Chinese: A Corpus Analytical Framework. Springer, Berlin Heidelberg.

He Yukai, Ge Ning. 2020. Zidongjiashiqicheanquan: conghangkongyejiejianjingyan [Autonomous Vehicle Safety: Lessons from Aviation]. Zhongguojisuanjixuehuitongxun [Communications of China Computer Federation], 2020, 16(10): 84.

Jiang Ting, Wang Fangye, Lu Dun. 2020. Suanfa, pingtaihe zhongzhupianjian [Algorithms, Platform and Ethnic Bias]. Zhongguojisuanjixuehuitongxun [Communications of China Computer Federation], 2020, 16(2):76

John Tukey. 1977. Exploratory Data Analysis. Addison Wesley, Cambridge, Mass.

Kaibao Hu, Feng Pan, Xin Li. 2015. Jiyuyuliaokude jizhezhaodaihui hanyingkouyiyanjiu [A Corpus-based study of Chinese-English Conference Interpreting]. Foreign Language Teaching and Researching Press, Beijing.

Kaibao Hu. 2015. Jiyuyuliaokude shashibiyaxiju hanyi yanjiu [A Corpus-based study of English-Chinese Translation of Shakespeare's Plays]. Shanghai Jiaotong University Press, Shanghai.

Kaibao Hu. 2016. Introducing Corpus-based Translation Studies. Springer, Berlin Heidelberg.

Kirsten Malmkjær. 2011. Translation Universals in The Oxford Handbook of Translation Studies Edited by Kirsten Malmkjær and Kevin Windle. Oxford University Press, Oxford, UK.

Lala, Jaynarayan & Landwehr, Carl & Meyer, John. 2020. Autonomous vehicle safety: lessons from aviation. Communications of the ACM. 63. 28-31. 10.1145/3411053.

Liang Qichao. 1879. Lunyishu in Zhongguokejifanyishiliao [On Translating books in The History of Technical Translation in China] Edited by Li Nanqiu. Hefei: University of Science and Technology of China Press, 1996.

Libo, Huang. 2007. Jiyu hanying/ying han pingxing yuliaoku de fanyi gongxing [Translation universals research based on Chinese-English/English-Chinese parallel corpus]. Fudan University Press, Shanghai.

Libo Huang. 2015. Style in Translation: A Corpus-Based Perspective. Springer, Berlin Heidelberg.

Li wei (Eds.). 1986. Kejifanyigongzuoshouce [Manual on technical translation]. Tianjin Technology Translation Publication Co., Ltd., Tianjin.

Moisl, Hermann, Warren Maguire and Will Allen. 2006. Phonetic Variation in Tyneside: Exploratory Multivariate Analysis of the Newcastle Electronic Corpus of Tyneside English. In Frans Hinskens (ed.) Language Variation – European Perspectives: Selected Papers from the Third International Conference on Language Variation in Europe (ICLaVE 3), Amsterdam, June 2005. 127-41. John Benjamins.

Moisl Hermann. 2015. Cluster Analysis for Corpus Linguistics. De Gruyter Mouton, Hague.

Paul Baker and Jesse Egbert. 2016. Introduction. In Triangulating Methodological Approaches in Corpus-Linguistic Research Edited by Paul Baker and Jesse Egbert, Routledge, New York and London.

Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020.

Richard Xiao, Xianyao Hu. 2015. Corpus-Based Studies of Translational Chinese in English-Chinese Translation. Springer, Berlin Heidelberg.

Silva, Selena and Kenney, Martin. 2019. Algorithms, platforms, and ethnic bias. Communications of the ACM. 62. 37-39. 10.1145/3318157.

Toury, Gideon. 2004. Probabilistic explanations in Translation Studies. 10.1075/btl.48.03tou.

Yanfu. 1898. Tianyanlunyiliyan in Zhongguokejifanyishiliao [On Translating Natural Selection in The History of Technical Translation in China] Edited by Li Nanqiu.. Hefei: University of Science and Technology of China Press, 1996.

Wang Hailong. 2021. Hanyubiaoshuzhongdaicizhuyushenglvexianxiangde yuyixuetantao: guojigeyiwenyuguhanyudaicizhuyushenglvexianxian g [Semantic Discussion on Subject Pronoun Ellipsis in Chinese Expression: Translation of The Internationale and Ellipsis of Pronoun Subject in Ancient Chinese]. Journal of Jiansu Normal University (Philosophy and Sciences Edition). 2021,47(2):22-37.

Wang Kefei, Qin Hongwu. 2009. Ying yi hanyuyan tezheng tantao —— jiyu dui ying yuliaoku de hongguan fenxi [A parallel corpus-based study of general features of translated Chinese]. Waiyu xuekan [Foreign Language Research] (1): 102–105.

Wang Kefei, Hu Xianyao. 2008. Jiyu yuliaoku de fanyihanyutihui tezhenyanjiu [Corpus-based study of lexical features of translated Chinese]. Zhongguofanyi [Chinese Translators Journal], 29(06):16-21+92.