Translation and Scientific Terminology

A Corpus-Based Multilingual Study

Meng Ji
University of Tokyo, Japan

Abstract

It is well known that the establishment of modern scientific language in the late nineteenth century was instrumental in the making of China's and Japan's early modern scientific identity. The current study aims to offer an original empirical investigation of early modern Chinese scientific terminology from a cognitive and functional perspective, which has been rarely explored before. Through a detailed corpus-based linguistic investigation, the present study probes the complex historical process of cross-cultural and cross-linguistic scientific exchange between the West and China and Japan in the late nineteenth century. The new insights brought about by the novel use of statistical methods point to potentially prolific directions for future research in contrastive historical linguistics and science history.
Keywords: corpus linguistics, translation studies, multilingual studies (Chinese, Dutch, English, French, German, Japanese), scientific terminology, corpus statistics.

1. Introduction

The main purpose of this paper is to explore the complex historical process of cross-cultural and cross-linguistic interaction between China, Japan and the West in the late nineteenth century. Instead of following a prescriptive approach to the subject matter which characterizes many past studies, the present paper attempts to formulate and test theoretical hypotheses through the generation and exploration of quantitative textual data retrieved from newly created large-scale online multilingual databases of historical texts. It represents an important effort in the development and innovation of research methodologies and analytical techniques in historical linguistics and related fields such as science history and cross-cultural studies.

The establishment of a working model for the translation of scientific terminology played a central role in the introduction and assimilation of imported Western scientific concepts into the traditional cultural and philological systems of China and Japan in the late nineteenth century. It was a daunting task which entailed a thorough understanding on the part of cross-cultural mediators (or translators) of the source and target languages, as well as an in-depth understanding of Western sciences practised at the time. It is our hypothesis that the huge difficulties implied in devising a working system for cross-cultural scientific exchange between China, Japan and the West were not only ideological and conceptual, but also linguistic and highly technical in terms of the borrowing and adaptation of subject-specific Western scientific expressions into traditional Chinese and Japanese (Ji, 2010a).

It should be noted that by the late nineteenth century, the network of cross-cultural scientific exchange developed in East Asia was already entrenched and largely dynamic that studies purely focused on the etymology of individual words would hardly be able to capture the whole picture and thus render less prolific (Shen, 2001). This paper therefore attempts to approach the topic from a novel cognitive and functional perspective that will bring us valuable insights into the formation of a modern scientific terminology in the character-based writing system of Chinese and Japanese. That is to say, instead of exhausting historical material on the etymological origin of scientific terms coined at the time, we concentrate on the micro-structural features and cognitive functions of relevant linguistic events

highlighted in the database, which in turn will prepare solid empirical ground for further discussions on the evolution of the Chinese and Japanese lexis in the late nineteenth century.

2. Database for textual analysis

The raw linguistic data used here were extracted from a large-scale database of early Chinese and Japanese translations of imported Western scientific works. It was originally developed and maintained by the University of Heidelberg, Germany (Lacker, et al, 2001). The textual material examined in this paper comprises five important pieces of bilingual lexicographical works produced and widely circulated in the late nineteenth century, which represented the frontier of contrastive scientific terminology between Chinese, Japanese and main European languages at the time. Concretely speaking, they are the *English Chinese Dictionary with Punti and Mandarin Pronunciation* (1866-9) by the German missionary Wilhelm Lobscheid; *A Vocabulary and Handbook of the Chinese Language* (1872-3) by the American Board missionary Justus Doolittle; *Tetsugaku Jii* (*A Dictionary of Philosophical Terms*, second edition) (1884) by Inoue Tetsujiro and Ariga Hisao; *Dictionnaire Français-Chinois* (contenant les expressions les plus usitées de la langue mandarine) (1884) by the French missionary Seraphim Couvreur; and *Nederlandsch-Chineesch Woordenboek met de Transcriptie der Chineesche Karakters in het Tsiang-Tsiu Dialekt* (1886) by the Dutch sinologist Gustave Schlegel.

Table 1 Five bilingual dictionaries of translated scientific terminology

NO.	AUTHOR	DATE	TRANSLATION DIRECTION	CORPUS SIZE (IN CHARACTER WORDS)
1.	Wilhelm Lobscheid	1866-9	(German) English → Chinese	5, 121
2.	Justus Doolittle	1872-3	(American) English → Chinese	2, 136
3.	Inoue Tetsujiro	1884	English → Japanese	3, 912
4.	Seraphim Couvreur	1884	French → Chinese	500
5.	Gustave Schlegel	1886	Dutch → Chinese	3, 782

Table 1 shows some basic information of the five bilingual lexicographic works selected for the current study. The reason for highlighting these five works from the rich referential material collected by the Heidelberg database was threefold. First of all, they were chosen under the consideration that they represent six languages heavily involved in the cross-cultural scientific exchange in East Asia at the time, namely, American English, (German) English, Chinese, Dutch, French and Japanese. In contrastive linguistics, especially translation studies, the source language may well serve as a controlling factor in the exploration of textual patterns underlying the Chinese and Japanese translations of Western scientific works. Furthermore, the representativeness of the language pairs studied might also be of help in verifying the existence of translation universals from a historical perspective (Baker, 2004; Laviosa, 1998).

Secondly, given the gap in the publication dates, the five contrastive lexicographic works may be roughly classified into two groups: (1) the mid-nineteenth century group containing Lobscheid's and Doolittle's translations and (2) the late-nineteenth century group including Inoue's, Couvreur's and Schlegel's translations. Such time intervals may be explored to examine the continuity or even possible legacy of earlier works on later translations, and the independence among works produced around the same time such as the former two or the latter three. This would also help elucidate the different linguistic modes and patterns developed by the five translators when working on early Chinese and Japanese scientific terminology or nomenclature in the late nineteenth century.

Lastly, as the last column of Table 1 shows, despite the different sizes of the five texts, they are all large enough to

make a likely statistical comparison among them. A distinctive feature of the current study consists in its use of quantitative linguistic data in the study of historical scientific translation. Textual features gathered in sufficient quantity make up a prerequisite for any corpus-based language study (Ji, 2010b). In processing the numerous textual traits extracted from the database, the use of statistical methods will help bring to light revealing patterns in the textual material which would appear inconspicuous to the naked eye. The analytical technique reported here represents a major methodological innovation in terms of the introduction of statistics from the social sciences to humanistic research such as science history and historical contrastive lexicography.

3. Textual phenomena under study

Due to the pilot nature of this paper, we single out two specific textual features of translated scientific terminology, with a view to testing the productivity of the methods developed for corpus-based historical contrastive lexicography. The two textual features highlighted are token length and functional particle. Token length refers to the number of characters making up a proper character word in historical Chinese and Japanese. It is an important linguistic feature in the study of textual genres. For instance, within the context of scientometrics, token length provides an effective measurement of the succinctness and referability of modern scientific language. In modern corpus linguistics, token length has been explored extensively for the purpose of natural language processing and automatic text genre disambiguation (Van Gijsel et al, 2006; Stamatatos, 2001). Its validity and usefulness for historical contrastive lexicography, however, remains to be tested.

Past studies show that modern Chinese lexis is predominately disyllabic (i.e. composed by two characters), whereas in ancient or classical Chinese, it was mono-syllabic words which prevailed (Needham and Robinson, 2004; Ji, 2010c). It is hoped that through a corpus-based statistical comparison of the five Chinese and Japanese translations of Western scientific works, the present study would help illustrate the various attempts made by influential cross-cultural mediators at the time, including foreign missionaries, sinologists and native Chinese and Japanese scholars, at forging a viable modern scientific lexis for the character-based writing system of Chinese and Japanese. In this paper, we shall first attempt to map out the distributional patterns of the running tokens in each text in terms of their varying lengths, in an effort to gauge the similarities and dissimilarities among the five translations.

Another origin observation made in this paper regards the creative use of functional particles in the five historical translations. Within the context of the current study, a functional particle is defined as a basic semantic or grammatical unit of translated terminology or nomenclature. It is often attached to the main part of a character word to indicate the abstract concept or metaphorical reference conveyed by the linguistic expression. The type of functional particles described here resembles to suffixes or prefixes of inflectional languages. It represented a creative use of classical Chinese and Japanese, given the fact that these two historical languages were essentially non-inflectional, in contrast with most European languages. A whole set of functional particles, as the current study will elucidate, were already coined and put into practice by the five translators in the late nineteenth century, which eventually became an integral part of modern Chinese and Japanese scientific terminology.

The experimentation with these functional particles proved instrumental in the establishment of a working modern scientific lexicon in Chinese and Japanese. That is because, at the conceptual level, due to the huge social and cultural differences between the West and the traditional Chinese and Japanese societies, many scientific terms or expressions imported from the West were totally unknown to the target audience at the time, severely hindering the introduction and further penetration of Western scientific ideas and concepts among the target readership. One possible way of solving the problem, as the five translated texts seem to suggest, was to contrive a linguistic technique or device that

might allow some conceptual commensurability between the two major language systems in question. Since by the late nineteenth century, modern scientific register in main European languages had already achieved a level that was far more sophisticated and elaborated than the traditional scholarly writing style of Chinese and Japanese, the borrowing of functional expressions from the West had turned out to be most cost-effective.

The development of linguistic devices in Chinese and Japanese comparable to functional expressions in modern scientific register in Western European languages was not that straightforward as it might have been for cognate languages. It first and foremost involved a thorough and painstaking re-examination of the target knowledge body, searching for expressions of metaphorical references parallel to their Western counterparts. In cases where lexical borrowing of readily-available words was deemed unfeasible, a number of functional particles had been coined to facilitate the transmission of concepts and ideas between the two knowledge systems. We shall endeavor to identify various recurrent types of functional particles found in the five translations. The findings uncovered in this paper will throw new light on the complex process in which the existing knowledge structure of the traditional Chinese and Japanese societies was gradually transformed through an influx of scientific translation in the late nineteenth century.

4. HCA (Hierarchical Cluster Analysis) of token lengths distribution in the five translations

Table 2 Comparative statistics of token lengths

Token Lengths	Lobscheid	Doolittle	Inoue	Couvreur	Schlegel
1	262	194	48	48	487
2	2, 158	1, 033	2, 680	334	1842
3	1, 111	501	569	75	579
4	1, 100	269	410	25	561
5	264	91	100	4	161
6	125	27	19	1	105
7	57	9	7	1	32
8	23	8	1	2	10
9 or above	21	4	3	1	5

Table 2 exhibits the comparative statistics of token lengths in the five translations of scientific terminology. In the left column of the table, the token length of a word is measured by its constituent character (s) ranging from one to nine or above. For example, in the Chinese translation of *aerometry* 度−氣−體−之−理 (measure-air-form of existence-of-theory) (du-qi-ti-zhi-li) (Lobscheid, 1866-9), the length of the newly coined Chinese scientific term is five-character long; or in the Chinese translation of *nominative*即−變−詞−格−之−第−一−法−也 (thus-change-lexis-rules-of-the-first-law-affirmative particle in classical Chinese) (ji-bian-ci-ge-zhi-di-yi-fa-ye) (Schlegel, 1886), the novel Chinese linguistic expression is distinctively formed by nine individual characters, implying the experimental attempts made by the translator at introducing a new concept or an unknown linguistic phenomenon to the target readership. A first glance at the table finds that in all five translations, the vast majority of word tokens are disyllabic, followed by tri-syllabic and quadric-syllabic words. Monosyllabic and penta-syllabic words display a sharp decrease in number when compared to the top three, forming the third layer of the database in terms of word token lengths. Poly-syllabic words composed by six characters or more make up the fourth tier of the word token hierarchy, totalling no more than three per cent of the entire database. To visualize the patterns indicated in the cross-tabulation, a linear histogram representing the distribution of varying word tokens in the five translated texts was plotted and given below.

Diagram 1 Distribution of token lengths in the five historical translations of scientific terminology

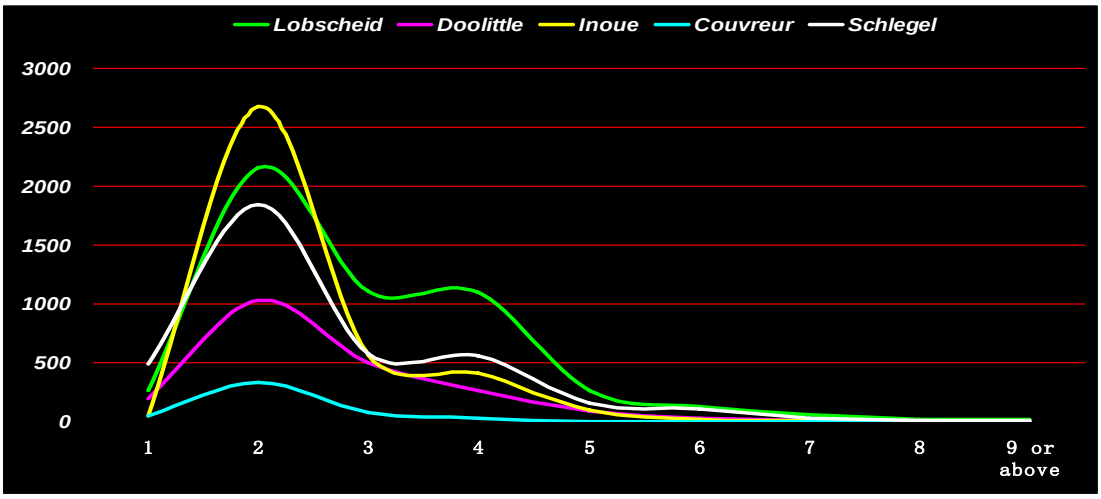


Diagram 1 gives us more direct access to the information hidden behind the abstract figures. In the first three categories, namely word tokens composed by one to three characters, the shape of the bell curve in coloured lines resembles each other. Corresponding to what was found in Table 1, disyllabic words form the sharp crest of the bell curve in all colours. The number of tokens picks up at the fourth category of quadric-syllabic words, after which the bell-shaped curve proceeds further and further into a long tail along the X axis. Despite the striking similarities in the general shape of the bell curve representing each of the five translations, subtle yet important variations seem to emerge to the surface. This regards the proportions between different token length categories within each translation. For instance, although the frequency of occurrence of disyllabic words in Inoue Tetsujiro's work (1884) features the highest on the graph, the number of quadric-syllabic words found in his work is far less conspicuous compared to other translations. It is worthwhile asking whether such variation within each translation should be taken into consideration when measuring the similarities among the five translated texts.

Table 3 Agglomeration schedule of HCA of intra-textual dissimilarities (based on token length)

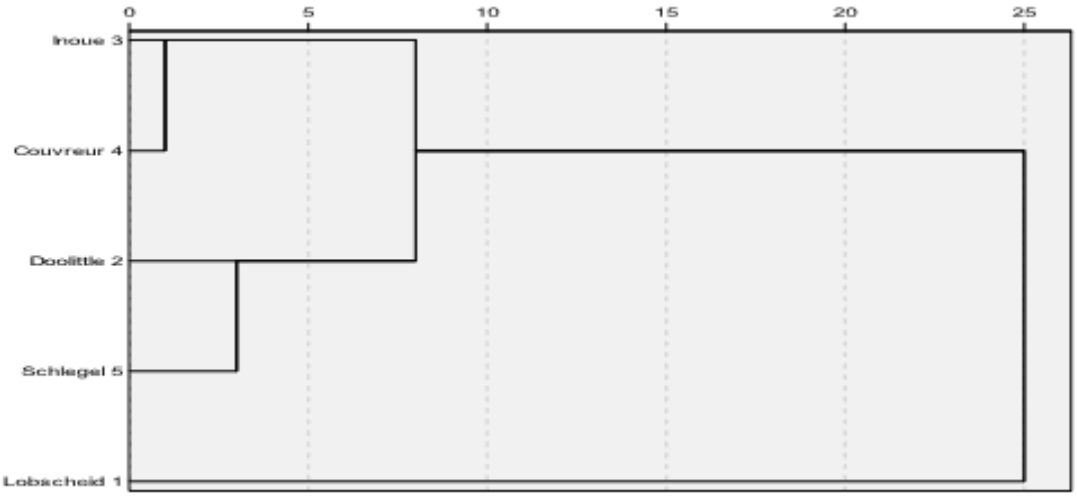
Stage	Cluster Combined		Distance Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	4	.987	0	0	3
2	2	5	.978	0	0	3
3	2	3	.958	2	1	4
4	1	2	.889	0	3	0

To reveal the latent structure of the database regarding the distribution of token lengths, we introduce the statistical procedure of Hierarchical Cluster Analysis (HCA) in this part of our study. HCA is an important exploratory technique in statistics, which has been used widely in the social sciences. To the best of my knowledge, its usefulness and validity for historical contrastive linguistics has been rarely tested before. Within the context of the current study, HCA is used to group translated texts based their computed similarities and dissimilarities. Due to the exploratory nature of the procedure, the preliminary result obtained at this stage remains to be tested in future research with textual material collected at a larger scale. The new insights brought about at this stage, however, will help deepen

our understanding of the underlying structure of the pilot database, giving rise to the formulation of new theoretical hypotheses regarding the correlation and independence among the five translations under investigation.

Table 3 shows the agglomeration schedule of the HCA based on the computed dissimilarities among the five translations. It is a numerical summary of the cluster solution. As may be seen from Table 3, at the first stage of the HCA, case 3 (Inoue) and case 4 (Couvreur) are combined, for they have the smallest distance. The cluster created by their joining next appears in Stage 3. At Stage 3, the clusters created at Stage 1 and 2 are joined. The resulting cluster next appears in Stage 4. In assessing the statistical result, a good cluster solution sees a sudden gap in the distance coefficient, and the solution before the gap indicates the good solution. From the distance coefficients column, it is easy to find that the largest gap in distance coefficient occurs between Stage 3 and 4, indicating a two-cluster solution to the classification of the five translations in comparison.

Diagram 2 HCA dendrogram of intra-textual dissimilarities (based on token length)



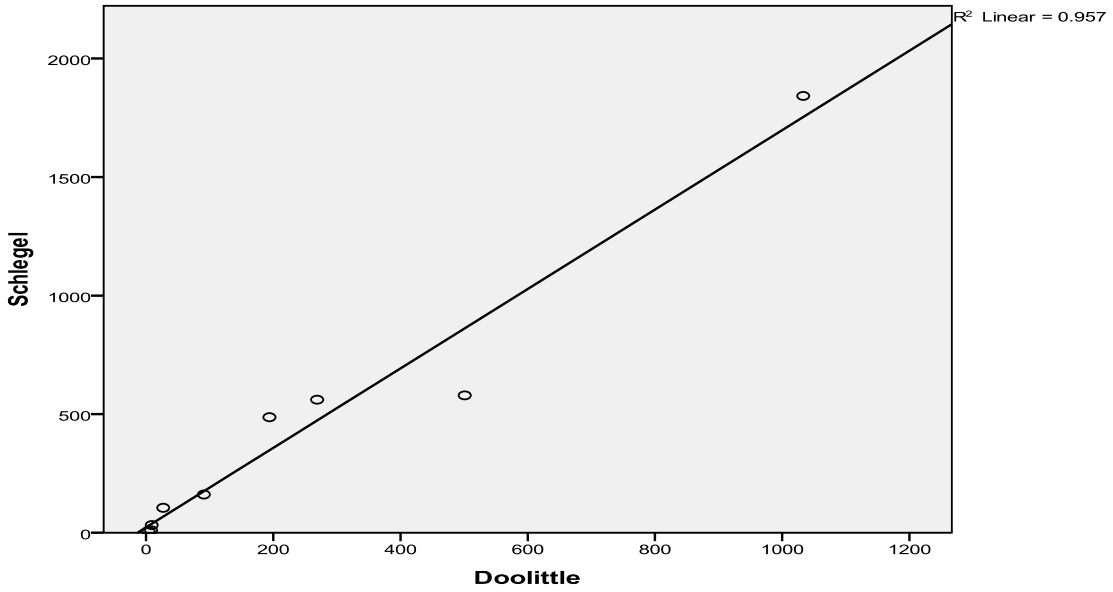
The decisiveness of this classification scheme is reflected in the dendrogram, which is a graphic demonstration of the clustering process conducted by the HCA. From right to left, the initial splitting of the tree forms two clusters. The upper cluster contains four later translations including Inoue, Couvreur, Doolittle and Schlegel, whereas the lower cluster refers to Lobscheid's work only. The top-level cluster suggests that despite the apparent similarities among the five translations, Lobscheid's translating style distinguishes itself from the rest in a most decisive manner with regards to the use of length-specific expressions throughout his work. One possible explanation for this statistically detected difference is his idiosyncratic use of Chinese four-character expressions as suggested in Diagram 1.

The rationale behind Lobscheid's particular approach to scientific terminology in his Chinese translation may well been due to his background as a native speaker of German, though he chose English as his working language, presumably to increase the influence of the dictionary that he was working on. It is noted that the extensive use of multiword expressions forms a distinctive feature of German scientific vocabulary (Keppler, 1955). Lobscheid's intuitive use of compound phrases in his Chinese translation of English scientific nomenclature thus lets his work stand out from the other four translations under investigation. Lobscheid's preference of Chinese multiword expressions may also be due to his own understanding of the Chinese language, its traditional scholarly writing style, as well as the ideological principle underpinning his cross-cultural enterprise. Four-character expression is a unique type of phraseology in Chinese which tends to be associated with the stylistic elegancy and formality of classical

Chinese. Moreover, it is a highly conventionalized morpho-syntactic structure charged with idiomaticity. For instance, it is estimated that more than ninety per cent of Chinese idioms are composed by four characters, causing the common mis-understanding that any four-character expression is idiomatic in Chinese. Lobscheid's heavy use of four-character phrases in his dictionary may well be an indication of his target-oriented approach to the translation and introduction of Western scientific ideas and concepts into the existing knowledge base of China.

The upper cluster is then further split into two smaller clusters containing case 3 (Inoue) and case 4 (Couvreur), and case 2 (Doolittle) and case 5 (Schlegel). As mentioned at the outset of the study, the deliberately built diachronic framework of the five translations may be explored to examine the continuity or possible legacy of earlier translations on later works. The statistical result computed by the HCA seems to suggest that there is a higher level of similarity between Inoue's (1884) and Couvreur's work (1884), and between Doolittle's (1872-3) and Schlegel's work (1886). To examine the possible legacy of Doolittle's translation on Schlegel's work, we proceed to compute the Linear Regression (LR) test, which is widely used in quantitative linguistics to analyze the relationship between the controlling factor and the dependent variable. Within the context of the current study, LR is used to explore the influence of earlier Chinese or Japanese translations of Western scientific works on latter translations. To illustrate the computer-assisted textual analysis, we choose to study the detected correlation between Doolittle's and Schlegel's translations as an example.

Diagram 3 Histogram of linear relationship between Doolittle's and Schlegel's work



As with most parametric statistical tests, in sight of the limited range of textual data collected in the current study, we first proceed to check the normality of the linear relationship between the two variables under study. A commonly used technique for such purpose is the graphic demonstration of intra-textual relationship. The simple scatter-plot given above depicts the mode and pattern in which the two factors correlate with each other in terms of the distribution of the various token length categories. It becomes clear from the scatter-plot that in most cases, Schlegel's use of length-specific tokens varies in proportion to Doolittle's use of corresponding lexical categories. This is shown in the graph as most scatters cluster round the best fitting line with a regression coefficient as high as 0. 957. This linearity check suggests that despite the limited amount of data collected in this case study, the two

variables under investigation, i.e. Doolittle's and Schlegel's translations, exhibit a sound linear relationship that is suitable for linear regression analysis.

Table 4 ANOVA test of Doolittle's and Schlegel's work

ANOVA[b]						
Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	2610967.821	1	2610967.821	154.964	.000[a]
	Residual	117941.735	7	16848.819		
	Total	2728909.556	8			
a. Predictor: Doolittle; b. Dependent Variable: Schlegel						

Table 4 reports the result of the ANOVA test as an integral part of the linear regression analysis. It tests the acceptability of the model from a statistical perspective. The regression row displays information about the variation accounted for by the linear regression model thus built. Accordingly, the Residual row shows information regarding the variation that is not covered by the statistical model. As one may see from the table, the amount of the Sum of Squares explained by the Regression model is much higher than that left out in the Residual half. This suggests that the linear regression model has captured much of the variation in the dependent variable, i.e. Schlegel's use of length-specific words, in relation to Doolittle's use of corresponding phrasal categories in his translation. The value of F statistics is significantly lower than the threshold value of five per cent, suggesting that the variation explained by the model is not due to chance. The ANOVA summary validates the efficiency of the model. It however does not directly address the strength of correlation between the model and the dependent variable.

Table 5 Summary of linear regression model

Model Summary [b]				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.978[a]	.957	.951	129.803
a. Predictors: (Constant), Doolittle; b. Dependent Variable: Schlegel				

Table 5 shows the model summary which reports the strength of the relationship between the model and the dependent variable. R stands for the correlation coefficient which measures the linear relationship between the observed and model-predicted values of the dependent variable. Within the context of the present study, R value indicates the level at which the linear regression has rightly predicted the frequency of occurrence of length-specific tokens in Schlegel's work based on the features of Doolittle's translation. R value ranges from zero to one: the higher the predicting power of the regression model, the higher the R value. In an idealized setting, the R value would reach one signifying the identification of the observed and predicted values of the dependent variable. From Table 5, it becomes clear that the original R value is as high as 0.978, pointing to a strong relationship between the model and the dependent variable. R Square refers to the coefficient of determination. It is the squared value of the correlation coefficient. It shows that more than ninety per cent of the variation in Schlegel's use of length-specific scientific terms has been rightly predicted by the regression model, i.e. the computed similarities between Doolittle's and Schlegel's works. The validated strong relationship between Doolittle's and Schlegel's works requires further detailed textual analysis of the two translations. Nevertheless, the preliminary findings presented here would seem to point to a shared approach to Chinese scientific terminology between the two translations, despite the different source

languages involved, i.e. American English in the case of Doolittle and Dutch in Schlegel's case, as well as the distinctive social and cultural circumstances surrounding the publication of each translation in different time periods.

5. HCA of the distribution of functional particles in the five translations

Table 4 Comparative statistics of functional particles (e.g. suffixes and prefixes)

No.	Functional particles categories	Lobscheid	Doolittle	Inoue	Couvreur	Schlegel
1.	GMF	56	0	41	4	0
2.	SCF_AC	30	10	103	82	96
3.	SCF_CC	19	53	19	34	15
4.	SCF_CP	5	22	127	8	7
5.	SCF_DE	54	158	10	90	118
6.	SCF_DF	21	12	103	10	39
7.	SCF_ML	53	23	133	96	49
8.	SCF_PA	96	51	59	160	211
9.	SCF_PE	102	251	3	82	60
10.	SCF_PQ	11	5	14	24	14
11.	SCF_QA	5	14	112	14	18
12.	SCF_SCP	32	1	58	32	43
13.	SCF_SQ	190	22	9	40	79
14.	SCF_TSR	35	31	6	38	32
15.	SCF_SS	37	321	45	36	52
16.	SCG_TG	40	5	68	36	78
17.	SCF_TS	47	16	90	32	89
18.	NA	167	5	0	182	0
19.	Total	1, 000	1, 000	1, 000	1, 000	1, 000

Table 4 offers a tentative classificatory framework of several recurrent functional particles identified in the five translations. The annotation system developed in the current study includes seventeen marking-up labels. They may be divided into two main categories of GMF and SCF, in accord with the grammatical and semantic-cognitive functions they assume in early Chinese and Japanese scientific nomenclature. GMFs refer to prefixes or suffixes indicating the grammatical function of a translated term. For example, a typical GMF instance in the five translations is the use of the Chinese character 的 (of) (de) as the ending component of the translated scientific term. It serves to indicate the grammatical function of the expression as an adjective. Empirical evidence collected from non-translated historical Chinese corpora shows that the use of 的 is predominant in modern Chinese (1368-1911), especially in the penultimate dynasty Ming (1368-1644), with a higher frequency of occurrence in general literary fiction.[1] The newly invested use of this functional particle in early Chinese scientific translation as an emerging textual genre in the late nineteenth century points to an important linguistic strategy adopted by early language workers at the time, i.e. an inventive transformation of the Chinese language attained within its existing linguistic and philological system.

[1] The non-translated Chinese historical Chinese corpus used is the Sheffield Corpus of Chinese, which is accessible online at http://www.hrionline.ac.uk/scc. Last access was on 23 November 2010.

Within the context of the current study, SCFs may be construed as prefixes or suffixes indicating the semantic or conceptual function of a translated scientific term. They play a central role in the introduction and institutionalization of Western scientific expressions in China. For example, in Lobscheid's English Chinese Dictionary (1866-9), a number of English scientific terms covering biology, physics, chemistry, law and international relations, religion, etc. were translated into compact penta-syllabic words into Chinese, which were invariably marked by the use of 者 (meaning: unspecific reference to things, agents or abstract concepts) (zhe) as the ending morpheme. In non-translated Chinese historical corpora, distribution statistics shows that the use of the SCF者 is most significant in archaic Chinese (12th AC – 220AD), with a percentage as high as 60.4%, compared to 25.1% in medieval Chinese (220-1368) and 14.5% in modern Chinese (1368-1911). In terms of textual genre distribution, the most common use of the SCF者 occurs in philosophical texts (24.9%), followed by historical texts (15.2%) and biographical essays (13.7%).

A comparison of the Chinese scientific terms coined by Lobscheid and the original expressions in English shows that the use of the SCF 者 symbolizes an initial attempt made by the German translator at transferring highly conventionalized nominalised expressions in Chinese philosophical texts to its early modern scientific register. Nominalisation is an important linguistic device in modern scientific discourse, which has been studied extensively. Nominalised terms provide a wealth of easily referable linguistic resources which may help describe complex physical process, chemical change, human behaviour and socio-cultural phenomena. The five translators, especially Lobscheid, were obviously aware of the conventionalized use of the SCF 者 and thus took full advantage of this particular linguistic feature of the target language when rendering Western scientific texts into their Chinese counterparts.

It is, however, interesting to notice that the use of functional primitive like 者 seems to point to a gradual transformation of historical Chinese at a cross-genre level within the traditional Chinese language system, rather than at a cross-linguistic and cross-cultural level as previously thought. This may be inferred from the statistics given above as the functional term 者 is now heavily exploited in early modern scientific translations, despite its predominate occurrence in Chinese philosophical texts of ancient times. Based on our original observation, it may be said that the exploration of such linguistic devices represents a tactical manipulation of the traditional Chinese linguistic system, with the intention of extending the boundaries of its existing knowledge body into unknown conceptual territories.

In an effort to attain a deeper understanding of the significance and nature of such linguistic strategies in scientific translations produced at the time, we proceed to identify several recurrent SCF categories in the pilot parallel corpus. It should be noted that due to the experimental nature of the current study, the list of SCFs given below is far from exhaustive; it however helps indicate the scale and the patterns of various highly productive functional particle types in the five scientific translations, which in turn were representative of the cross-cultural scientific exchange practiced at a much larger scale in China and Japan in the late nineteenth century. Through the online corpus mining tool, a range of functional particles falling under the SCF category were identified and summarized below:

[1] SCF_AC: abstract concepts (e.g. patterns of change or ideologies);
[2] SCF_CC: units of classification or types of measurement;
[3] SCF_CP: chemical process, physical change or ways of human behaviour (e.g. behavioural verbs);
[4] SCF_DE: devices or apparatus (including those of the human body) (concrete or abstract devices);

[5] SCF_DF: disciplinary fields;

[6] SCF_ML: research methodologies (specific or abstract) or logic;

[7] SCF_PA: practitioners, institutions of social and cultural activities and the forms of their organization (e.g. schools of thoughts, associations, learned societies, etc.);

[8] SCF_PE: physical entities, chemical substance, natural phenomena;

[9] SCF_PQ: physical qualities of things including shape, weight, size, quantity, etc.;

[10] SCF_QA: qualities of animate or inanimate things in response to external changes (natural, social, physical, etc)

[11] SCF_SCP: social, cultural and religious practice;

[12] SCF_SQ: nominalization: persons, objects, concepts or socio-cultural phenomena of special qualities;

[13] SCF_TSR: temporal or spatial reference;

[14] SCF_SS: systems of scientific symbols and basic disciplinary concepts (including medical, clinical terms, etc.)

[15] SCF_TG: textual genres: oral or written discourse; particular linguistic events or phenomena;

[16] SCF_TS: theoretical systems (论，理，法，说): to describe more established Western sciences with existing Chinese terms; these theoretical systems largely form the foundation of discrete modern scientific disciplines in China; the distribution of SCF_TS depicts the modes and patterns underlying the introduction and assimilation of different scientific fields in China through translated scientific terminology.

Diagram 3 Distribution of functional particles in the multilingual database of translated scientific terminology

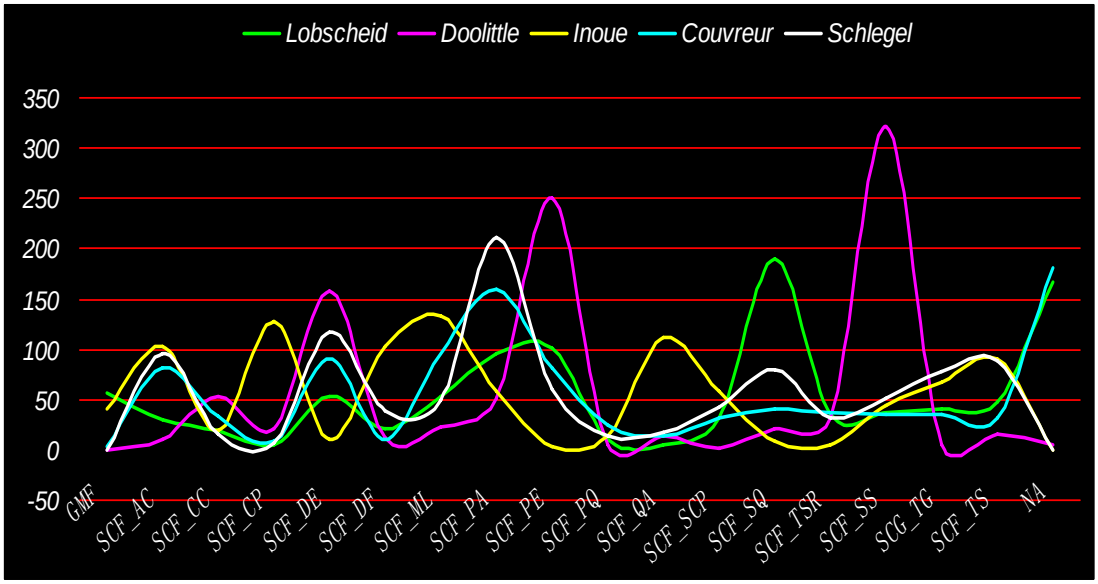


Diagram 3 shows the distribution of the various types of functional particles in the five translations. The graph depicted here turns out more complex than Diagram 1, in which the distribution of tokens of varying lengths shows striking similarities among the five texts under study. As mentioned above, Hierarchical Cluster Analysis (HCA) is especially useful when dealing with the classification of a small number of variables. It explores the underlying structure of the entire database through extracting a limited number of clusters based on their computed (dis-) similarities. At the initial stage of our corpus-based study, a variety of functional particles have been identified based on the researcher's own observation of the database. Those categories are then used to mark up the quantitative linguistic events retrieved from the five translations. It is likely that there is certain overlapping or even redundancy among the originally proposed tagging categories. To streamline the later textual analysis based on the initial hypothesis, it is essential to explore the latent structure of the experimental framework of classification. The refined

model of categorization may be then used to improve the initial theoretical hypothesis with new evidence collected from the database. In the current study, after extracting a large amount of translated scientific terms from the five translations, we proceed to tag the raw data in accord with the categorizing framework explained above. The marked-up linguistic data are then sorted out and put into a cross-tabulation shown in Table 5. Due to the internal complexity of the framework and the relatively large amount of data, it is hard to discern any underlying patterns by looking at the figures only. At this point, we again resort to the HCA to refine the initial hypothetical model.

Table 5 Agglomeration schedule of HCA of functional particles

Stage	Cluster Combined		Distance Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	10	.089	0	0	5
2	15	16	.247	0	0	6
3	2	9	.423	0	0	4
4	2	13	.689	3	0	12
5	3	5	1.032	1	0	11
6	11	15	1.634	0	2	10
7	1	6	2.452	0	0	10
8	4	8	4.119	0	0	9
9	4	14	7.429	8	0	13
10	1	11	*10.873*	7	6	11
11	1	3	*16.193*	10	5	12
12	1	2	*24.727*	11	4	15
13	4	12	*34.539*	9	0	14
14	4	7	*46.957*	13	0	15
15	1	4	*75.000*	12	14	0

Table 5 shows the agglomeration schedule of HCA of the tagged functional particles. As can be seen from the table, the decisiveness of the HCA is reflected in the increasing gaps in distance coefficients after stage 10 (italics are mine). Since the largest difference in distance coefficients occurs between stages 14 and 15, the cluster solution may be safely put at two. This has greatly simplified the initial model by conflating the originally proposed sixteen particles with semantic and cognitive functions into two major categories. The lesser differences detected within each of the two main categories may be then used to identify new subcategories of functional particles. In this way, the initial framework of classification is effectively streamlined and improved to benefit further quantitative data analysis. To have a more visualized access to the statistical information summarized in the agglomeration schedule, we proceed to plot the dendrogram of HCA, which illustrates the arrangement of clusters in a tree graph.

Diagram 4 Dendrogram of HCA of functional particles

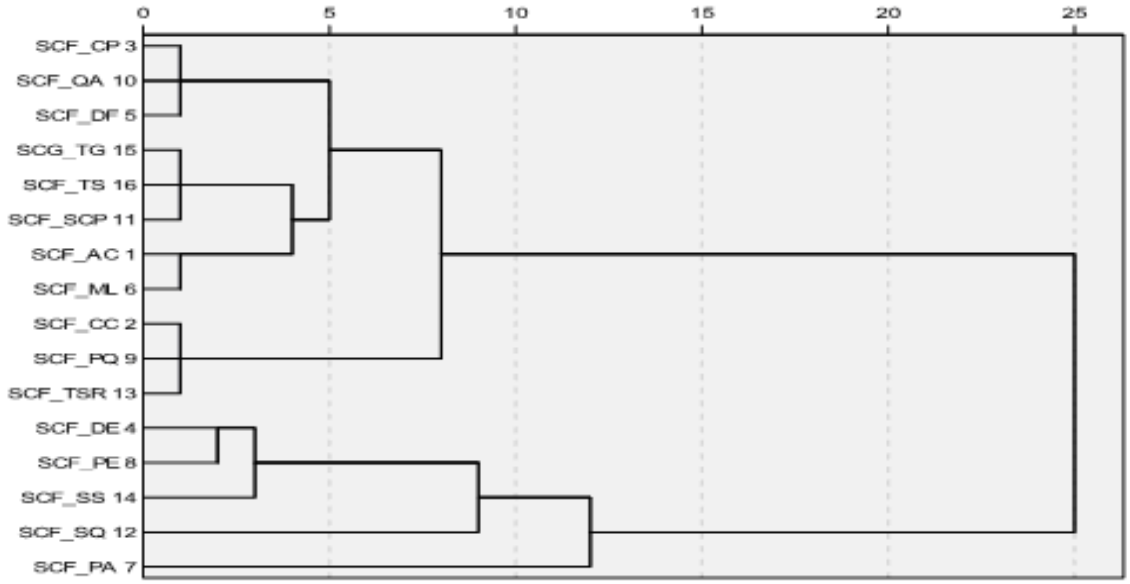


Diagram 4 displays the dendrogram of HCA of functional particles in the five translations. Along the vertical axis, each type of functional particles is marked with a numerical label, which is its case number used in the statistical processing. It is easy to find in the graph that from right to left, at the top level, the original sixteen SCF categories have been conclusively split into two clusters. The upper cluster includes eleven functional particles from SCF_CP to SCF_TSR, whereas the lower cluster is substantiated by five functional particles from SCF_DE to SCF_PA. Within each of the two main clusters, subcategories of functional particles are identified based on their computed degrees of dissimilarities. For example, in the upper cluster, at the root level in the left, four basic clusters are first extracted which are (UC1) SCF_CP/ SCF_QA/ SCF_DF, (UC2) SCF_TG/ SCF_TS/ SCF_SCP, (UC3) SCF_AC/ SCF_ML, (UC4) SCF_CC/ SCF_PQ/ SCF_TSR. Clusters UC2 and UC3 are then converged to join UC1. Their joint cluster meets UC4 to form the upper cluster. In a somehow contrastive manner, the lower cluster is formed through the expansion of the basic cluster containing SCF_DE and SCF_PE to annex SCF_SS, SCF_SQ and SCF_PA, successively.

The quantitative analysis thus conducted requires further qualitative interpretation of the refined model to have a better understanding of the underlying structure of the database. Within the upper cluster, we attempt to identify the common features of the variables forming the constituent clusters, i.e. UC1 to UC4. For example, the three functional particles forming UC1 invariably refer to the nature of things from modes of physical changes to general qualities. Similarly, the three variables substantiating UC2 may be termed as naming systems from theoretical through socio-cultural to scientific spheres. UC3 comprises SCF_AC and SCF_ML which jointly point to metaphorical concepts. Lastly, the tripartite UC4 provides useful resources for measurements and directional locations. In a similar fashion, the lower cluster as a whole may be defined as referential specificity, for the five constituent functional particles consistently refer to specific and concrete objects or agents.

The qualitative analysis provided above offers possible explanations of the rationale behind the corpus-based statistical analysis. It reveals the dichotomy of abstractness vis-à-vis specificity dividing the original sixteen SCF categories at a general level. Within each of the two main categories, a number of SCF subgroups were identified based on their shared functions in interpreting the variations detected in the dependent variables, i.e. the five early

Chinese and Japanese scientific translations. It helps elucidate the complex process in which early cross-cultural mediators attempted to develop a new scientific language through an innovative use of linguistic devices as part of historical Chinese. It forms part of the transformation of traditional Chinese writing style of scientific texts, which makes important preparations for the systematic introduction and assimilation of Western scientific concepts and ideas into the Chinese existing knowledge body at later times.

6. Conclusion

This paper offered an original investigation of translated scientific terminology in China and Japan in the late nineteenth century from a solid empirical perspective. In an effort to delve into the complex historical process of the early introduction of Western scientific ideas and concepts in China and Japan, we concentrated on the generation and statistical processing of quantitative linguistic events extracted from a large-scale database of translated terminology. The two linguistic features highlighted are token length and functional particles in newly coined scientific terms. In the study of token lengths, the HCA singled out the idiosyncratic use of idiomatic expressions by the German missionary Wilhelm Lobscheid; and in the case of functional particles, the HCA has greatly streamlined the initial framework of analysis, suggesting a refined dichotomous model of abstractness vis-à-vis concreteness which underlies the original sixteen SCF categories of functional particles. In this paper, we attempted to test the validity and usefulness of statistical procedures for historical contrastive linguistics. Last but least, due to the limited size of the database used in the current study, the findings reported here remain largely exploratory and experimental. At a later stage of our study, as the database enlarges, the newly refined model will be subjected to further experimentation to test its robustness and wider applicability. The very interesting and revealing patterns emerged in our investigation, however, would help point to potentially prolific directions for future research.

Reference

[1] Baker, M. (2004) 'The Treatment of Variation in Corpus-based Translation Studies', in Karin Aijmer and Hilde Hasselgård (eds) Translation and Corpora (Göteborg Studies in English), Göteborg University, Sweden, 7-17

[2] Couvreur, S. (1884) *Dictionnaire Français-Chinois* (contenant les expressions les plus usitées de la langue mandarine)

[3] Doolittle, J. (1872-3) *A Vocabulary and Handbook of the Chinese Language*

[4] Inoue, T. and Ariga, H. (1884) *Tetsugaku Jii* (*Dictionary of Phisophical Terms*, second edition), Tokyo

[5] Ji, M. (2010a) "A Corpus-Based Study of Linguistic Variation in Modern Chinese Scientific Writing", in *Journal of Gender Equality and Multicultural Conviviality*, no.2, Tohoku University, pp. 106-15

[6] Ji, M. (2010b) *Phraseology in Corpus-Based Translation Studies*, Oxford and Bern: Peter Lang

[7] Ji, M. (2010c) "A corpus-based study of lexical periodization in Chinese historical corpora", in *Literary and Linguistic Computing*, Oxford University Press, vol. 25, no.2, pp. 199-213

[8] Keppler, K. (1955) "Characteristics and difficulties of the German scientific vocabulary", in *The German Quarterly*, vol. 28, no.3, pp. 152-8

[9] Lackner, M. et al (eds.) (2001) *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China*, Leiden: Brill

[10] Laviosa, S. (1998) "The corpus-based approach: a new paradigm in translation studies", in *Meta*, 43, pp. 474-9

[11] Lobscheid, W. (1866-9), *English Chinese Dictionary with Punti and Mandarin Pronounciation* (1866-9)

[12] Needham, J. and Robinson, K. (2004) *Science and Civilization in China*, vol.7, no.2, Cambridge University Press

[13] Schlegel. G. (1886) *Nederlandsch-Chineesch Woordenboek met de Transcriptie der Chineesche Karakters in het Tsiang-Tsiu Dialekt* (*Dutch Chinese Dictionary with Transcript of Chinese Characters in the Jiang Su Dialect*)

[14] Shen, G. W. (2001) "The creation of technical terms in English Chinese dictionaries from the nineteenth century", in M. Lackner, et al (eds.) *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China*, Leiden: Brill, pp. 287-304

[15] Stamatatos, E. et al (2010) "Automatic text categorization in terms of genre and author", in *Computational Linguistics*, vol.26, no.4, pp. 471-96

[16] Taylor, G. and Chen, T. (1991) "Linguistic, cultural and sub-cultural issues in contrastive discourse analysis: Anglo-American and Chinese scientific texts", in *Applied Linguistics*, vol.12, no.3, pp. 319-36

[17] Tsien, T. (1954) "Western impact on China through translation", in *The Far Eastern Quarterly*, vol. 13, no.3, pp. 305-27

[18] Van Gijsel, S. et al. (2006) "Locating lexical richness: a corpus linguistic, sociovariational analysis", in the *Proceedings of JADT 2006: 8es Journées internationales d'Analyse statistique des Données Textuelles*, pp. 953-64

[19] Wright, D. (1998) "The translation of modern western science in nineteenth century China", in *ISIS*, vol. 89, pp. 653-73