GEMC
Tohoku University
ISS, University of Tokyo

A CORPUS-BASED STUDY OF LINGUISTIC VARIATION IN MODERN CHINESE SCIENTIFIC WRITING

JI Meng

■ Abstract

This paper presents a progress report on a corpus-based study of modern Chinese scientific writing in terms of the linguistic patterns characterizing the textual organization of science reportage in the country between the early 1950s and the early twentieth-first century. It aims to offer a detailed investigation of three major linguistic features, i.e. clausal conjunction, nominalization and mixed terminology, which seem to have largely informed the way modern Chinese scientific language was gradually evolved to sophistication, in spite of the pervasive political discourse influencing many aspects of scientific research in the country. At the current stage, our study is mainly focused on the design and construction of a pilot database of modern Chinese scientific texts which would then be subjected to empirical textual analysis to yield insights into the nature and manner of scientific reportage during the specific period of time under consideration.

Ⅰ. Introduction

In pre-modern Chinese culture, the study of natural sciences occupied a secondary position to traditional philology, which lay at the heart of classical culture and informed the way knowledge was built into cultural identity (Elman, 1984). The expansion of industrial capitalism in the late nineteenth century forced China to diverge from its own cultural evolutionary trajectory. As part of a broader cultural transformation, the injection of western scientific concepts into China's existing base played a major role in this process. China's contacts with western science triggered cultural conflicts centring on the way translators introduced and assimilated modern scientific terms in line with traditional cultural and philological canons (Xu, 2005). To some extent, the history of modern science in China is a textual experience in which language workers are engaged in mapping the systematic differences between China's traditional cultural agenda and that of the West through giving culturally-loaded words new dimensions of meaning or coining novel expressions with the available linguistic resources.

An analysis of the linguistic legacy left by scientific translations will permit new ways of analysing a range of socio-cultural phenomena

crucial to the history of modern science in China. Many past studies have documented the influence of individual scientists or scientific institutions (see Bowers et al, 1989; Neushul, 2000; Schneider, 2003; Amelung, 2004 for the study of individual scientists or institutions), rather than exploring how science was practiced generally in the modern context of cross-cultural and cross-linguistic interactions.

A new perspective can be offered by investigating the quantitative linguistic evidence generated by this historical process (Reardon-Anderson, 1991). The advantage of prioritizing the analysis of language as the fundamental element of cross-cultural scientific communication consists in the insights brought into the nature and manner that the target cultural system has absorbed foreign concepts and ideas. Elman (2005) provides a successful example of how socio-linguistic research can contribute to the early history of Chinese science. Instead of falling back on recounting historical events and personal anecdotes (Wang, 2007), his work goes back to the primary material – language – which serves as the essential carrier of concepts in the dissemination of modern science in China. It queries China's classical classificatory schemes for natural objects, and their incompatibility with western scientific systems, revealing their fundamentally different cosmological views of nature.

However, the linguistic evidence Elman provides is still quite limited and fragmentary. Elman's work, like many past studies in the field (Needham, 1954), also focus on the pre-modern period, rather than on the twentieth century which remains under-explored. In the current study, the sixty years under investigation, i.e. from 1949-2009 witness intensive social and cultural movements, which have greatly changed the landscape of scientific research in the country at an unprecedented level. The resilience of the language used in modern scientific practice is put to test, which has given rise to a new text genre in modern Chinese to match the scope and pace of scientific research in the country. Despite the crucial role played by language in such a historical process, there is rather limited systematic research on the various linguistic changes observed in modern Chinese science writing, which seems to suggest a lack of consciousness in the development and regulation of scientific language in the country.

This project will probe the intriguing process of cross-cultural transmission of scientific concepts, ideas and norms as reflected in the making of a new textual genre in modern Chinese, i.e. modern Chinese science writing. This will be achieved through an empirical study of textual material produced during that historical process which will be collected by following a deliberately structured framework of data sampling. It is believed that a broad selection of linguistic data embodied in actual texts will help generate theoretical hypotheses on the development of modern Chinese science writing.

II. Database

The study of the linguistic characteristics of scientific writing from a detailed sociolinguistic perspective allows an objective evaluation of the nature of scientific language used at the time as the essential carrier of meaning and thoughts. To ensure the representativeness of our pilot corpus-based study, the textual material used is from the *Chinese Science Bulletin*, which is the official journal published by the Chinese Science Academy since 1950. As one of the most

established science journals in mainland China, the *Chinese Science Bulletin* provides valuable first-hand material for investigation on the modes and patterns followed by science reportage in the country since 1949. In our pilot corpus, the texts are arranged chronologically to reflect the general trends of development of highlighted linguistic features in the corpus.

In order to optimize the research result, we selected texts from subject fields which share similar epistemological requirements such as physics, chemistry and maths. This has not been an easy task with the screening of scientific reports from political narratives which seem to populate early issues of the journal until the mid-1970s. Furthermore, as may be easily detected in the corpus, a unique feature of early scientific writing is that there is not a clear-cut line between technical reports and political discourse. Comments reflecting ideological conflicts and class struggles are often intermingled with scientific observation and assessment and largely inform the way scientific arguments are constructed in that particular socio-cultural setting. As a result of the nature of the early corpus texts, subjectivity has been inevitable in separating texts showing varying levels of technical content.

Another technical problem that prevents us from processing the texts automatically is that earlier texts are invariably written in traditional Chinese and are preserved as PDF files in the digital library. This makes the conversion of the texts from PDF to word files extremely difficult, since the scanned documents in traditional Chinese are recognized by the OCR software as pictures rather than proper texts. Hand coding highlighted textual features becomes necessary, which permits a deeper though somehow limited analysis of the corpus texts. Manual tagging is a truly laborious and time-consuming process. In order to improve the cost-effectiveness of the manual text annotation, we decide to choose the texts randomly after giving each text a numeric label. It is expected that the processing of more recent corpus texts which are machine-readable should be much easier, with natural language tools of high precision rate readily available.

Ⅲ. Theoretical framework for corpus-based textual analysis

As a pilot corpus-based study, we have chosen clausal conjunction, normalization and mixed terminology as the main subjects of the textual analysis. This is partly based on previous studies on the evolution of the language used in modern English science writing (Schnel, 2009; Taavitsainen and Pahta 2004). As will be explained in details below, the particular semantic and pragmatic functions of the three highlighted textual features may help us probe the linguistic organization, i.e. logical structure, configuration of complex natural phenomena and assimilation of imported scientific concepts and ideas, of the textual data under investigation. It is envisaged that in future research, we would incorporate more linguistic features or meaningful textual units to enrich the corpus analysis, with a view to establishing a working theoretical framework for genre analysis of modern Chinese science writing.

1. Clausal conjunctions

The study of the language of science has a long tradition in English, whereby classic texts like Darwin's *Origin of Species* or Newton's

Treatise on Opticks are often used to illustrate the particular hermeneutical style or rhetoric of modern scientific writing (Bulhof, 1992; Halliday and Martin, 1993). It is not until recent times with the fast development of computational linguistics and natural language processing that large-scale investigation on the collective behaviour of science writing as a key component of social interaction and communication among scientists has been made possible.

A new feature of the emerging field of scientometrics is an increasing emphasis on the observation and analysis of quantifiable linguistic events extracted from naturally occurring contexts (Argamon et al, 2008). To trace the evolution of modern Chinese science writing as from 1949 onwards, we focus on three specific linguistic features highlighted in the pilot corpus, i.e. clausal conjunctions, nominalization and mixed terminology to describe a developing style of science writing which may be observed at a cross-disciplinary level.

The selection of these three textual features is based on the following considerations. Firstly, the choice for clausal conjunction is derived from past studies on English linguistic variation within the systemic functional framework (Halliday and Martin, 1993; Halliday, 1994). The study of clausal conjunction, which is often used to construct the logical structure of a scientific text, provides valuable textual clues to the nature of scientific argumentation, i.e. the way scientists observe natural phenomena, evaluate scientific findings and communicate with their intended interlocutors about the research results.

Despite the potentially wide applicability of the systemic functional theory, its validity with Chinese scientific language studies has not been tested before, so far as we are aware. Our study shall therefore attempt to explore modern Chinese scientific writing from a systemic functional perspective, with a special focus on the types of clausal conjunctions which may be used to measure the varying levels of cohesiveness in the corpus texts under investigation. Cohesion holds the key to an effective writing style in modern scientific practice among other textual features. It facilitates the development of rational lines of argument in the text that may be easily followed through and commented upon by the readers.

In our case study, the typological framework employed for data classification is the one specified in Argamon et al. (2008), which is further stratified into extension, enhancement and elaboration (3E system hereafter, Table 1). It is expected that experimentation of the 3E system with chronologically arranged Chinese scientific texts in the corpus will contribute towards the construction of an exploratory conceptual apparatus, with a view to quantifying the variation and changes characterizing the evolving logical structure in modern Chinese science writing.

Table 1. 3E system of textual cohesion

Subsystem	Definition	Examples
Extension	Function words linking different arguments	and, but, furthermore …
Enhancement	Further qualify the information contained in one clause by adding extra information in subsequent clauses;	Similarly, therefore…
Elaboration	Deepens the clause by clarification and exemplification	In other words, more precisely ...

2. Nominalization

Secondly, nominalization is another core feature of modern scientific language, which has been studied in depth in several past studies (Banks, 2004; Montgomery 2006). As Halliday and Martin (1993) observe, this is a crucial linguistic strategy developed in modern English scientific writing. With English, nominalization is defined as lexical-syntactical changes which turn adjectives and verbs into noun phrases describing complex physical processes. Its main function is to provide an efficient and easily referable description of complex physical process in modern scientific practice. It is part of our hypothesis that the strong linguistic influence from modern English scientific discourse on modern Chinese scientific writing should be somehow detectable at a micro-linguistic level, i.e. through the study of a range of highly compact noun phrases in Chinese, which are somehow comparable to nominalised expressions in English.

We contend that due to the inherent differences between the two language systems, it is not always feasible to keep the morph-syntactic structure of an English expression intact when it is translated into Chinese, whereas syntactical borrowing is more common among cognate languages, such as French, Spanish, English and German. Our corpus-based study of nominalization in modern Chinese science writing has thus to be highly pragmatic, given the developing nature of the textual feature under study, as well as the considerable lexcio-grammatical differences between the two languages. To be specific, in defining a certain term as a potential nominalised expression in modern Chinese, we have to take into consideration not only the unique lexico- grammatical patterns of modern Chinese (see Table 2 and 3), but also the particular contextual circumstances which have given rise to that expression – that is to focus on the communicative function of the terms in individual texts to facilitate the configuration and description of complex scientific processes in easily-referable expressions.

Table 2. A tentative classification of nominalised expressions in Chinese science writing: lexicalized expressions

Lexico-grammatical patterns	Examples
…. 机制	统计加速机制; 突变分子机制;
….性	各向异性;
…. 化	核退化; 卵同化; 天体演化; 波函数重正化; 归一化; 催化;
…. 过程	衰变过程; 高能天体过程; 基本粒子转化过程;
….反应	热核反应;
…..变	突变; 裂变; 诱变;
…. 现象	能隙现象; 超导现象;
…. 运动	质心运动; 四粒子团运动;
….作用	同化作用; 基本粒子高能相互作用; 引力相互作用; 自作用; 光合作用
….行为	高能散射渐近行为;
….形成	化学键形成;
…. 效应	驰豫效应;

Realizing this point is especially important for a better understanding of the current state of Chinese science writing, which when compared to English, is much less mature in terms of the linguistic complexity and methodological sophistication achieved so far. An established system of modern Chinese science writing is still very much in the balance, where Chinese scientists make great efforts to articulate their thoughts and ideas with linguistic resources available to them, such as nominalization. In the absence of an official language policy to regulate the use of such imported linguistic devices, the efforts made by Chinese scientists spontaneously to improve the quality of their scientific output in writing can hardly contribute towards the establishment of a working system of science language to serve the purpose of cross-linguistic science communication. Our corpus-based study aims to make up for the insufficiency in this particular research area by investigating the various forms of nominalised expressions which have been created in modern Chinese science writing.

Another interesting phenomenon observed in the corpus is the use of quotation marks to highlight a particular physical process as though the scientist is suggesting the potential nominalization of that process characterizing the behaviour of the physical entity under investigation. In other words, the defining feature of the process may be captured and referred to by a specific term both accurately and efficiently. For example, in R. E. 艾连戈恩 M. N. 列平尼娜 (1953), the Chinese translator used a long quotation in the text to describe the distributional feature of masculine elements inside the embryo in development: "….在那个时期, 就是说在胚胎生成四个核的时期, 就发生了'在相当大的卵的原生质的区域里雄性成分的繁殖和分布'……." (p.72). The use of quotation at this point has a different pragmatic function from that of others appeared in the text: while in other places of the text, quotations are used either to give emphasis to a specific term or to imply the metaphorical or analogical function of that term being quoted, in this sentence, the long quotation

Table 3. A tentative classification of nominalised expressions in Chinese science writing – grammatical changes

Subcategory	Patterns of grammatical changes	Examples
Post-position of verbs	noun <+ function word 的> (may be omitted) + verb	…. 的产生; …. 的俘获; …. 的湮灭; …. 的减数; …. 的解体; …. 的瓦解; …. 的孤立; …. 的溃缩;
Schematic four-character expressions	Adj. + V. + Adj. + V.	偏振辐射; 广延簇射; 级联簇射;
	N. + V.	动量守恒; 绝热引入; 绝热撤出;
	N.+ N.+ V. + V.	新陈代谢;
	V.+ V.	氧化还原;

used seems to suggest that the particular feature of the phenomenon under investigation, i.e. the universal presence of a large amount of masculine elements in an embryo of Sphingidae when it is into developed four kernels, might be summarized and referred to by a subject-specific term.

3. Mixed terminology

Last but not least, we study the use of mixed terminology in modern Chinese scientific writing, which represents a fascinating area of research that lies at the heart of the modernization of Chinese science and technology (Zhang, 2007; Wu, 2008). A mixed terminology may be defined as Chinese transliterations or expressions annotated with English (or Russian, especially in the 1950s and 1960s) terms in parentheses. The significance of establishing a consistent language policy in China to regulate the use of scientific terminology suggest that the difficulties involved in the complex process of cross-cultural scientific communication are not only ideological, conceptual, but also linguistic and technical, sometimes at a rather considerable level.

As different from Japanese, which has developed the writing system of katakana to transcribe words or phrases from foreign languages, Chinese has always been composed by character words only. There have been various attempts at translating foreign scientific terms either semantically or phonetically or more often, a combination of the two, with a view to establishing an integrated linguistic system that might be used to absorb foreign scientific concepts efficiently and in time. Within such a general picture, the use of mixed terminology may be seen as a compromised solution to the technical problems implied in Chinese modern scientific writing, where the development of the language does not seem to keep pace with the increasing internal needs for a more effective linguistic system to assimilate western scientific ideas and concepts.

Table 4. A tentative classification of mixed terminology in Chinese science writing

Item	Subcategory	Examples
1	Person name	达尔文氏; M.R. 索洛维; 费米 (Femi, E.)
2	Places or laboratories named after scientists' names:	布鲁海文实验室;
3	Physical process named after scientists' name	费米型加速; 康普顿散射; 细胞融合 (cytomixis); 四粒子团运动 (quartet);
4	Hypothetical models named after scientists' name	泡利原理; 希伯脱空间; Lee 模型;
5	Specific entities named after scientists' name	契连科夫光; 格氏液 (Ringer's solution); 贝尼霍夫带; 哈密顿量 H;
6	Chemical elements	含铁氧化还原素 (FdH); 2- 氨基嘌呤 (AP);
7	Plant or animal Latin names after their popular Chinese names	半被子植物 (Hemiangiospermae); 原大 (Populus primoova)
8	New metaphors	密码 (code); 梯状式 (scalariform type); 膜壁小孔 (Pits); 子遗植物 (relics);
9	New concepts or methods	异态 (variation); 以太 (ether);

Our pilot study based on the journal articles published in the *Chinese Science Bulletin* aims to map the distributional patterns of the use of mixed terminology in modern Chinese scientific texts. This is to provide an indication of the different levels of penetration achieved by this specific linguistic strategy in modernizing Chinese science writing. As our corpus-based study will show, the use of mixed terminology in different contextual situations may well have been prompted by various communication needs. For example, a mixed terminology may be used for clarification when describing physical processes or entities less known to the audience; or for the purpose of specification when introducing new methods and concepts to the scientific community; or standardization when annotating the Chinese vernacular or vulgar names of plants or animals with their Latin names; or disambiguation when the creation of new technical terms when translated to Chinese might lead to conceptual confusion due to the particular semantic and logical structure of the target language (Table 4, here we just provide a tentative framework which is still expanding).

Ⅳ. Methodology

It is argued that the making of modern Chinese science writing can be effectively measured and examined in quantitative terms, i.e. through the annotation and computer-assisted analysis of the massive text evidence which embodies the evolution of China's modern science history. This will be achieved through a well-designed selection and incorporation of original scientific texts in modern Chinese in the pilot corpus. The quantitative data gathered will then be used for empirical sociolinguistic analysis to uncover the underlying socio-cultural factors which may have given rise to the highlighted linguistic and cultural phenomena in the corpus.

Through a well-grounded empirical sociolinguistic analysis, the project will elucidate the complex and changing historical circumstances surrounding the making of modern Chinese science as a means to demonstrate its contemporary significance in an era of growing globalization, as illustrated by global inter-dependence in terms of scientific and technological exchange and collaboration. Instead of adopting a top-down or prescriptive approach to the research question, this project will be conducted from a descriptive and highly interdisciplinary perspective drawing on research methodologies widely tested in quantitative social sciences, corpus/computational linguistics, , science history, and other methodologies as the research requires.

The project will pioneer the study of China's science history by deploying a number of statistical methods in the analysis of the corpus texts to enquire into the structure or patterns of interconnection linking the various sociolinguistic factors playing a part in constructing China's modern science writing. To be specific, this will be achieved through the development of an experimental framework of annotation which aims to highlight the various linguistic strategies devised by language workers as motivated by different socio-cultural factors and ideologies prevalent at the time.

The construction of theoretical models will be achieved by adapting and modifying methods and techniques that have been successfully tested in English sociolinguistics (Biber and Finegan, 1994; Biber et al, 1998; Cheshire, 1991; Conrad and Biber, 2001), in studying the regional variations and evolution of modern English, and the rationale

behind such a historical process. By adapting this methodology to Chinese cultural and language studies, this project will make the first step towards establishing the link between the making of modern science in China and the evolution of modern Chinese at various linguistic levels.

V. Conclusion

The project will break new ground by constructing theoretical models for Chinese science writing from a solid socio-linguistic perspective. Through the experimentation of methodologies which have been successfully tested in the study of English variation, this project can be expected to yield valuable and novel insights into the relationship between the evolution of China's modern science history recorded in massive textual evidence and the various socio-cultural and ideological factors shaping that historical process.

At the current stage, our work focuses on the documentation and digital processing of the primary textual material gathered so far. This is a laborious however highly rewarding process which will lead to the creation of a large-scale database of modern Chinese science writing. In the future research, we plan to subject the corpus to textual analyses to uncover the linguistic patterns in the texts. For example, through cluster analysis, we will try to locate the relative position of each scientific text as a vector of numeric feature values in the high-dimensional space created for the purpose of text classification. In this way, our study may help identify the textual patterns which seem to distinguish the language style characterizing the official scientific publication in the 1950s, 1960s and 1970s in China.

The findings revealed in our project will be of significant value in the study of the socio-cultural patterns underlying the scientific and economic development of China in both modern and contemporary times. Our project can be expected to make major contribution to our understanding of the historical shaping of and contemporary nature of the identity of the Chinese state in the context of conflicting notions of science, highlighting national views of scientific concepts, were eroded fundamentally by globalization processes, as illustrated by the impact of western concepts of science on China's modern scientific writing.

Acknowledgements

I would like to extend my thanks the GCOE of Gender Equality and Multicultural Conviviality, Tohoku University, for their sponsorship of this fellowship. I would also like to thank the two anonymous reviewers for their advice and suggesetions given to my paper.

References

Amelung, I. (2004) "Naming Physics: The Strife to Delineate a Field of Modern Science in Late Imperial 'China'," in Lackner, M. and Vittinghoff, N. (eds.) *Mapping Meanings: The Field of New Learning in Late Qing China*, Leiden: Brill, pp. 381-422.

Argamon, S. et al. (2008) "Language Use Reflects Scientific Methodology: A Corpus-Based Study of Peer-Reviewed Journal Articles," in *Scientometrics*, 75 (2), pp. 203-238.

Banks, D. (2004) "On the Historical Origins of Nominalized Process in Scientific Texts," in *English for Specific Purposes*, 24 (3), pp. 347-357.

Biber, D. and E. Finegan, (eds.) (1994) *Sociolinguistic Perspectives on Register*, New York: Oxford University Press.

Biber, D. et al (1998) *Corpus Linguistics: Investigating Language Structure and Use*, New York: Cambridge University Press.

Bowers, J. et al (eds.) (1989) *Science and Medicine in Twentieth-Century China: Research and Education*, Ann Arbor: University of Michigan Press.

Bulhof, I. N. (1994) *The Language of Science: a Study of the Relationship between Literature and Science in the Perspective of Hermeneutical Ontology, with a Case Study of Darwin's Origin of Species*, Leiden: Brill.

Chen, P. (1999) *Modern Chinese: History and Sociolinguistics*, Cambridge: Cambridge University Press.
Cheshire, J. (ed.) (1991) *English around the World: Sociolinguistic Perspectives*, Cambridge: Cambridge University Press.
Conrad, S. and D. Biber, (2001) *Variation in English: Multi-dimensional Studies*, Princeton: Pearson Education Limited.
Elman, B. (1984) *From Philosophy to Philology: Intellectual and Social Aspects of Change in Late Imperial China*, Cambridge: Harvard University Press.
Elman, B. (2005) *On Their Own Terms: Science in China 1550-1900*, Cambridge: Harvard University Press.
Halliday, M. A. and J. R. Martin, (1993) *Writing Science: Literacy and Discursive Power*, London: Falmer.
Halliday, M. A. (1994) *An Introduction to Functional Grammar* (Second Edition), London: Edward Arnold.
Montgomery, S. L. (2006) "Translation of Scientific and Medical Texts," in *Encyclopedia of Language and Linguistics*, Oxford: Elsevier Science, pp. 65-69.
Needham, J. (1954) *Science and Civilization in China*, New York: Cambridge University Press.
Neushul, P. (2000) "Between the Devil and the Deep Sea: C.K. Tseng, Mariculture, and the Politics of Science in Modern China," in *ISIS*, 91, pp.55-88.
Reardon-Anderson, J. (1991) *The Study of Change: Chemistry in China, 1840-1949*, Cambridge: Cambridge University Press.
Schneider, L. (2003) *Biology and Revolution in Twentieth-Century China*, Lanham, Maryland: Rowman and Littlefield Publishers.
Schnel, H. (2009) *The Evolution of the English Scientific Register*, Munich: GRIN Verlag.
Taavitsainen, I. and P. Pahta, (eds.) *Medical and Scientific Writing in Late Medieval English*, Cambridge: Cambridge University Press.
Xu, J. Z. (2005) "Brief History of Science Translation in China," in *Meta*, 50 (3), pp. 1010-1021.
Wang, Z. Y. (2007) "Science and the State in Modern China, in *ISIS,* 98, pp. 558-570
Wu, F. M. (2008) "From Ge Zhi to Science: Discussion on the Inheritance of Chinese Traditional Culture," in *Chinese Science Terminology*, No. 3, pp. 74-77.
Zhang, H. (2007) "Standardization of Terminology is a Key Issue in the Development of Science and Technology," in *Chinese Science Terminology*, No.2, pp. 18-21.