International Journal of Applied Linguistics & English Literature
ISSN 2200-3592 (Print), ISSN 2200-3452 (Online)
Vol. 2 No. 2; March 2013

Language Varieties of the Abstracts in Journal Articles Written by Chinese and American Scholars: A Contrastive Corpus Analysis using the Gramulator

Danmin Ye (Corresponding author)
School of Foreign Languages & Cultures, Nanjing Normal University
122 Ninghai Road, Nanjing 210097, China
Tel: 86-25-8359-8581 E-mail: demiyedm@gmail.com

Dongzhu Wang
School of Foreign Languages & Cultures, Nanjing Normal University
122 Ninghai Road, Nanjing 210097, China
Tel: 86-25-8359-8581 E-mail: wdzhnsd@yahoo.com.cn

Received: 14-11- 2012 Accepted: 23-12- 2012 Published: 01-03- 2013
doi:10.7575/aiac.ijalel.v.2n.2p.15 URL: http://dx.doi.org/10.7575/aiac.ijalel.v.2n.2p.15

Abstract

In this study, we identify the systematic language varieties and discourse characteristics that are indicative of the academic writings of Chinese and American scientists. We conduct a Contrastive Corpus Analysis using the computational tool, the Gramulator, to identify indicative features in Chinese science journal abstracts as compared to American science abstracts. The results suggest that the Chinese scientists tend to employ different linguistic features from their American counterparts. Specifically, Chinese science abstracts can be characterized as non-standard varieties of English by the choice of the three items: the agent, the tense, and two major types of reporting verbs. We conclude that the results may account for the interpretation of Chinese academic writings of English as non-prototypical in terms of discourse style. This study sheds light on language varieties and methodology that may be helpful to English Language Learners as well as materials developers in countries such as China.

Keywords: language varieties, science abstracts writing, contrastive corpus analysis, the Gramulator

1. Introduction

The last three decades have seen a growing number of discourse studies on written academic genres, especially research articles produced by scientists (e.g., Hyland, 2000; Swales, 1990, 2004) and by graduate students (e.g., Bunton, 2002, 2005; Dong, 1998; Ridley, 2000; Swales,2004; Thompson, 1999, 2001, 2005). Studies such as these have ranged from explorations of the discourse macrostructure (e.g., the introduction, discussion, and conclusion) to various patterns of lexical features. In the recent years, interdisciplinary methods have been applied to this area of genre analysis, especially with the development of applied natural language processing (ANLP) tools and techniques (e.g., Crossley&Louwerse, 2007; Gibbs et al, 2002; Graesser& McNamara, 2011). The writing of science abstracts, another unique section of the discourse structure, has also attracted the interests in ANLP studies (e.g. Cho, 2009; McCarthy et al, 2009; Min & McCarthy, 2012, in review) as well as scientists in different countries especially non-English speaking countries who seek to publish their research in international journals. As a large portion of these non-native-English-speaking-researchers, Chinese scientists find it frustrating to have their research articles resubmitted or rejected for language problems in this specialized area of science journals (Yu & Liang, 2006). Of course, there are many books that are helpful with basic academic writing in a second or foreign language (e.g., Tang, 2012). But for the level of scientific journal texts, as McCarthy and colleagues (2009) demonstrate, relatively little research has compared the texts of non-native-English-speaking scientists (or *Outsiders* as they are referred to in Min & McCarthy) to those written by their native-English-speaking counterparts (or *Insiders* as they are referred to in Min & McCarthy) so as to identify linguistic varieties using computational tools. This issue is of importance because the degree to which an English-language text differs from an expected model (e.g. American English) may negatively affect the chances of the non-native English speakers having their manuscripts accepted (Flowerdew, 2001; Hewings, 2006).

To facilitate the production of Outsiders with the issue of academic English writing, McCarthy et al. (2009) analyzed English articles written by Japan, Britain, and American scientists. Their study provided evidence that Japanese authors used significantly more locational and temporal items, high frequency words, high familiarity words, cardinal numbers and higher syntactical similarity in sentence structures as compared to their American counterparts. Building on this study, Duncan and Hall (2009) analyzed the journal articles of three groups of scientists: Americans; Koreans publishing articles in Korea; and Koreans publishing articles in America. They found that the journal articles of Koreans publishing-in-Korea were the most distinct, and therefore, the least prototypical as compared to the other two groups. Recently Min and McCarthy (in review) compared the journal texts of Korean scientists and American scientists. Their findings suggested that American scientists preferred personal pronouns, present tense, and active voice in the use of verbs of reporting, as compared to the Korean scientists. Korean scientists share commonalities on the choice of words with Americans but they differed in terms of how they presented the words. Specifically, they preferred fewer personal pronouns, the past tense, and passive voice in the use of verbs of reporting when compared to their American counterparts. Building on this research, the current study further investigates the issue of English writing in journal articles by assessing the English science abstracts written by Chinese scientists. The purpose of our study is to discover and assess language varieties used in science abstract writings of Chinese scientists as compared to American scientists. In this study, we seek to identify the variation of linguistic features within the text (i.e., the linguistic choices in terms of words and groups of words). Through such an approach, we aim to address the following primary research questions:

Question one: Do the findings of this study support Min and McCarthy (2012, in review) results concerning Korean scientists?

Question two: Do Chinese scholars employ distinct language varieties in academic science abstracts writings in comparison to a prototypical model from American scientists?

If so, how different do they use these non-standard language varieties compared with their American counterparts?

Question three: Do Chinese and American scientists have different preference of linguistic features while writing journal abstracts?

Hypothesis one: The findings of this study support Min and McCarthy (2012, in review) results concerning Korean scientists.

Hypothesis two: Chinese scholars employ distinct language varieties in comparison to a prototypical model from American scholars in their academic science abstracts writings.

Hypothesis three: Chinese and American scholars have different preference of linguistic features while writing journal abstracts.

2. Methods of the present study

2.1 Contrastive Corpus Analysis

The origin of Contrastive Corpus Analysis (Cobb, 2003; Granger, 1998) can be dated back to the Brown corpus (Kučera& Nelson, 1967) in that its first collection of texts (500) enabled numerous studies, among which the most famous presumably is Biber’s (1989), to understand text types as much by where they overlapped as where they did not. The principle of CCA is that any discourse unit (e.g., text-type, register, genre, variety, or section of text) is best understood, and perhaps only understandable, within the context of its contrast to some other discourse unit (McCarthy, Watanabe, & Lamkin, 2012). CCA differs from traditional corpus analyses because it emphasizes on what two (or more) correlative corpora can reveal when their commonalities are excluded by computational and statistical techniques. In the field of second language learning (SLL), Cobb describes CCA as the comparison of two corpora through which what is present and what is not present can be derived. Thus, in two corpora that are highly related but differ minimally (e.g. scientific writing in English by Chinese scientists vs. scientific writing in English by American scientists), the linguistic features that are characteristic of one corpus, but non-characteristic of the sister corpus, is what is *indicative* of the text type.

2.2 The Corpus

Our corpus comprises 672 abstracts taken from 31 science journals published in either China or the United States respectively. The contents cover three genres: the so called “hard sciences” of biology, chemistry, and physics. These journals, all with high impact factors rank in the top five in each area of the three subjects. And all of the articles are published in the last five years (i.e., from 2007 to 2012). In addition, the journals are compiled as parallel as possible to ensure the comparability. For example, we have *Chinese journal of Inorganic Chemistry* in the Chinese corpus in parallel with *Inorganic Chemistry* in the American corpus. From these texts, two individual sub-corpora were compiled: (1) Chinese scientists in China (CSC) and (2) American scientists in America (ASA). The Chinese English corpus comprises Chinese scientists’ abstracts (n = 335), published exclusively in 15 different Chinese journals. The American English corpus (the assumed prototypical model) comprises U.S. scientists’ abstracts (n = 337), published exclusively in 16 U.S. journals.

To ensure the original nationalities of the authors, the model of McCarthy and colleagues (2009) was followed (see also Duncan & Hall, 2009 and Min and McCarthy 2012, in review). This model has two major criteria: (1) the first author (generally the person who writes most of the paper or leads the projects in the field of science) and the last author (generally the supervisor) should be from universities or institutes within the same country (e.g. in this study, the first and the last authors of the Chinese English and the American English corpora should be from the Chinese and the American universities or institutes respectively). (2)The names of the primary and final authors must be ‘typical’ of the country of the classification. That is, the primary and final authors in the Chinese and American corpora represent the typical names for Chinese and Americans respectively. Of course, this model cannot always ensure the authenticity of the authors’ nationalities, but these criteria of classification are effective in determining the language backgrounds of the writers. For the classification of Chinese authors, it is not hard to do the task because the first author of this study is Chinese and the names are always written in Chinese characters in Chinese science journals. For the classification of the American authors, we ensure that both the first and the last authors are working in American universities or institutes, which means that they are American-based authors.

Following Min and McCarthy (2012, in review), the current study focuses on the abstracts of journal articles written by Chinese and American scientists to find the distinctive features of each corpus. As a unique section of the discourse structure, science abstracts are the first reviewed and most frequently read part of the journal articles. They are also representative of the entire research, always available on internet, and easy to collect. These points make science abstracts a reasonable point of departure for the current study.

	Chinese corpus	American corpus
Journals	15	16
Abstracts	335	337
Years	2007-2012	2007-2012
Content	biology, chemistry, physics	biology, chemistry, physics

2.3 The Gramulator

In this study, we analyzed the Chinese and American corpora using the natural language processing tool, the Gramulator (McCarthy, Watanabe, & Lamkin, 2012). The Gramulator is a qualitative and quantitative computational textual analysis tool, freely available on internet. It is designed to identify differential linguistic features of correlative sister corpora. The tool's primary unit of analysis is the *n-gram*: adjacently positioned lexical items in a text. In this study, we focus on two-word *n-grams* (or, *bigrams*); and more exactly, on *differentials*, the lexical features that are most commonly occurring to one corpus (i.e., among the 50% most frequent bigrams), but are *un*common to the contrasting corpus (i.e., *not* among the 50% most frequent bigrams). By identifying differentials, we reveal the language varieties in the specific contexts that are most characteristic to the Chinese corpus but are least characteristic in the American corpus and explore further the reason why they are present in the discourse of abstracts. In our Gramulator analysis, we write the differentials of the two corpora as *Chinese (American)* and *American (Chinese)* differentials respectively.

In the current study, we followed Min and McCarthy's research and focus on the two Korean differentials they found: *was/were not*, and *verbs of reporting*. The analysis is mainly about the *differentials* in two forms: the *bigram* (i.e. the actual two words) and the *flexi-gram* (i.e. the underlying or theoretical form of the bigram, for example the flexi-gram for *his dog* and *his cat* is *his pet*). We verify these differentials using the Chinese corpus to assess whether Korean scientists' preferences generalize to Chinese scientists in terms of choice of linguistic features.

We employed the method of Contrastive Corpus Analysis in this study. Here are two samples of the science abstracts from the Chinese and American corpus respectively:

A patch of open coal mining land was reclaimed for ecological rehabitation 17 years ago in Antaibao, Shanxi Province, China. Under the rehabilitated Robinia pseudoacacia + Pinus tabuliformis mixed forest, herbacious plants in 320 quadrats(1 m 1 m) in the 0.8 hm2 plot were surveyed for the species composition, spatial patterns and other community properties. The results showed that the land was rich in herbaceous species, containing 44 species, which belonged to 30 genera under 16 families. The dominant families were Poaceae and Asteraceae, and the dominant species included Artemisia annua, Elymus dahuricus and Artemisia sieversiana. The initially planted Bromus inermis deteriorated badly. The important value, species abundance and frequencies were different among the families or species. The dominant families and species were commonly distributed, and the spatial patterns were obviously spatial heterogeneous. (Chinese_biology_073)

Mussels have a remarkable ability to attach their holdfast, or byssus, opportunistically to a variety of substrata that are wet, saline, corroded, and/or fouled by biofilms. Mytilus edulis foot protein-5 (Mefp-5) is one of several proteins in the byssal adhesive plaque of the mussel M. edulis. The high content of 3,4-dihydroxyphenylalanine (Dopa) (30 mol %) and its localization near the plaque Csubstrate interface have often prompted speculation that Mefp-5 plays a key role in adhesion. Using the surface forces apparatus, we show that on mica surfaces Mefp-5 achieves an adhesion energy approaching Ead = 14 mJ/m2. This exceeds the adhesion energy of another interfacial protein, Mefp-3, by a factor of 4C5 and is greater than the adhesion between highly oriented monolayers of biotin and streptavidin. The adhesion to mica is notable for its dependence on Dopa, which is most stable under reducing conditions and acidic pH. Mefp-5 also exhibits strong protein Cprotein interactions with itself as well as with Mefp-3 from M. edulis. (American_biology_062)

3. Results

Analysis 1: Comparison to Min and McCarthy Results

Highest ranking flexigram preference

In Min and McCarthy (2012, in review), the most frequently used flexigram for the Korean scientists was *was/were not*. More specifically the Koreans appeared to prefer *was/were + not* as compared to their American counterparts' preference for *am/is/are + not*. When we apply this Korean flexigram to the Chinese and American corpora from the current study, we find no significant difference in frequency of use (Chinese: 7 instances across 7 files out of 335 abstracts, 2.09%; American: 9 instances across 8 files out 337 abstracts, 2.37%; p = 1.000). Converting to the present tense (as preferred by the American scientists), the flexigram *am/is/are + not* again failed to reach a level of significance (Chinese: 15 instances across 14 files out of 335 abstracts, 4.18%; American: 18 instances across 17 files out of 337 abstracts, 5.05%; p = .713). Taken together, the results do not provide evidence that the primary findings of the Korean/American scientists' linguistic preferences generalize to Chinese scientists.

Verbs of reporting

The second flexigram of note from Min and McCarthy (2012, in review) was entitled *verbs of reporting*. The flexigram included various forms of the lemmas for seven verbs (e.g. *show*, *demonstrate, report, find, identify, use* and *present)*. Following Min and McCarthy's findings, we searched for the verbs of reporting in the differentials of the Chinese and American corpora. We found 5 verbs of reporting, *show*, *demonstrate*, *present*, *find*, and *report* in the American differentials and took them as the prototypical words for reporting findings. To assess whether Min and McCarthy's finding generalizes to Chinese scientists, we searched for these verbs (in any form) across the array of *typicals* (i.e. the above average frequency occurring *n-grams)* from both the Chinese and American corpus. We use *typicals* instead of *differentials* because the former contains lexical features regardless of any co-occurrence across corpora while the latter, by definition, can only appear in lexical features from one corpus. We found that the highest ranked verb of reporting occurring in both of the corpora is the lemma *show*, and the second highest ranked example is the lemma *find* (see Table 1). For the other three reporting verbs, Chinese used none of them in any form, but American scientists used them diversely. Our findings suggest that Chinese scientists may be simplifying their choice of verbs of reporting by (over)using the word *show* and *find* as compared to the diverse lexicon of the presumably prototypical American scientists.

Table 1. Rank, percentile and z-score of the verbs of reporting in Chinese and American *typicals*

verbs of reporting	**Chinese**			**American**		
	rank	**percentile**	**z-score**	**rank**	**percentile**	**z-score**
show	19	92	0.716	31	89	0.687
shows	69	71	0.133	/	/	/
showed	10	96	2.108	/	/	/
shown	204	13	0.008	52	81	0.272
demonstrate	/	/	/	67	76	0.177
demonstrates	/	/	/	/	/	/
demonstrated	/	/	/	134	52	0.043
present	/	/	/	96	65	0.083
presents	/	/	/	/	/	/
presented	/	/	/	237	14	0.005
find	/	/	/	100	64	0.079
finds	/	/	/	/	/	/
found	31	87	0.431	54	81	0.26
report	/	/	/	116	58	0.061
reports	/	/	/	/	/	/
reported	/	/	/	/	/	/

Analysis 2: Comparison of American (Chinese) and Chinese (American) Differentials

To more broadly assess the differentials, we follow the common practice of corpus data investigation (McCarthy &Boonthum-Denecke, 2012; McNamara, Graesser, McCarthy, &Cai, in press; Witten & Frank, 2005). Specifically, we randomly divided the texts of each corpus into two-thirds training-set data (Chinese: 224 texts; American: 226 texts) and one-third test-set data (the remaining 111 texts for each corpus). This investigation model of analysis allows us to run preliminary findings using the training-set data and confirmatory analysis using the test-set data. Such an approach helps guard against type 1 errors.

American Differentials

The Gramulator analysis on the American training-set data produced 134 differentials bigrams. Of these differentials, the highest ranked example is *we have*. The Americans employ this differential bigram for32 instances across 25 of 226 texts whereas the Chinese employ it for only 7 instances across 5 of 224 texts (American:11.60%, Chinese: 2.23%; $p<$ .001) Taking a closer look at the context, we find that the bigram *we have* is most often employed by Americans to show *the conclusion* of respective studies in their abstracts. More specifically, they tend to use the active voice and present perfect form to report their conclusion of research (e.g. *we have carried out, we have used, we have studied, we have measured,* and *we have developed*). By contrast, we find that the 4th ranked of the 105 differentials from the Chinese training-set data is *by using* (e.g. *by using this method*, *by using a specific model*).The Chinese employ it for 23 times across 19 files out of 224 texts while the Americans employ only 6 times across 6 files out of 226 texts (Americans:

2.66%; Chinese: 8.48%; p = .007). The results indicate that the Chinese appear to use the bigram *by using* to show *the method* of research but Americans employ the active and present perfect *we have* to show *the conclusion* in their abstracts writings. To verify this finding, we used the test-set data to compare the bi-grams and found that for *we have*, Americans use it for 4 times across 4 files out of 111 texts while Chinese employ 0 instance out of 111 texts (American: 3.60%; Chinese: 0%; p = .122). But for *by using*, Americans employ only 2 instances across 2 files out of 111 texts while Chinese use it for 12 times across 11 files out of 111 texts (American:1.80%; Chinese: 9.91%; p = .019). We then reran the bigrams on the entire corpus and found that Americans employ *we have* more frequently than their Chinese counterparts (American: 8.61%; Chinese: 1.49%; p < .001) whereas Chinese scientists employ *by using* much more frequently than Americans (American: 2.37%; Chinese: 8.96%; p < .001). Taken together, the results suggest that Americans prefer the present perfect structure of *we have* to show *the conclusion* as opposed to Chinese's preference for *by using* to show *the method* of their research in abstracts.

Returning to the results of the training-set data, our analysis suggests that there is a systematic difference between American and Chinese scientists on how to use linguistic features to report their findings. Americans appear to make a greater use of the flexigram *we* + [*verbs of reporting*] as compared to their Chinese counterparts. That is, we found 5 verbs of reporting with high ranks among the 134 differentials in the American training-set, with all of them combined with the agent pronoun *we*. For instance, the bigram *we show* ranked highest (11th), with *we present* (20^{th}), *we report* (23^{rd}) *we demonstrate* (31^{st}) ranking similarly and *we find* (83^{rd}) ranking relatively low. American scientists employ *we show* much more frequently in 18 instances across 16 files out of 226 texts while their Chinese counterparts use it for only 2 times across 2 files out of 224 texts (American:7.08%; Chinese:0.89%; p < .001). And *we present* is used by Americans for 14 times in14 files out of 226, but only 1 time in 1 file out of 224 by Chinese (American:6.20%; Chinese:0.45%; p = .001). The same result is found in *we report* (American: 5.75%; Chinese: 2.23%; p = .090), *we demonstrate* (American: 4.87%; Chinese: 0%; p = .001). The frequency of *we find* is in the direction of Americans, but is not significantly different (American: 3.54%; Chinese: 1.79%; p = .381).

The 5 verbs of reporting described above can be combined into the flexigram *we* + [*verbs of reporting*]. Considered as a single instantiation, test-set analysis suggests that the difference in frequency between the sister corpora is significant: Americans employ 42 instances across 31 files and Chinese employ 13 instances across 8 files out of 111 texts (American: 27.92%;Chinese:7.21%; p< .001).

Tabel 2. The flexigram of *we* + [*verbs of reporting*] in American and Chinese training-set data

we +[*verbs of reporting*]	American	Chinese	*p*-value
we show	16	2	<.001*
we demonstrate	11	0	.001*
we present	14	1	.001*
we report	13	5	.090*
we find	8	4	.381

*Note: * significant at .05.*

The result of flexigram *we* + [*verbs of reporting*] called for further analysis as to whether Americans' preference of *we*+ [*verbs of reporting*] is consistent when the verbs are in the past tense. The results also prompted further investigation as to whether the American scientists prefer the present tense of these verbs of reporting in comparison with the Chinese scientists. That is, do the two groups of scientists differently employ the verbs of reporting in terms of tense? We reran the training-set and found that the Chinese scientists appear to prefer the bigram *we found* to their American counterparts (Chinese: 3.125%; American: 0.442%; p = .037).

Tabel 3. The flexigram of *we* + [*verbs of reporting*] in past tense in American and Chinese training-set data

we + [*verbs of reporting*]	American	Chinese	*p*-value
we showed	0	2	.247
we demonstrated	0	0	1.000
we presented	0	0	1.000
we reported	0	1	.498
we found	2	7	.037*

*Note: * significant at .05.*

Taking all the results together, we can conclude that Americans prefer the present perfect structure of *we have* to show *the conclusion* as opposed to Chinese's preference for *by using* to show *the method* of their research in abstracts writing. In addition, American and Chinese scientists tend to employ different choice of verbs to report their findings: Americans prefer the flexigram of *we*+ [*verbs of reporting*] as compared to their Chinese counterparts, who appear to prefer less pronoun *we* with the verbs of reporting in the present tense but they prefer the bigram *we found*.

Chinese Differentials

The analysis on American differentials also called for further studies on how Chinese scientists are functionally performing the act of reporting in science journals. That is, what is the preferred flexigram employed by the Chinese scientists as compared to the American scientists' preference of *we* + [*verbs of reporting*]? We will address this issue in detail in this section, specifically assessing the Chinese scientists' preference for the agent *result/s*.

Of the 105 differentials produced using the training set data, the most frequently employed Chinese differential bigram is *showed that*, which is employed by Chinese for 44 instances across 36 files, while it is employed by the Americans for only 1 instance across 1 file (Chinese: 16.07%; American: 0.44%; $p < .001$). The second most frequently used bigram is *results show*. It is employed in 23 instances across 23 files by the Chinese scientists but only 1 instance across 1 file by the American scientists (Chinese: 10.27%; American: 0.44%; $p < .001$). A similar differential, *results showed*, the 6th most frequently employed bigram, was found in 20 instances across 19 files in the Chinese training-set data, while 0 instance in the American texts (Chinese: 8.48%; American: 0%; $p < .001$). Combining this differential (*results showed*) with the most common differential (*showed that*) gives us the trigram *results showed that*. This trigram features in 18 instances across 18 files of the Chinese training-set data , but does not feature at all in American training-set data (Chinese: 8.04%; American: 0%; $p < .001$).

In American (Chinese) differentials, we found that Americans appear to make greater use of the flexigram *we* + [*verbs of reporting*]. Based on this result, we predicted that the Chinese scientists may tend to forego the pronoun *we* and instead use the flexigram *result/s* + [*verbs of reporting*]. That is, we predicted that the Chinese and American scientists may prefer different agents: the inanimate *result/s* for the Chinese and animate *we* pronoun for the Americans. Additionally, we broaden the analysis of the 5 verbs of reporting in different tense (the present and the past tense). We conducted a series of Fisher's Exact Tests to assess the differences between the Chinese and American data sets. The results suggest that Chinese scientists tend to avoid using the animate agent pronoun *we* with the verbs of reporting in the present tense. The negative percentage difference for Chinese indicates that they employ much less agent *we* with verbs of reporting in the present tense compared with their American counterparts. The difference is significant for the verbs of *show* (Chinese: 0.893%; American: 7.08%; $p = .001$)*, demonstrate* (Chinese: 0%; American: 4.867%; $p = .001$)*, and present* (Chinese: 0.446%; American: 6.195%; $p = .001$). However, we find that the Chinese employ the agent *we* for verbs of reporting in the past tense, especially in the bigram *we found* (Chinese: 3.125%; American:0.442%; $p = .037$).Turning to the agent of *result(s)*, the Chinese demonstrate a preference for employing the verb of *show* in both present and past tense (see Table 4).

Table 4. Chinese-American percentage difference of employing the 5 verbs of reporting in the training-set data

show		show	shows	showed
	We	-6.187*	/	0.893
	[Result(s)]	9.826*	0.893	8.482*
demonstrate		demonstrate	demonstrates	demonstrated
	We	-4.867*	/	0
	[Result(s)]	-1.319	0.893	0.892
present		present	presents	Presented
	We	-5.749*	/	0
	[Result(s)]	0	0	-0.885
find		find	finds	found
	We	-1.754	/	2.683*
	[Result(s)]	0	0	0
report		report	reports	reported
	We	-3.52	/	0.446
	[Result(s)]	0	0	0

*Note: * significant at .05.*

Based on these results, we made a further hypothesis that the three items (i.e., the agent, the different types of verbs of reporting, and the tense), which we can label as a *register phenotype* may account for the interpretation of Chinese academic writings of English as non-prototypical in terms of discourse style. That is, in addition to the agent, the different types of reporting verbs and the tense are also characteristic of the linguistic features of the Chinese and American texts. Specifically, we have two major types of reporting verbs: Free Verbs of Reporting (i.e. verbs with the agent of either pronoun or results) *show* and *demonstrate*, and Restricted Verbs of Reporting (i.e. verbs with the agent of pronoun only) *present, find* and *report*. The different choice on the tense of these verbs may also lead to the different manifestation of language varieties.

From the training-set data (see Table 5), we find that for Free Verbs of Reporting (or FVR, i.e. *show* and *demonstrate*), Chinese scientists employ the flexigram of *we* + [*FVR in present tense*] significantly less frequently than Americans (Chinese: 2 instances in 2 files out of 224 texts, 0.89%; Americans: 30 times in 24 files out of 226 texts, 10.62%; *p*

<.001). But for FVR with the agent of *result/s*, the Chinese scientists show their preference in employing both the flexigram *result/s* + [*FVR in present tense*] (Chinese: 25 times in 25 files out of 224 texts, 11.16%; American: 6 times in 6 files out of 226, 2.66%; $p < .001$) and *result/s* + [*FVR in past tense*] (Chinese: 19 files out of 224, 8.48% ; American: 0 file out of 226, 0%; $p < .001$). Meanwhile, for the Restricted Verbs of Reporting type (or RVR, i.e. *present*, *find* and *report*), the Chinese scientists demonstrate the preference of using the flexigram of *we* + [*RVR in past tense*] (Chinese: 8 files out of 224 texts,3.57% ; American:1 file out of 226 texts, 0.44%; $p = .020$) while their American counterparts employ the flexigram of *we* + [*RVR in present tense*] (Chinese: 9 files out of 224 texts, 4.02%; American: 31 files out of 226 texts, 13.72%; $p < .001$).

Table 5. The use of three items, the agent, two different types of reporting verbs, and the tense, in Chinese and American training-set texts

Agent	Verbs of Reporting	Tense	Chinese	American	*p*-value
We	FVR (show,demonstrate)	present	2	24	<.001*
		past	2	0	.247
	RVR (present, find and report)	present	9	31	<.001*
		past	8	1	.020*
Result/s	FVR (show, demonstrate)	present	25	6	<.001*
		past	19	0	<.001*
	RVR (present, find and report)	present	0	0	1.000
		past	0	2	.499

*Note: * significant at .05.*

Building on these results, we further predicted that for Free Verbs of Reporting (FVR), Chinese scientists prefer *result/s* while American scientists prefer the pronoun *we*. However, for Restricted Verbs of Reporting (RVR), Chinese prefer the past tense, Americans prefer the present tense.

To test our hypothesis, we conducted a Fisher's Exact Test using the test-set data to assess the frequencies between the Chinese-English and American-English corpora (see Table 6). We find that for the FVR, the difference in frequency is significant: Chinese prefer the agent of *result/s* and both the present and past tense (for FVR in the present tense, Chinese:11 files out of 111 texts, 9.91%; American: 2 out of 111, 1.80%; $p = .019$; for FVR in the past tense, Chinese:11 files out of 111 texts, 9.91% ; American:1 out of 111, 0.90%; $p = .005$), while Americans prefer the agent of pronoun *we* and the present tense(American:16 files out of 111 texts, 14.41%; Chinese : 4 out of 111, 3.60%; $p = .008$). However, for the RVR, Americans prefer the present tense (American: 21 files, 18.92%; Chinese: 6 files, 5.41%; $p = .003$), while both Chinese and American scientists show low frequency and no significance in the past tense (Chinese: 4 files, 0.90%; American: 1 file, 3.60%; $p = .369$).

Table 6. The use of three items, the agent, two different types of reporting verbs, and the tense, in Chinese and American test-set texts

Agent	Verbs of Reporting	Tense	Chinese	American	*p*-value
We	FVR	present	4	16	.008*
		past	0	1	1.000
	RVR	present	6	21	.003*
		past	4	1	.369
Result/s	FVR	present	11	2	.019*
		past	11	1	.005*
	RVR	present	0	0	1.000
		past	0	0	1.000

*Note: * significant at .05.*

We reran this assessment on the entire corpus to verify the results (see Table 7). For the FVR, the difference in frequency is significant: Chinese prefer the agent of *result/s* and both the present and past tense (for the FRV in the present tense, Chinese:36 files out of 335 texts, 10.75% ; American: 8 out of 337, 2.37%, $p < .001$; for the FRV in the past tense, Chinese :30 files out of 335 texts, 8.96% ; American:1 out of 337, 0.30%, $p < .001$), while Americans prefer the agent of pronoun *we* and the present tense (American:40 files out of 337 texts, 11.87%; Chinese :6 out of 335, 1.79%; $p < .001$). However, for the RVR, Chinese prefer the past tense (Chinese: 12 files, 3.58%; American: 2 files, 0.59%; $p = .007$), while Americans prefer the present tense (American: 52 files, 15.43%; Chinese: 15 files, 4.48%; $p < .001$).

Table 7. The use of three items, the agent, two different types of reporting verbs, and the tense, in the entire Chinese and American corpora

Agent	Verbs of Reporting	Tense	Chinese	American	*p*-value
We	FVR	present	6	40	<.001*
		past	2	1	.623
	RVR	present	15	52	<.001*
		past	12	2	.007*
Result/s	FVR	present	36	8	<.001*
		past	30	1	<.001*
	RVR	present	0	0	1.000
		past	0	2	.499

*Note: * significant at .05.*

Taking the above results together, our analysis suggests that the Chinese scientists employ similar but perhaps oversimplified choice of verbs to report their findings. They employ the flexigram of *result/s + verbs of reporting* in their abstracts writing, and the register phenotype (the agent, the tense, and the two major types of verbs of reporting) characterizing the academic writing of Chinese scientists as non-standard varieties of English compared with their American counterparts. Specifically, for verbs of reporting with the agent of either pronoun or results (FVR), Chinese scientists prefer results, while American scientists prefer the pronoun *we*. However, for verbs of reporting with the agent of pronoun only (RVR), Chinese appear to prefer the past tense while Americans prefer the present tense.

4. Discussion

In this study, we assess whether Chinese scientists employ distinct language varieties in academic science abstracts writings in comparison to a prototypical model from American scientists. By using the computational tool, the Gramulator, we discuss and assess the language varieties in a full length. Collectively our results suggest that the Chinese scientists tend to use different linguistic features for the register phenotype of the agent, the tense, and two different types of reporting verbs.

This study addressed the three primary research questions: 1) *Do the findings of this study support Min and McCarthy (2012, in review) results of Korean scientists?*2) *Do Chinese scientists employ distinct language varieties in academic science abstracts writings in comparison to a prototypical model from American scientists? If so, how different do they use these non-standard language varieties compared with their American counterparts?* 3) *Do Chinese and American scientists have different preference of linguistic features while writing journal abstracts?*

Addressing the first question, our response is that this study supports Min and McCarthy's findings of Korean scientists (2010, in review) in that they both employ non-standard varieties of English in academic writing of science abstracts. But our findings differ from theirs for the register phenotype (i.e. the three items of the agent, the tense, and two major types of verbs of reporting) Chinese and American scientists tend to employ in their science abstract writings.

To answer our second question, our response is that Chinese scientists appear to employ distinctive linguistic features in academic science abstracts writings which characterize them as a non-prototypical language variety as compared to their American counterparts. Specifically, Chinese tend to use similar but oversimplified choice of verbs to report their findings and different bigrams to show the method of their research as opposed to their American counterparts.

Addressing the last question, we find that Americans prefer the present perfect tense structure of *we have* to show the method of their studies as compared to Chinese's preference for *by using*.

Moreover, Chinese scientists employ the flexigram of *result/s + verbs of reporting* in their abstracts writing in comparison to the flexigram of *we + verbs of reporting* used by their American counterparts. And the *register phenotype* (i.e. the agent, the tense, and the two major types of verbs of reporting) characterizes their academic abstracts writing as non-standard varieties of English. Specifically, for verbs of reporting with the agent of either pronoun or results (FVR), Chinese scientists prefer results, while American scientists prefer the pronoun. For verbs of reporting with the agent of pronoun only (RVR), however, Chinese prefer the past tense, Americans prefer the present tense.

Although our study provided these findings, future research needs to be done on the breadth of prototypical and non-prototypical varieties. For instance, how well do these findings generalize to different language learners other than Chinese (e.g. Japanese, Indians, etc.)? And how well do these findings generalize to the different academic areas of journal articles (e.g., geography articles, computer science articles etc.). Furthermore, how well do these results generalize to different sections of research articles (e.g., the introduction section, the discussion section etc.)? Future experiment also needs to be conducted to assess whether changes made to journal articles as to the *register phenotype* of the agent, verbs of reporting, and tense, has a positive effect on reviewers and the subsequent success of publication.

Acknowledgements

This research was supported by the Project of Jiangsu Graduate Students Scientific Research and Innovation (CXZZ11_0864), and by the Priority Academic Program Development of Jiangsu Higher Education Institutions (20110101). The authors also acknowledge Hyunsoon C. Min and Dr. Philip M. McCarthy at the Institute for Intelligent Systems, University of Memphis, United States of America, for their help during the writing process of this paper.

References

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Bizzell, P. (1992) *Academic discourse and critical consciousness.* Pittsburgh: University of Pittsburgh Press.

Bunton, D. (2002). Generic moves in PhD thesis introductions. In J. Flowerdew (Ed.), *Academic discourse* (pp. 57-75). Harlow: Pearson.

Bunton, D. (2005). The structure of PhD conclusion chapters. *Journal of English for Academic Purposes*, 4, 207-224.

Cho, D.W. (2009). Science journal paper writing in an EFL context: The case of Korea. *English for Specific Purposes*, *4*(28), 230-239.

Cobb, T. 2003. Analyzing Late Interlanguage with Learner Corpora: Québec Replications of Three European Studies. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes* 59(3): 393-423

Conrad, S. (1996). Investigating academic texts with corpus-based techniques: an example from biology. *Linguistics and Education*, *8*, 299-326.

Crossley, S.A., Greenfield, J., & McNamara, D.S. (2008). Assessing text readability using psycholinguistic indices. TESOL Quarterly*, 42*, 475-493.

Crossley, S.A. & Louwerse, M. et al. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal, 91,* 15-30.

Dong, Y. R. (1998). Non-native speaker graduate students' thesis/dissertation writing in science: Self-reports by students and their advisors from two US institutions. English for Specific Purposes, 17, 369-390.

Duncan, B., & Hall, C. (2009). A coh-metrix analysis of variation among biomedical abstracts. Proceedings of *the twenty second International Florida Artificial Intelligence Research Society Conference* (pp. 237-242). Menlo Park, California: The AAAI Press.

Ferris, D., & Hedgcock, J. (2005). *Teaching ESL composition*. New Jersey: Lawrence Erlbaum Associates.

Flowerdew, J. (2001). Attitudes of journal editors to nonnative speaker contributions. *TESOL Quarterly, 35*, 121-150.

Gamon, M., Leacock, C., Brockett, C., Gao, J., & Klementiev, A. (2009). Using statistical techniques and web search to correct ESL errors. *CALICO Journal*, *26* (3), 491-511.

Glanville, R., Sengupta, S., & Forey, G. (1998). A (cybernetic) musing: Language and science and the language of science. *Cybernetics and Human Knowing*, *5*(4), 61-70.

Granger, S. eds. (1998). *Learner English on computer*. London: Longman.

Graesser, A.C. & McNamara, D.S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 2,* 371-398.

Graesser, A.C., Olde, B. A., & Klettke, B. (2002). How does the mind construct and represent stories? In M. Green, J. Strange, and T. Brock (Eds.), *Narrative Impact: social and cognitive foundations* (pp.57-69). Mahwah: NJ: Erlbaum.

Hyland, K. (2000). Disciplinary discourses: Social interactions in academic writing. Essex: Pearson Education.

Kučera, H. & Nelson, F. (1967). *Computational Analysis of Present-day American English*. Brown University Press.

McCarthy, P.M., Hall, C., Duran, N.D., Doiuchi, M., Duncan, B., Fujiwara, Y., & McNamara, D.S. (2009). A Coh-Metrix Analysis of Discourse Variation in the Texts of Japanese, American, and British Scientists *The ESPecialist, 30,* 141-173.

McCarthy, P.M., Watanabe, S.&Lamkin, T.A. (2012). The Gramulator: A Tool to Identify Differential Linguistic Features of Correlative Text Types. Applied Natural Language Processing: Identification, Investigation and Resolution. IGI Global.

Min, H.C & McCarthy, P.M (2012). Insiders and Outsiders: A Gramulator Analysis of the Journal Text of American and Korean Scientists. *TESOL Quarterly*. in review

Oakhill, J., & Cain, K. (2007). Issues of causality in children's reading comprehension. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp.47-71). Mahwah, New Jersey:

Lawrence Erlbaum Associates.

Perelman, C., & Olbrechts-Tyceta, L. (1969). *The new rhetoric: A treatise on argumentation*. Notre Dame: University of Notre Dame Press.

Porter, J. (1992). *Audience and Rhetoric: An archaeological composition of the discourse community*. New Jersey: Prentice Hall.

Reid, J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing*, *1*(2), 79-107.

Stubbs, M. (1996). *Text and corpus analysis*. Oxford: Blackwell.

Swales, J.M. (1990) *Genre analysis: English in academic and research Settings*. Cambridge: Cambridge University Press.

Tang, R. (2012). Academic writing in a second or foreign language: Issues and challenges facing ESL/EFL academic writers in higher education contexts. London: Continuum.

Trebits, A. (2009). The most frequent phrasal verbs in English language EU documents: A corpus-based analysis and its implications. *System*, *37*, 470-481.

van Dijk, T.A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Yu L.& Liang Y(2006). A Study of the Writing Model of English Scientific Papers. *Foreign Language Education*. 27(1). 34-37

Zwaan, R.A. (1993). *Aspects of literary comprehension*. Amsterdam: John Benjamins.