International Linguistics Research; Vol. 3, No. 4; 2020
ISSN 2576-2974 E-ISSN 2576-2982
https://doi.org/10.30560/ilr.v3n4p94

# A Comparative Study of Linguistic Features of English Agricultural Journal Abstracts Written by American and Chinese Scientists

Wenli Xu[1] & Yi Tang[2]

[1] School of English for International Business, Guangdong University of Foreign Studies, Guangzhou, China

[2] School of Foreign Languages, Huazhong Agricultural University, Wuhan, China

Correspondence: Yi Tang, School of Foreign Languages, Huazhong Agricultural University, No.1 Shizishan Street, Wuhan 430070, P.R. China. Tel: 0086-189-7156-4698. E-mail: 47361645@qq.com

Received: November 5, 2020; Accepted: November 30, 2020; Online Published: December 2, 2020

## Abstract

The present study investigated the variations in linguistic features of English academic writing by American and Chinese scientists by building a corpus of 600 English agricultural journal abstracts and using the natural language processing tool Coh-Metrix. Through a one-way Analysis of Variance (ANOVA) and a discriminant function analysis (DFA), we statistically analyzed the corpus texts based on their lexical, syntactic and cohesive features and generated 8 distinguishing linguistic indices. The results indicated that Chinese scientists tended to write abstracts with more frequent words, more similar sentence structures, more modifiers per noun phrase and more agentless passive voice forms, while the American counterparts tended to write abstracts with a wider range of vocabulary, more specific terms, more words with multiple senses and more adversative connectives. These findings offer good guidance for Chinese scientists to write in a style closer to the agricultural research field and the native speakers so as to get their manuscripts better reviewed and more easily published. These findings also have practical implications for the development of agricultural English teaching materials as well as the curriculum design.

**Keywords:** Coh-Metrix, linguistic features, agricultural English, English academic writing, cross-culture study

## 1. Introduction

When it comes to the prime directive of academia "Publish or Perish", researchers may devote all their time and energies to innovations. Unfortunately, these innovative scientists often fail to attach enough importance to the fact that publications should be in English. In the majority of science fields, the most prestigious journals and particularly those with the highest reputation and impact all over the world only publish in English. Therefore, it is not surprising that about 74% to 90% of international journals only accept scholarly work in English (Lillis & Curry, 2006). Moreover, with the increasing growth of these percentages (Tardy, 2004), the centralization on English as the language of academic articles gains momentum. It cannot be denied that there are also distinguished and illustrious journals which are published in other languages (Belcher & Connor, 2001; Canagarajah, 2002). Nevertheless, success in most fields depends on publishing in English. As a consequence, most non-native English speaking researchers have no choice but to submit their articles for publication in English.

As the abbreviated, accurate representation of all contents of an article, abstract is the very first part that journal editors would check and consider whether it can be accepted. What's more, some principal international retrieval authorities only capture the abstract of an article rather than the full text. And in most data bases, only abstracts are provided freely. Therefore, a concise, cohesive and focused abstract with appropriate and natural English writing, plays an extremely important role in publication, which, however, is a significant challenge for scientists, especially for non-native English speakers due to the fact of linguistic heterogeneity between L1 and L2. So, a full understanding of linguistic variation has important implications for academic practice. Over the past decades, numerous studies have uncovered evidence that linguistic features of texts vary across cultural backgrounds. These studies have helped illuminate and model the specific linguistic differences, which to a certain degree can serve as answers to the challenges. What follows is a brief review of the relevant empirical literature.

It is commonly assumed that greater differences exist along the lines of English genres (e.g. argumentative or expository texts) than along the lines of linguistic variation across cultural backgrounds (e.g. American or Chinese writings) within a specific written genre (Johansson, 1985), which has restricted the number of studies comparing different varieties of English. The few researches that have been conducted tend to highlight only those wellestablished aspects at superficial levels like phonological, lexical, and morphological level. The first multi-dimensional study across English language varieties was carried out by Biber (1987). Biber found evidence that British texts were more formal but less interactive and abstract than do their American counterparts. But Biber also pointed out that these differences were neither large nor consistent and it was necessary to figure out underlying linguistic differences. Later, these findings were enhanced by Helt (2001) who updated Biber's approach and found similar differences within the spoken registers. Following Biber's research model, Connor (1995) made a comparative analysis of essays written by British, New Zealand and American students, and found that British and New Zealand essays contained a higher type/token ratio than their American counterparts. Likewise, researchers at home have also done some similar studies. For example, Ma (2002) compared linguistic features between EFL (English as foreign language) essays written by Chinese students and ENL (English as native language) essays written by American students, and the results indicated that 9 out of 66 linguistic features (e.g. second person pronouns, discourse conjunctions) were significantly different between the two groups. It can be inferred from these findings that traditional shallow metric approaches to the study of distinguishing English varieties do not extend beyond word-level features (e.g. grammatical class and frequency) and these superficial analyses are defective. Underlying differences that may occur at higher-order text components such as in cohesion should be taken into account.

The difficulty of investigating linguistic features at deeper and more macro levels has not been overcome properly until the development of computational linguistics and the emergence of natural language processing tools (NLP tools), which can transform massive and complex texts into mathematical notations and expressions in an unimaginably short time. Discourse analyses based on NLP tools can quantize and compute various linguistic indices which are extremely difficult to get by hand, consequently exploring linguistic features in a more global level (Graesser & McNamara, 2011). For example, McCarthy and his colleagues (2009) conducted an interdisciplinary study comprising two complementary analyses on a corpus of English science journal abstracts written by American, British, and Japanese researchers, using the computational tool Coh-Metrix to assess texts at discourse level and another computational tool Gramulator to compare the frequency of n-grams. The results of this study showed 15 linguistic indices including word familiarity, syntax similarity, and argument overlap, et al., were significantly different across the three sources of abstracts. Likewise, Ye (2015) also used two NLP tools, Coh-Metrix and Gramulator, to analyze the specific and distinctive features of English Biological, Physical, and Chemical abstracts written by Chinese, American, and Korean scientists and she found 8 linguistic indices such as word familiarity, tense and aspect repetition index, and CELEX frequency, suggesting significant differences between abstracts written by American and Chinese scientists. From the later researches on linguistic variations across cultural backgrounds using NLP tools, it is shown that these computational tools can go beyond those surface components and go some distance in automating deeper and more global levels of text and language analysis.

From the prior research, we see that writers or researchers are usually influenced by their cultural context and hence use linguistic devices in their own ways. However, empirical investigations into the linguistic variation of expert academic writing such as research articles across cultural backgrounds are scarce with the few existing ones exploring research articles more in other fields than in agriculture. Also, little work has focused exclusively on Chinese researchers' academic writing. With the above limitations in mind, this paper intends to make a comparative study of the linguistic features of English agricultural journal abstracts written by American and Chinese scientists with the purpose of offering guidance for scientists to write in a style closer to the agricultural research field and the native speakers so as to get their manuscripts better reviewed and more easily published.

Five key components of this study are introduced briefly as follows. The very first part gives an overview of linguistic features and the research tool Coh-Metrix. The second part is about the research methods of the present study, discussing in detail what research questions are to be solved, how linguistic features are selected, how the corpus is constructed and how the data is statistically analyzed. In the third part, research results are reported which are then comparatively analyzed in part four from the levels of lexicon, syntax and cohesion. In the last part, we make a conclusion of the study and provide implications for academic English teaching and learning.

## 2. Linguistic Features

The definition of linguistic features is of much debate. Theoretically, linguistic features contain both shallow and local components such as mean number of syllables in words, and deep and global indices like WordNet verb overlap, and age of acquisition for content words. Broadly, hundreds of measures of language can be classified as linguistic features. But in a narrow sense, we commonly discuss only three main features: lexical, syntactic and cohesive features. Lexical features, just as its name implies, only focus on word information which refers to the idea that each word is assigned a syntactic part-of-speech category (e.g. nouns, prepositions) (Graesser et al., 2004).

Lexical features are comprised of a great number of measures. Take five measures as examples: word frequency calculating how frequent particular words occur in the English language based on certain word corpus (e.g. incidence of adjectives, average word frequency for content words), psycho-linguistic information of words (e.g. age of acquisition for content words, meaningfulness for content words), hypernymy (e.g. hypernymy for nouns, hypernymy for nouns and verbs), polysemy (e.g. polysemy for content words), and lexical diversity which refers to the range of vocabulary in a text based on different values (e.g. type-token ratio, MTLD lexical diversity). Following the theories of assigning words into syntactic part-of-speech category as in lexical level, syntactic features also group words into phrases or constituents (e.g. verb phrases, adverbial clauses), and construct syntactic tree structures for sentences. Specifically, take three measures as examples: syntactic similarity (e.g. syntactic similarity between all sentences), syntactic complexity (e.g. number of words before main verb), and syntactic pattern density (e.g. incidence of noun phrase, incidence of preposition phrase).

Compared with the former two relative transparent features, cohesive features require further explanation. Linguistic indices reported by computational tools can be used in a variety of ways to estimate the *cohesion* of the explicit text and the *coherence* of the implicit text. *Cohesion* refers to characteristics of the texts, while *coherence* is a characteristic of the reader's mental representation of the text content. As an object component of the explicit language, cohesion contains explicit words, phrases and sentences that can help readers in decoding the substantive ideas, and in connecting ideas with more global units (e.g. themes and theses). These cohesive indices provide hints for readers on how to form a coherent representation. While, coherence is an achievement of psychological representations and processes, which refers to interactions between linguistic and knowledge representations constructed in readers' mind. It depends on the skills and knowledge that readers bring to the situation, for example, it is probably impossible for a reader with little knowledge about the subject background to form a coherent mental representation. In a word, cohesion is a textual construct, whereas coherence is a psychological construct (Louwerse, 2001). As a result, cohesive features refer to explicit cohesive indices (e.g. noun overlap between adjacent sentences, incidence of logic connectives) which can make contributions to readers' coherence.

A considerable number of studies have been done by researchers at home and abroad during the past several decades to investigate the potential factors associated with the writing quality of English productions. Among all the possible factors, linguistic features, including tense (e.g. present and past), voice (e.g. active and passive), person (e.g. first and third), cohesion (e.g. connectives and co-referentiality), hedge (e.g. approximators and shields), and et al., have been heavily examined and proved to be linked to proficient writing. For example, Schleppegrell (2001) has analyzed some school-based texts and found that technical and specific lexis, and explicitly stated logical relations were required for the presentation of information; the choice of declarative mood, and the use of grammatical and lexical resources were the reflection of authoritativeness; and the elaboration of noun phrases could realize a high degree of structure. Later, McNamara, Crossley, and McCarthy (2010) have evaluated a corpus of expert-graded essays and found three most predictive indices of syntactic complexity, lexical diversity and word frequency, to distinguish the differences between high- and low-proficiency essays. Researchers at home have also conducted similar studies to explore the relations between linguistic features and the writing quality of English productions. For instance, Zeng and Hu (2005) found that hedges in abstracts of English academic papers reflected a rational way to deal with the interactions between authors and readers, which could inspire readers' sense of identity. With regard to passive voice, Fan (2005) found that the over-use of passive voice in Chinese medical journal articles led to the consequence of inflexible structure and semantic ambiguity, which falls short of latest international writing standards of scientific articles.

To sum up, linguistic features can be divided into three levels: lexical, syntactic, and cohesive features. Lexical features only focus on word information such as word frequency and lexical diversity, and syntactic features refer to constituents and syntactic tree structures in sentences. Cohesive features refer to a higher level of explicit cohesion indices to help with readers' coherence. And from the prior researches, indeed, it is reasonable to conclude that the writing quality of English productions is characterized by their linguistic features, which provides scientific support for our study.

## 3. Coh-Metrix

Rapid development of computational linguistics and discourse processing makes it possible to take into consideration of global text attributes and conceptual information. At the leading edge of the emerging computational techniques is a freely available, web-based software tool named Coh-Metrix (Graesser et al., 2004) developed at the University of Memphis.

Integrating a set of text analyzing tools, Coh-Metrix functions through various modules including latent semantic analysis (LSA) (Landauer & Dumais, 1997), part-of-speech taggers (Brill, 1995), and syntactic parsers (Charniak, 1997). Its word relationship indices root in the WordNet lexical database (Fellbaum, 1998), and conceptual information indices in the MRC database (Coltheart, 1981). Based on these modules and databases, Coh-Metrix has had three online versions since it came into service, and altogether it can generate over 600 indices of language and text (Graesser et al., 2004). In this study, we used the Coh-Metrix online version 3.0 which provided 106 linguistic indices (All indices are provided in Appendix A). In addition to these intricate indices, Coh-Metrix also provides a range of traditional superficial indices (e.g. number of paragraphs) and the readability scores of *Flesch Reading Ease* and *Flesch_Kincaid Grade Level* (Klare, 1974). Compared with previous work by hand, Coh-Metrix's output is more objective, consistent, accurate and efficient, and Coh-Metrix has been validated and shown to be an ideal tool for investigating linguistic variation.

More than 50 published researches have demonstrated that Coh-Metrix indices are effective enough to detect subtle linguistic differences between texts. For instance, McCarthy et al. (2006) found evidence that Coh-Metrix could successfully identify authorship even though the individual authors recorded significant shifts in their writing styles. Later, McCarthy et al. (2007) reported that Coh-Metrix could differentiate sections in typical science texts, such as introduction, method, result, and discussion. McNamara et al. (2006) used Coh-Metrix to distinguish high- and low-cohesion texts. Hall et al. (2006) demonstrated that Coh-Metrix could distinguish linguistic differences between law texts written by American and English/Welsh scientists. Crossley et al. (2007) investigated linguistic structure differences between sampled simplified texts and authentic reading texts using Coh-Metrix. Domestic researchers have done plenty of studies using Coh-Metrix as well. Jiang (2016) summarized the applications of Coh-Metrix in foreign language teaching and research. Du and Cai (2013) made a Coh-Metrix-based study of linguistic features and generated a model to predict the writing quality of argumentative essays written by English learners in China. Liang (2006) conducted a comprehensive study on the cohesion of EFL learners' written productions using Coh-Metrix.

On the whole, Coh-Metrix has taken a part in many research endeavors, ranging from distinguishing different types of texts to the application in foreign language teaching and learning. The wide range and variety of successful applications of Coh-Metrix provide us with a rich array of support in this analysis.

## 4. Methodology

In this section, the research methods of this study are discussed in detail. Firstly, guided by the prior research, three research questions are put forward. The second part presents the rationale of how to select appropriate indices from the 106 indices generated by Coh-Metrix online version 3.0. Then the third part introduces how to classify authors according to their cultural backgrounds and how to build the corpus of English agricultural journal abstracts. Finally, we discuss how to analyze the data generated by Coh-Metrix using SPSS, a data processing software tool.

### *4.1 Questions*

This study addresses the following three questions:

1) Which linguistic features are significantly different in English agricultural journal abstracts written by American and Chinese researchers?

2) What are the potential reasons for these significant linguistic differences?

3) What should agricultural scientists do to write more natural English abstracts in order to get their manuscripts better reviewed and more easily published?

This paper not only makes a comparative analysis of linguistic features of American and Chinese abstracts, but also inquiries into reasons for these significant linguistic differences. At the same time, it provides practical implications for the development of agricultural English teaching materials as well as the curriculum design.

### *4.2 Linguistic Features Selection*

In this study, Coh-Metrix online version 3.0 was adopted, which can generate 106 linguistic indices (All indices are provided in Appendix A). Since some of the 106 indices represent similar linguistic features and some indices proved not related to writing quality by past researches, only fourteen sets (altogether 62 indices) of Coh-Metrix measures were selected to reflect respectively lexical features (word frequency, lexical diversity, psycho-linguistic information of words, hypernymy and polysemy), syntactic features (syntactic complexity, syntactic similarity and syntactic pattern density) and cohesive features (lexical co-referentiality, semantic co-referentiality, givenness, connectives, temporal cohesion, and causal cohesion). In the remainder of this section, a brief description of these measures is presented. An extensive analysis of Coh-Metrix theories, modules and measures can be found in Graesser et al. (2004).

**Word frequency.** This measure includes 13 indices from No. 82 to No. 94 in Coh-Metrix online version 3.0. Word frequency refers to the likelihood of a particular word being familiar to the audience and subsequent likelihood of having been encountered by him previously, due to its frequency count taken from CELEX, a 17.9 million-word corpus (Baayen et al., 1995). Higher score in word frequency suggests that this text contains more frequent words which can facilitate quicker decoding (Rayner et al., 2012).

**Lexical diversity.** This measure includes 2 indices from No. 48 to No. 49 in Coh-Metrix online version 3.0. Lexical diversity (LD) measures the variety of particular word types deployed by the author in relation to the total number of words. Higher LD score is widely held as an indication of more linguistic skills, more textual difficulty, and a higher level of competence and socioeconomic status of the speaker (Grela, 2002; Avent & Austermann, 2003; McCarthy, 2005). Coh-Metrix offers several LD indices but in this study we only chose MTLD (McCarthy, 2005) and D (Malvern et al., 2004) values.

**Psycho-linguistic information of words.** This measure includes 5 indices from No.95 to No. 99 in Coh-Metrix online version 3.0. Psycho-linguistic information of words computes word information on five psychological dimensions: age of acquisition, familiarity, concreteness, imageability, and meaningfulness with the MRC psycho-linguistic database (Coltheart, 1981), a collection of human ratings of 150,837 words and providing information for 26 linguistic properties of these words. Take imageability index for an example, *butterfly* is a high imagery word as compared with the low imagery score for *likewise*.

**Hypernymy.** This measure includes 3 indices from No. 101 to No. 103 in Coh-Metrix online version 3.0. Hypernymy calculates word specificity with WordNet module (Fellbaum, 1998), an online lexicon tool which locates each word on a hierarchical scale measuring the number of superordinate words above and subordinate words below the target word. As a consequence, a higher value indicates an overall use of more specific words.

**Polysemy.** This measure includes only 1 index No. 100 in Coh-Metrix online version 3.0. Polysemy indicates word ambiguity by measuring the number of senses a word has based on the same module as hypernymy index is, the WordNet (Fellbaum, 1998). Higher polysemy score relates to the potential for a greater number of lexical interpretations.

**Syntactic complexity.** This measure includes 5 indices from No. 67 to No. 71 in Coh-Metrix online version 3.0. Syntactic complexity shows whether sentences contain simple or complex syntax by measuring the degree to which words, phrases and clauses are embedded in the text. Greater syntactic complexity indicates greater structural density, more textual ambiguity and ungrammaticality (Graesser et al., 2004). Sentences with complex syntax tend to be more difficult to process.

**Syntactic similarity.** This measure includes 2 indices from No. 72 to No. 73 in Coh-Metrix online version 3.0. Coh-Metrix reports syntactic similarity value by analyzing the structural representation of a sentence in a parse tree. Therefore, syntactic similarity measures the consistency and uniformity of the syntactic constructions in a text. A higher score for syntactic similarity is held to be indicative of more consistency in style and form.

**Syntactic pattern density.** This measure includes 8 indices from No. 74 to No. 81 in Coh-Metrix online version 3.0. Syntactic pattern density is informed by the density of particular word types, phrase types and syntactic patterns. The density of every index plays an important role in the processing difficulty of a text. For example, if a text has a higher score in agentless passive voice incidence, it is more likely to contain a complex syntax which is more difficult to process.

**Lexical co-referentiality.** This measure includes 8 indices from No. 28 to No. 36 except for No. 35 in Coh-Metrix online version 3.0. Lexical co-referentiality refers to an approximation of the conceptual redundancy between sentences. It includes four forms of lexical co-reference between sentences: noun overlap, argument overlap, stem overlap, and content word overlap. Take noun overlap index for an example, it is a proportion of all sentence pairs that share one or more common nouns.

**Semantic co-referentiality.** This measure includes 3 indices of No. 38, 40, and 42 in Coh-Metrix online version 3.0. Different from lexical co-referentiality, semantic co-referentiality measures conceptual overlap and semantic similarity between sentences and paragraphs by a sophisticated computational module for word meaning: Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997). LSA represents word meaning by analyzing the type of contexts where that word tends to occur. For example, the word *computer* will be highly associated with words of same functional context, such as *keyboard, mouse, software* and *screen*.

**Givenness.** This measure includes only No. 44 index in Coh-Metrix online version 3.0. Givenness refers to the proportion of new information each sentence provides within LSA values. Compared with new information, given information is thought to be recoverable from previous discourse (Halliday, 1967) and does not require activation (Chafe, 1976), bringing less cognitive load on readers.

**Connectives.** This measure includes 9 indices from No. 50 to No. 58 in Coh-Metrix online version 3.0. Coh-Metrix provides an incidence score (occurrence per 1000 words) for a large variety of connectives. This measure calculates the density of five general types of connectives: causal (e.g. because, so), logical (e.g. and, or), adversative/contrastive (e.g. although, whereas), temporal (e.g. first, until) and additive (e.g. and, moreover) (Halliday & Hasan, 1976). Connectives can increase the cohesion of a text by linking ideas in an explicit way, thus facilitating both comprehension and learning (Halliday & Hasan, 1976).

**Temporal cohesion.** This measure includes only No. 66 index in Coh-Metrix online version 3.0. It measures the repetition score for tense and aspect. The repetition score for tense is averaged with that for aspect.

**Causal cohesion.** This measure includes only No. 62 index in Coh-Metrix online version 3.0. It is the ratio of causal particles (e.g. because, consequence of) to causal verbs (e.g. kill, pour). The causal particle count depends on a defined set of causal particles while causal verb count is identified through WordNet (Fellbaum, 1998).

### *4.3 Corpus Collection*

The corpus for this study is comprised of 600 English agricultural journal abstracts written by American and Chinese scientists. The very first problem we have to solve is the classification of cultural backgrounds. So, we employed and adjusted the two criteria for text classification which are the common practice of many studies (e.g. McCarthy et al., 2009). Firstly, the first author is required to be affiliated with an institute within the country of classification. Secondly, the first author's name is required to be "typical" of the country of classification. While such a technique is by no means perfect, 46 nationality confirmation emails (Sample emails in English and Chinese are provided respectively in Appendix B and C) were sent to random authors in order to figure out whether the majority of abstracts would be appropriately categorized using these guidelines. Eight emails got replies and all the replied emails showed a correct confirmation.

To make sure that the chosen abstracts are representative, we searched the 2018 version of the Journal Citation Reports (JCR) list in the discipline of agriculture, which includes four sub-disciplines: agricultural dairy and science, agricultural economics and policy, agricultural engineering, and agricultural multidisciplinary. Firstly, 5 journals with the highest impact factors were selected respectively from the above four sub-disciplines. Secondly, articles published between 2011 and 2015 were searched respectively from the chosen 20 journals in the Web of Science database and ranked according to their impact factors from the highest to the lowest. Thirdly, under this sequence, the first 15 articles written by American researchers were selected out of the search results based on aforementioned criteria and their abstracts were downloaded. If there were no 15 articles meeting our requirements in one journal, articles from the next journal would be selected to make sure the total number of texts in the American corpus is 300. For example, if there is no article meeting our requirements in the first journal which is supposed to provide 15 articles, we then would select 30 articles from the second journal. Likewise, 300 abstracts written by Chinese researchers were downloaded as well. Thus, the selected journals and articles can be considered as the most representative ones in the discipline of agriculture. So, the corpus created in this study is consisted of the comparable American and Chinese corpora. Table 1 shows the composition of our corpus.

Table 1. Composition of the corpus

| **Journal** | **Number of abstracts (mean length)** | | **Size** |
|---|---|---|---|
| | **AMr** | **CNr** | |
| Annual Review of Animal Biosciences | 18(148.28) | 2(131.50) | 2,932 |
| Genetics Selection Evolution | 8(321.38) | 12(323.17) | 6,449 |
| Journal of Animal Science and Biotechnology | 19(249.42) | 21(250.14) | 9,998 |
| Journal of Dairy Science | 15(336.00) | 25(230.28) | 10,797 |
| Poultry Science | 15(383.87) | 15(282.67) | 9,997 |
| Food Policy | 15(217.00) | 6(123.17) | 3,994 |
| American Journal of Agricultural Economics | 15(112.27) | 1(166.00) | 1,850 |
| Annual Review of Resource Economics | 15(131.13) | 0(0.00) | 1,967 |
| Journal of Agricultural Economics | 11(118.00) | 0(0.00) | 1,298 |
| Agricultural Economics | 19(180.26) | 2(185.50) | 3,796 |
| Bioresource Technology | 15(156.93) | 81(140.99) | 13,774 |

| | | | |
|---|---|---|---|
| Industrial Crops and Products | 15(221.93) | 15(168.87) | 5,862 |
| Biomass & Bioenergy | 15(203.87) | 15(180.93) | 5,772 |
| Biosystems Engineering | 12(203.08) | 15(205.87) | 5,525 |
| Journal of Irrigation and Drainage Engineering | 18(225.67) | 15(199.60) | 7,056 |
| Agriculture Ecosystems & Environment | 15(232.07) | 15(296.60) | 7,930 |
| Agricultural Systems | 15(297.87) | 6(229.67) | 5,846 |
| International Journal of Agricultural Sustainability | 3(197.33) | 0(0.00) | 592 |
| Agriculture and Human Values | 16(200.00) | 1(225) | 3,425 |
| Journal of Agricultural And Food Chemistry | 26(178.00) | 53(181.62) | 14,254 |
| **Total corpus size** | 300(62,494) | 300(59,095) | 600(121,589) |

### *4.4 Data Analysis*

Firstly, the 600 journal abstracts were input one by one into Coh-Metrix online version 3.0 to get indices of linguistic features automatically. All results were extracted from the selected abstracts by the computational tool and no work by hand. Secondly, the corpus was split into two equally sized groups randomly: the training set (including 150 American abstracts and 150 Chinese abstracts) and the test set (the rest 300 abstracts). The training set was designed to identify which of the 62 Coh-Metrix indices showed significant differences between American and Chinese cultural backgrounds. Thirdly, a one-way Analysis of Variance (ANOVA) was carried out on the training set to identify which of the indices contained in the 14 selected measures could best distinguish the two cultural backgrounds. Fourthly, a stepwise discriminant function analysis (DFA) was conducted and then generated a model using these indices as variables to predict cultural backgrounds in the test set. Finally, we used three values of *Recall*, *Precision*, and *F1* to assess the accuracy and precision of this model.

## 5. Results

A one-way ANOVA was conducted using the selected Coh-Metrix measures as the dependent variables, and cultural backgrounds from the training set as the independent variables (American researchers and Chinese researchers). Variables with the largest effect size were selected as the representative variable for each measure. With the exception of six measures: psycho-linguistic information of words, lexical co-referentiality, semantic co-referentiality, givenness, temporal cohesion and causal cohesion, the rest 8 measures contained at least one index which demonstrated significant differences between cultural backgrounds.

Pearson's correlations were conducted between the selected 8 variables to ensure that no index pair correlated above r => .70, which was a conservative standard on the issue of multi-collinearity (e.g. Duran et al., 2007; McNamara et al., 2010). The problem of multi-collinearity refers to situations where two or more variables correlate at approximately r => .70. Model with such correlated variables could not reflect accurate relationship between independent variables and dependent variables (Brace et al., 2002). The results of Pearson's correlations showed that no variable pair was correlated above r => .70. Therefore, all the 8 selected variables were not highly correlated and should be kept in the analysis.

Because each set (training and test) was comprised of 300 abstracts, a maximum of 15 variables could be selected for the analysis in case of over-fitting of the model, according to the typical ratio (20:1) of statistical analyses of this kind (e.g. Duran et al., 2007, McNamara et al., 2010). Thus, none of the 8 variables should be removed. Descriptive statistics for the 8 variables are presented in Table 2 and described below.

Table 2. Descriptive statistics of linguistic variation between American and Chinese researchers

| **Variable** | **Amr** | | **CNr** | | **F(1,298)** | $\eta^2_p$ |
|---|---|---|---|---|---|---|
| | **M** | **SD** | **M** | **SD** | | |
| Hypernymy for nouns | 6.38288 | 0.612483 | 5.54014 | 0.787089 | 107.105*** | 0.264 |
| Polysemy for content words | 3.35403 | 0.384752 | 3.04663 | 0.401193 | 45.873*** | 0.133 |
| Number of modifiers per noun phrase | 1.26851 | 0.275249 | 1.43723 | 0.244961 | 31.450*** | 0.095 |
| Incidence of adversative connectives | 9.65544 | 7.637183 | 5.20089 | 6.220266 | 30.679*** | 0.093 |
| Incidence of agentless passive voice | 14.16017 | 10.501401 | 20.64749 | 10.858653 | 27.665*** | 0.085 |
| Average word frequency for all words | 2.71357 | 0.126631 | 2.78849 | 0.136980 | 24.190*** | 0.075 |
| Syntactic similarity between all sentences | 0.08629 | 0.025569 | 0.09521 | 0.028613 | 8.105** | 0.026 |
| Lexical diversity index of MTLD | 87.97969 | 29.557015 | 80.25335 | 25.746949 | 5.828* | 0.019 |

(Note: *** significant at $p < .001$; ** significant at $p < .01$; * significant at $p < .05$)

**Hypernymy for nouns.** The values for hypernymy for nouns indicate that American abstracts contain more specific nouns than Chinese abstracts do. The results suggest that Chinese abstracts hold words with a lower level in a conceptual, taxonomic hierarchy, which means it is easier for readers to comprehend.

**Polysemy for content words.** The results generated for polysemy indicate that American writers appear to use significantly more high polysemy words than do the Chinese counterparts. This means that American writers prefer to use words with high ambiguity, which may cause greater processing difficulty to readers.

**Number of modifiers per noun phrase.** The number of modifiers per noun phrase index reflects that Chinese abstracts contain sentences that have significantly more modifiers per noun phrase than American abstracts do. This result indicates that sentences in Chinese abstracts hold more embedded, complex syntax which potentially lead to greater processing difficulty.

**Incidence of adversative connectives.** The American researchers appear to put significantly heavier reliance on the use of adversative connectives (e.g. although, whereas). With such an outcome, we may infer that American abstracts tend to provide more adversative clues about text organization, which can be easier for readers to process and ease the burden on their comprehension.

**Incidence of agentless passive voice.** The Chinese writers employ significantly more agentless passive voice forms than do American writers. The Chinese abstracts' more density of such a particular syntactic pattern leads to more syntactic complexity, which may increase the processing difficulty of the texts.

**Average word frequency for all words.** The results for word frequency index indicate that American researchers use significantly more low frequency words than do Chinese researchers. Such an outcome suggests that American researchers may prefer to use more specialized terms with low word frequency, while Chinese researchers may be taking more care in the choice of lexicon, which indicates that American abstracts is more difficult for audience to read and process.

**Syntactic similarity between all sentences.** The results generated from the syntactic similarity index suggest that Chinese researchers write significantly more syntactically similarly constructed sentences than American researchers do. This shows Chinese reluctance or inability to express ideas in a variety of ways, which can decrease the demands on readers' working memory and thus facilitate reading.

**Lexical diversity index of MTLD.** The results in this analysis indicate that the Chinese scientists appear to use a significantly narrower range of vocabulary than do the American counterparts. The higher lexical diversity of American abstracts can place heavier demands on the readers' working memory and subsequently cause greater difficulty in comprehension.

To test the accuracy of the aforesaid findings, a stepwise discriminant function analysis (DFA) was conducted on the training set, using the cultural backgrounds (American researchers and Chinese researchers) as the dependent variables and the selected 8 linguistic indices as the independent variables. The Wilks' Lambda for the function of the model was significant ($\Lambda = .608$, $^{2}(8) = 146.470$, $p < .001$). The structure matrix with the classification function coefficients is shown in Table 3.

Table 3 Structure matrix of the discriminant functions for linguistic indices and constant for cultural backgrounds

| **Linguistic indices** | **Cultural background** | |
|---|---|---|
| | **AMr** | **CNr** |
| Lexical diversity index of MTLD | 0.391 | 0.386 |
| Incidence of adversative connectives | 0.754 | 0.680 |
| Number of modifiers per noun phrase | 39.964 | 41.752 |
| Syntactic similarity between all sentences | 66.337 | 77.115 |
| Incidence of agentless passive voice | -0.169 | -0.140 |
| Average word frequency for all words | 221.197 | 223.280 |
| Polysemy for content words | 6.030 | 5.036 |
| Hypernymy for nouns | 22.035 | 20.669 |
| (Constant) | -429.108 | -426.397 |

Then we used the above DFA model from the training set to make a prediction of group membership of cultural backgrounds in the test set. The accuracy of our findings can be appraised by the correspondence between the actual cultural backgrounds (either American or Chinese) and the predictions reported by the DFA in both training and test sets (see Table 4). At the same time, we conducted a chi-square test to report Kappa value to describe the range of deviation between the actual types and the predictions. If the Kappa value is between 0.41 and 0.60, it represents a moderate agreement. The results indicate that the DFA in training set correctly classified 238 of the total 300 abstracts as American or Chinese ( $^2$(1) = 103.419, p < .001) for an accuracy of 79.3%. The reported Kappa value (=.587) means a moderate agreement between the actual and predicted classification. Moreover, the DFA in test set correctly classified 223 of the total 300 abstracts as American or Chinese ( $^2$(1) = 71.132, p < .001) for an accuracy of 74.3%. The reported Kappa value (=.487) also shows a moderate agreement between the actual and predicted classification.

Table 4. Actual versus predicted cultural backgrounds results in both training and test set

| **Actual cultural backgrounds** | **Predicted cultural backgrounds** | |
|---|---|---|
| | **AMr** | **CNr** |
| **Training set** | | |
| **Amr** | 122 | 28 |
| **CNr** | 34 | 116 |
| **Test set** | | |
| **Amr** | 109 | 41 |
| **CNr** | 36 | 114 |

Table 5. Recall, precision and F1 scores for predicting cultural backgrounds

| **Cultural backgrounds** | **Recall** | **Precision** | **F1** |
|---|---|---|---|
| **Training set** | | | |
| **Amr** | 0.813 | 0.782 | 0.797 |
| **CNr** | 0.773 | 0.806 | 0.789 |
| **Test set** | | | |
| **Amr** | 0.727 | 0.752 | 0.739 |
| **CNr** | 0.76 | 0.735 | 0.748 |

However, only the above accuracy indices are nowhere near enough to estimate the accuracy of our DFA model in a more scientific and comprehensive way. Therefore, according to the typical practice of discriminate analysis studies (e.g. Hall et al., 2007), the accuracy of our findings is reported in terms of *recall*, *precision* and *F1 values*. Recall shows the number of *true positive* (the correct predictions) divided by the number of *true positive* plus *false negative* (the incorrect predictions as negative type). For instance, in the training set, the recall of the abstracts predicted as American is the number of abstracts correctly classified as American (122) divided by the number of correct predictions (122) plus those incorrectly classified as Chinese (28), in other words, the total number of abstracts in the actual American group (150). On the other hand, precision is the number of *true positive* (the correct predictions) divided by the number of *true positive* plus *false positive* (the incorrect predictions as positive type). Take the training set as an instance as well. The precision of the abstracts predicted as American is the number abstracts correctly classified as American (122) divided by the number of correct predictions (122) plus those incorrectly predicted as American (34). Nevertheless, recall and precision may be contradictory to each other in some instances where a DFA model classified every item correctly as the members of the positive group leading to a recall of 100% but a low score in precision, because there may be a large proportion of false predictions as the negative group. Therefore, we choose F1 value to report a more comprehensive assessment. F1 score is calculated with the formula F1=2*R*P/(P+R) by considering both recall and precision scores. Table 5 shows detailed scores of recall, precision and F1. To sum up, the average accuracy of our DFA model for the training set and test set is .793 and .744 respectively, which proves that this model is of relatively high precision.

## 6. Discussion

Supported by the results of above analyses, it proves that there is a wide variety of distinctions between English agricultural journal abstracts written by American and Chinese scientists. Given that these distinctions cover varied indices at each of the text analysis levels including lexical features, syntax features and cohesive features, an integrated discussion will be made from the perspective of the above three linguistic features. And in the remainder of this part, the authors intend to discuss potential factors for these significant linguistic differences from perspectives of social psychology, English language teaching and compensatory mechanism.

### *6.1 Lexical Features*

From the previous analysis, it is clear that four indices in lexical features show significant linguistic differences between American and Chinese abstracts. They are Average word frequency for all words, Lexical diversity index of MTLD, Hypernymy for nouns, and Polysemy for content words. And there is no significant linguistic difference in the measure of Psycho-linguistic information of words. To summarize the results of above four significantly different lexical features, we can see that Chinese researchers appear to use a narrower range of vocabulary with more frequent words, while American researchers prefer to use more specific terms and words with multiple senses. One possible explanation of this result is that Chinese researchers prefer to stick to words or structures they know reasonably well tending to realize grammatical correctness and avoid making errors, but American researchers may assume that readers are very familiar with the subject and even very professional so it is acceptable to use more specialized terms and words with high ambiguity. This hypothesis is further supported by the results of syntactic similarity (see 4.2). Considering only the lexical features, it can be inferred that Chinese abstracts are easier for readers to read and process due to their lower degree of lexical variety and sophistication, which corresponds to the findings by Silva (1993). However, American abstracts may better serve the expectations of journal reviewers due to their higher lexical diversity and wider range of vocabulary.

### *6.2 Syntactic Features*

The aforementioned analysis has demonstrated that three indices in syntactic features can distinguish American and Chinese abstracts. They are Number of modifiers per noun phrase, Syntactic similarity between all sentences, and Incidence of agentless passive voice. Considering the results of index of syntactic similarity, Chinese scientists prefer to express ideas in similar ways leading to easier processing and comprehension. Thus, we may infer that Chinese scientists prefer to stick to structures they know certainly well out of convenience or simplicity for fear of making errors, which can also be concluded from the results of lexical features. However, from the results of indices of modifiers and agentless passive voice, it seems that Chinese scientists try to embed more constituents to reach more complex syntax, making the texts more ambiguous and more difficult to comprehend. It can be assumed that Chinese scientists have to use more embedded phrases and clauses to make up the lack of lexical variety and specialized terms. Considering only the syntactic features, it is reasonable to infer that Chinese abstracts put more demand on readers' working memory and are more difficult to understand due to their more complex syntax.

### *6.3 Cohesive Features*

According to the results of above analyses, it is surprising to find that only one index can significantly distinguish American and Chinese abstracts, which is the Incidence of adversative connectives. The more adversative connectives in American abstracts may result in more text cohesion and thus facilitate readers' working memory and comprehension. However, there is no significant difference in the indices of Semantic co-referentiality and Givenness, which indicates that from the perspective of semantic co-referentiality, Chinese abstracts show no significant difference from American abstracts. According to De Beaugrande and Dressler (1996) and Liang (2006), superficial cohesive ties such as connectives do not contribute much to the cohesion of texts and it is the semantic cohesion that plays a fundamental role in discourse cohesion. Therefore, all linguistic indices considered, we can infer that Chinese abstracts hold good discourse cohesion. With regard to index of lexical co-referentiality, there is no significant difference between American and Chinese abstracts, in correspondence to Reynolds' (1995) study which demonstrated that it was impossible to use lexical repetition to distinguish L1 and L2 writers. Likewise, the result that the index of causal cohesion shows no significant difference between American and Chinese abstracts corresponds to findings of Graesser et al. (2004), which demonstrated that causality is generally not of great importance for texts expressing abstract logical arguments. Considering only the cohesive features, we can infer that Chinese abstracts are not significantly different from their American counterparts except for the adversative connective index.

To sum up, the scientific statistical analysis has proved that there is indeed a wide variety of distinctions between English agricultural journal abstracts written by American and Chinese scientists. These distinctions can be classified in three levels of lexicon, syntax, and cohesion. With respect to lexical features, Chinese abstracts contain a too narrow range of vocabulary with too many frequent words, while American abstracts hold more specific terms and words with multiple senses. In terms of syntactic features, Chinese scientists use too many similar sentence structures, and the overuse of modifiers per noun phrase and agentless passive voice forms leads to a consequence of more complex syntax than their American counterparts. At last, with regard to cohesive features, there is no significant difference between the two English varieties from the perspective of semantic cohesion, but Chinese abstracts lack adversative connectives compared with American abstracts. In a word, Chinese abstracts are easier to comprehend on the lexical dimension but put more demand on understanding on the syntactic and cohesive dimensions.

### *6.4 Reasons for the Differences*

It cannot be denied that language is the carrier of culture and the linguistic features are the reflection of cultural backgrounds. Likewise, during the writing of agricultural English journal abstracts, it is inevitable that the author would be deeply influenced by his cultural background, which is apparently a natural advantage for native English speakers (e.g. American) but a huge challenge for non-native English speakers (e.g. Chinese). In this section, it is the authors' intention to try to explain the above linguistic differences between American and Chinese abstracts from the aspects of social psychology, English language teaching and compensatory mechanism.

#### 6.4.1 Social Psychology

In academic circle, people hold different attitudes towards authority in Asian culture and in Western culture. It is widespread that the respect towards authority and the recognition to community is highly valued in Asian countries. According to Hyland (2012), Asian scholars appear to be so respectful to authorities that they prefer to duplicate and inherit their predecessors' knowledge and experience through memory and imitation. He also mentioned that scholars laid more stress on analyzing, criticizing and evaluating authorities in Western culture. These two different social cultures subsequently shape different individual values and psychological characteristics. In Western society, scholars pursue individuality and originality with the purpose of having their *voice* heard, so they are more likely to opt for active voice forms to emphasize the agents. However, in Asian society, scholars emphasize the pursuit of collectivity and authority which may lead to conflicts between creativity and the respect to authority. According to Lu (2016), passive voice forms possess the function of emphasizing and hiding information. As a result, Chinese researchers tend to use more agentless passive voice forms to avoid the conflicts with authorities by hiding the agents.

#### 6.4.2 English Language Teaching

The variety of English taught in school would subsequently be presented in compositions. As what has been mentioned in the previous section, Hyland (2012) pointed out that scholars in Western cultures held a criticized and evaluated attitude towards authorities, and therefore, students were encouraged to criticize and then rebuild the existing knowledge to form their own ideas. On the contrary, in China, under the pressure of examination-oriented education, a majority of English teachers and students overemphasize rote memorization of the correctness of grammars. Thus, students have got used to memorizing and imitating those fixed collocations in order to get higher grades (Dai, 2001). These two totally different English language teaching methods result in that Chinese abstracts contain more similar sentence structures and more embedded constituents. And on the other hand, Chinese traditional English teaching attaches little importance to lexicon learning, which can be an explanation of the results that Chinese abstracts contain a narrower range of vocabulary with more frequent words and lack adversative connectives.

#### 6.4.3 Compensatory Mechanism

A common feature of language use in instructional texts is linguistic compensation (Medimorec et al., 2015). The compensatory mechanism refers to that difficulty on one dimension (e.g. Incidence of agentless passive voice) would be compensated for by increasing text ease on another dimension (e.g. Hypernymy for nouns). The analysis results demonstrates that the decreases of reading difficulty in lexical features (e.g. Polysemy for content words) of Chinese abstracts are associated with the increases in syntactic and cohesive features (e.g. Number of modifiers per noun phrase). To be specific, Chinese scientists have to use more embedded phrases and clauses, and more complex syntax to make up the lack of specific words, lexical diversity, and adversative connectives.

## 7. Conclusion

This study explored the linguistic variation of English academic writing by American and Chinese scientists by building a corpus of 600 English agricultural journal abstracts and using the computational tool Coh-Metrix. Through a one-way Analysis of Variance (ANOVA) and a Discriminant Function Analysis (DFA), eight Coh-Metrix linguistic indices were generated to distinguish these two language varieties. Overall, the results indicated that Chinese scientists tended to write abstracts with more frequent words, more similar sentence structures, more modifiers per noun phrase and more agentless passive voice forms, while the American counterparts tended to write abstracts with a wider range of vocabulary, more specific terms, more words with multiple senses and more adversative connectives. Different social psychology and English language teaching environments between American and Chinese cultures can be the explanation of these linguistic differences. The results that the difficulty in syntactic and cohesive features of Chinese abstracts is compensated for by increasing text ease in lexical features, corresponding to the compensatory mechanism (Medimorec et al., 2015). The results also provide evidence for Coh-Metrix's effectiveness in distinguishing different language varieties in deeper and more global levels.

Even though this study cannot provide any evidence for the claim that the above linguistic differences would bring negative effects on readers, it is reasonable to assume that for Chinese agricultural scientists, these differences may get in the way of gaining optimal reviews. Journal reviewers have certain expectations as to the abstracts, which go beyond mere lexical and grammatical correctness. Diversity within lexicon and syntactic structures is required to enhance the reader's interest. More specific and specialized terms and relative simple syntax can serve the principle of scientific essays: technicality and professionalization. Therefore, Chinese agricultural scientists need to pay more attention to the above aspects in order to increase their chances of being accepted by journal reviewers. They may be faced with the prospect of learning native English expressions, and adapting their English writing styles according to the requirements of a certain journal. The model generated by our discriminant function analysis may go a long way to assist Chinese agricultural scientists in evaluating the degree to which their abstracts have met those standards.

The results are also of special importance for the development of academic English teaching materials. It is the difficulty of collecting suitable amount of material that gets in the way of ESP teaching (Orr, 2001). However, as demonstrated in this study, a large number of natural examples of target agricultural texts are available, which provides agricultural educators or researchers with free corpus to determine which linguistic aspects to teach or to learn. For example, from the results of our analysis, we can see that Chinese agricultural scientists should learn more specific and specialized technical terms rather than just pay special attention to the correctness of lexicon and grammar. Therefore, agricultural educators can customize teaching material with more technical terms for Chinese scientists, and minimize pure English knowledge teaching such as grammar lessons.

Though with so many meaningful implications, this thesis is impossibly perfect. Three of the major limitations of our thesis are the use of only one computational tool, the limitation of native English speakers' texts to American, and the focus only on agricultural journal abstracts. Various NLP tools such as Gramulator, and WordSmith can be used in future research to explore more underlying, consistent and comprehensive differences between English varieties. And future studies must also consider productions of other language groups such as British English in order to adapt to the integrated international market. Linguistic distinctions of other sections of academic articles (e.g. introductions, methods, results, and discussions) between inter-disciplinary English varieties can be future research focal points.

## Acknowledgments

I am so grateful that so many respectable professionals and lovely intimates have provided selfless help to this study. First of all, my supervisor, Professor Ping Liu's kind instructions and warmhearted encouragement deserves my heartfelt gratitude. My deep gratitude would go to all the other teachers in School of English for International Business for their direct and indirect help. I am also deeply indebted to those American and Chinese agricultural scientists for providing their abstracts for this research. I would like to thank my dear friends Wei Li and Yufan Qu for their patient guidance in the use of software SPSS.

## References

Avent, J., & Austermann, S. (2003). Reciprocal scaffolding : A context for communication treatment in aphasia. *Aphasiology*, *17*(4), 397-404. https://doi.org/10.1080/02687030244000743

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia: University of Pennsylvania.

Belcher, D., & Connor, U. (2001). *Reflections on Multiliterate Lives*. Clevedon: Bilingual Education and Bilingualism.

Biber, D. (1987). A Textual Comparison of British and American Writing. *American Speech*, *62*(2), 99-119. https://doi.org/10.2307/455273

Brace, N., Kemp, R., & Snelgar, R. (2002). *SPSS for psychologists: A guide to data analysis using SPSS for Windows* (Chinese tr). Taiwan: Wu-Nan Books Inc.

Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, *21*(4), 543-565. https://doi.org/10.5555/218355.218367

Canagarajah, S. (2002). *A Geopolitics of Academic Writing*. Pittsburgh: University of Pittsburgh Press.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. Li (Ed.), *Subject and Topic* (pp. 25-55). New York: Academic Press.

Charniak, E. (1997). Statistical Parsing with a Context-free Grammar and Word Statistics. *Proc National Conference on Artificial Intelligence*.

Connor, U. (1995). Examining syntactic variation across three English-speaking nationalities through a multi-dimensional approach. In D. L. Rubin (Ed.), *Composing Social Identity in Written Language*. Hillsdale: Lawrence Erlbaum.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, *91*(1), 15-30. https://doi.org/10.1111/j.1540-4781.2007.00507.x

Dai, W. D. (2001). Gou jian ju you zhong guo te se de ying yu jiao xue "yi tiao long" ti xi [The construction of the streamline ELT system in China]. *Wai Yu Jiao Xue Yu Yan Jiu [Foreign Language Teaching and Research].*, *33*(5), 322-327+399.

De Beaugrande, R., & Dressler, W. (1996). *Introduction to text linguistics*. New York: Longman.

Du, H. Y., & Cai, J. T. (2013). Ji yu Coh-Metrix de zhong guo ying yu xue xi zhe yi lun wen xie zuo zhi liang yu ce mo xing yan jiu [A Coh-Metrix-based model of linguistic features predicting argumentative writing quality of EFL learners]. *Xian Dai Wai Yu [Modern Foreign Languages]*, *36*(3), 293-300+331.

Duran, N., McCarthy, P., Graesser, A., & McNamara, D. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods*, *39*(2), 212-223. https://doi.org/10.3758/BF03193150

Fan, X. H. (2005). Lun yi xue lun wen ying wen zhai yao zhong de bei dong yu tai de lan yong [The overpassivization in English abstracts of Chinese medical journals]. *Zhong Guo Ke Ji Fan Yi [Chinese Science & Technology Translators Journal]*, *18*(4), 13-16. https://doi.org/10.16024/j.cnki.issn1002-0489.2005.04.004

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*(2), 371-398. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, *36*(2), 193-202. https://doi.org/10.3758/BF03195564

Grela, B. G. (2002). Lexibal verb diversity in children with Down syndrome. *Clinical Linguistics and Phonetics*, *16*(4), 251-263. https://doi.org/10.1080/02699200210131987

Hall, C., Lewis, G., McCarthy, P., Lee, D., & McNamara, D. (2006). Language in Law: Using Coh-Metrix to Assess Differences between American and English/Welsh Language Varieties. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2498.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English: Part 2. *Journal of Linguistics*, *3*(2), 199-244. https://doi.org/10.1017/S0022226700016613

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. New York: Longman.

Helt, M. (2001). A multi-dimensional comparison of British and American spoken English. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies*. Essex: Pearson.

Hyland, K. (2012). *Disciplinary identities: individuality and community in academic writing*. Cambridge: Cambridge University Press.

Jiang, J. (2016). Coh-Metrix gong ju zai wai yu jiao xue yu yan jiu zhong de ying yong [The Application of Coh-Metrix in Foreign Language Teaching and Research]. *Zhong Guo Wai Yu [Foreign Languages in China]*, *13*(5), 58-65. https://doi.org/10.13564/j.cnki.issn.1672-9382.2016.05.009

Johansson, S. (1985). Some observations on word frequencies in three corpora of present-day English texts. *International Journal of Applied Linguistics*, *67*, 117-126. https://doi.org/10.1075/itl.67-68.08joh

Klare, G. R. (1974). Assessing Readability. *Reading Research Quarterly*, *10*(1), 62-102. https://doi.org/10.2307/747086

Liang, M. C. (2006). Xue xi zhe shu mian yu yu pian lian guan xing de yan jiu [A study of coherence in EFL learners4 written production]. *Xian Dai Yai Yu [Modern Foreign Languages]*, *29*(3), 284-292.

Lillis, T., & Curry, M. (2006). Re-Farming notions of competence in multilingual scholarly writing. *Revista Canaria de Edtudios Inglese*, *53*, 63-78.

Louwerse, M. A. X. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, *12*(3), 291-315. https://doi.org/10.1515/cogl.2002.005

Lu, C. . (2016). Ying yu bei dong ju de yu pian gong neng ping xi [On Textual Functions of English Passive Voice]. *Jia Mu Si Zhi Ye Xue Yuan Xue Bao*, *160*(3), 340-341.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical Diversity and Language Development*. London: Palgrave Macmillan.

McCarthy, P. (2005). *An assessment of the range and usefulness of lexical diversity and the potential of the measure of textual, lexical diversity (MTLD)*. Memphis Tennessee: University of Memphis.

McCarthy, P., Briner, S., Rus, V., & McNamara, D. (2007). Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. In A. Kao & S. Poteet (Eds.), *Natural language processing and text mining* (pp. 107-122).

McCarthy, P., Hall, C., Duran, N., Doiuchi, M., Fujiwara, Y., Duncan, B., & McNamara, D. (2009). Analyzing Journal Abstracts Written by Japanese, American, and British Scientists Using Coh-Metrix and the Gramulator. *The ESPecialist: Research in Language for Specific Purposes*, *30*(2), 141-173.

McCarthy, P., Lewis, G., Dufty, D., & McNamara, D. (2006). Analyzing writing styles with Coh-Metrix. *Proceeding of the 19th Annual Florida Artificial Intelligence Research Society International Conference*.

McNamara, D., Ozuru, Y., Graesser, A., & Louwerse, M. (2006). Validating Coh-Metrix. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 572-578.

Mcnamara, D. S., Crossley, S. A., & Mccarthy, P. M. (2010). Linguistic Features of Writing Quality. *Written Communication*, *27*(1), 57-86. https://doi.org/10.1177/0741088309351547

Orr, T. (2001). English Language Education for Specific Professional Needs. *IEEE Transactions on Professional Communication*, *44*(3), 207-211. https://doi.org/10.1109/47.946467

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. (2012). *Psychology of Reading* (Second edi). New York: Psychology Press.

Schleppegrell, M. J. (2001). Linguistic Features of the Language of Schooling. *Linguistics and Education*, *12*(4), 431-459. https://doi.org/10.1016/S0898-5898(01)00073-0

Silva, T. (1993). Toward an Understanding of the Distinct Nature of L2 Writing: The ESL Research and Its Implications. *TESOL Quarterly*, *27*(4), 657-677. https://doi.org/10.2307/3587400

Tardy, C. (2004). The role of English in scientific communication: lingua franca or Tyrannosaurus rex? *Journal of English for Academic Purposes*, *3*(3), 247-269. https://doi.org/10.1016/j.jeap.2003.10.001

Ye, D. M. (2015). *Ji suan yu yan xue shi yu xia de zhong-mei-han san guo ke ji qi kan ying wen zhai yao zuo zhe shen fen yan jiu [A Computational Linguistic Research on Writer Identity in English Abstracts Written by Chinese, American, and Korean Scientists: An Analysis us*. Nan jing shi fan da xue [Nanjing Normal University].

Zeng, Y. F., & Wei, F. (2005). Ying yu xue shu lun wen zhai yao zhong de mo hu xian zhi yu [Hedges in Abstracts of English Academic Papers]. *Shan Dong Wai Yu Jiao Xue [Shandong Foreign Languages Teaching Journal]*, *105*, 40-42. https://doi.org/10.16482/j.sdwy37-1026.2005.02.021

## Appendix A

**Linguistic Indices in Coh-Metrix 3.0 Output File**

| | Label in Version 3.x | Description |
|---|---|---|
| **Descriptive** | | |
| **1** | DESPC | Paragraph count, number of paragraphs |
| **2** | DESSC | Sentence count, number of sentences |
| **3** | DESWC | Word count, number of words |
| **4** | DESPL | Paragraph length, number of sentences, mean |

| | | |
|---|---|---|
| **5** | DESPLd | Paragraph length, number of sentences, standard deviation |
| **6** | DESSL | Sentence length, number of words, mean |
| **7** | DESSLd | Sentence length, number of words, standard deviation |
| **8** | DESWLsy | Word length, number of syllables, mean |
| **9** | DESWLsyd | Word length, number of syllables, standard deviation |
| **10** | DESWLlt | Word length, number of letters, mean |
| **11** | DESWLltd | Word length, number of letters, standard deviation |
| **Text Easability Principal Component Scores** | | |
| **12** | PCNARz | Text Easability PC Narrativity, z score |
| **13** | PCNARp | Text Easability PC Narrativity, percentile |
| **14** | PCSYNz | Text Easability PC Syntactic simplicity, z score |
| **15** | PCSYNp | Text Easability PC Syntactic simplicity, percentile |
| **16** | PCCNCz | Text Easability PC Word concreteness, z score |
| **17** | PCCNCp | Text Easability PC Word concreteness, percentile |
| **18** | PCREFz | Text Easability PC Referential cohesion, z score |
| **19** | PCREFp | Text Easability PC Referential cohesion, percentile |
| **20** | PCDCz | Text Easability PC Deep cohesion, z score |
| **21** | PCDCp | Text Easability PC Deep cohesion, percentile |
| **22** | PCVERBz | Text Easability PC Verb cohesion, z score |
| **23** | PCVERBp | Text Easability PC Verb cohesion, percentile |
| **24** | PCCONNz | Text Easability PC Connectivity, z score |
| **25** | PCCONNp | Text Easability PC Connectivity, percentile |
| **26** | PCTEMPz | Text Easability PC Temporality, z score |
| **27** | PCTEMPp | Text Easability PC Temporality, percentile |
| **Referential Cohesion** | | |
| **28** | CRFNO1 | Noun overlap, adjacent sentences, binary, mean |
| **29** | CRFAO1 | Argument overlap, adjacent sentences, binary, mean |
| **30** | CRFSO1 | Stem overlap, adjacent sentences, binary, mean |
| **31** | CRFNOa | Noun overlap, all sentences, binary, mean |
| **32** | CRFAOa | Argument overlap, all sentences, binary, mean |
| **33** | CRFSOa | Stem overlap, all sentences, binary, mean |
| **34** | CRFCWO1 | Content word overlap, adjacent sentences, proportional, mean |
| **35** | CRFCWO1d | Content word overlap, adjacent sentences, proportional, standard deviation |
| **36** | CRFCWOa | Content word overlap, all sentences, proportional, mean |
| **37** | CRFCWOad | Content word overlap, all sentences, proportional, standard deviation |
| **LSA** | | |
| **38** | LSASS1 | LSA overlap, adjacent sentences, mean |
| **39** | LSASS1d | LSA overlap, adjacent sentences, standard deviation |
| **40** | LSASSp | LSA overlap, all sentences in paragraph, mean |
| **41** | LSASSpd | LSA overlap, all sentences in paragraph, standard deviation |
| **42** | LSAPP1 | LSA overlap, adjacent paragraphs, mean |
| **43** | LSAPP1d | LSA overlap, adjacent paragraphs, standard deviation |
| **44** | LSAGN | LSA given/new, sentences, mean |
| **45** | LSAGNd | LSA given/new, sentences, standard deviation |
| **Lexical Diversity** | | |
| **46** | LDTTRc | Lexical diversity, type-token ratio, content word lemmas |
| **47** | LDTTRa | Lexical diversity, type-token ratio, all words |
| **48** | LDMTLDa | Lexical diversity, MTLD, all words |

| | | |
|---|---|---|
| **49** | LDVOCDa | Lexical diversity, VOCD, all words |
| **Connectives** | | |
| **50** | CNCAll | All connectives incidence |
| **51** | CNCCaus | Causal connectives incidence |
| **52** | CNCLogic | Logical connectives incidence |
| **53** | CNCADC | Adversative and contrastive connectives incidence |
| **54** | CNCTemp | Temporal connectives incidence |
| **55** | CNCTempx | Expanded temporal connectives incidence |
| **56** | CNCAdd | Additive connectives incidence |
| **57** | CNCPos | Positive connectives incidence |
| **58** | CNCNeg | Negative connectives incidence |
| **Situation Model** | | |
| **59** | SMCAUSv | Causal verb incidence |
| **60** | SMCAUSvp | Causal verbs and causal particles incidence |
| **61** | SMINTEp | Intentional verbs incidence |
| **62** | SMCAUSr | Ratio of casual particles to causal verbs |
| **63** | SMINTEr | Ratio of intentional particles to intentional verbs |
| **64** | SMCAUSlsa | LSA verb overlap |
| **65** | SMCAUSwn | WordNet verb overlap |
| **66** | SMTEMP | Temporal cohesion, tense and aspect repetition, mean |
| **Syntactic Complexity** | | |
| **67** | SYNLE | Left embeddedness, words before main verb, mean |
| **68** | SYNNP | Number of modifiers per noun phrase, mean |
| **69** | SYNMEDpos | Minimal Edit Distance, part of speech |
| **70** | SYNMEDwrd | Minimal Edit Distance, all words |
| **71** | SYNMEDlem | Minimal Edit Distance, lemmas |
| **72** | SYNSTRUTa | Sentence syntax similarity, adjacent sentences, mean. |
| **73** | SYNSTRUTt | Sentence syntax similarity, all combinations, across paragraphs, mean |
| **Syntactic Pattern Density** | | |
| **74** | DRNP | Noun phrase density, incidence |
| **75** | DRVP | Verb phrase density, incidence |
| **76** | DRAP | Adverbial phrase density, incidence |
| **77** | DRPP | Preposition phrase density, incidence |
| **78** | DRPVAL | Agentless passive voice density, incidence |
| **79** | DRNEG | Negation density, incidence |
| **80** | DRGERUND | Gerund density, incidence |
| **81** | DRINF | Infinitive density, incidence |
| **Word Information** | | |
| **82** | WRDNOUN | Noun incidence |
| **83** | WRDVERB | Verb incidence |
| **84** | WRDADJ | Adjective incidence |
| **85** | WRDADV | Adverb incidence |
| **86** | WRDPRO | Pronoun incidence |
| **87** | WRDPRP1s | First person singular pronoun incidence |
| **88** | WRDPRP1p | First person plural pronoun incidence |
| **89** | WRDPRP2 | Second person pronoun incidence |
| **90** | WRDPRP3s | Third person singular pronoun incidence |
| **91** | WRDPRP3p | Third person plural pronoun incidence |

| | | |
|---|---|---|
| **92** | WRDFRQc | CELEX word frequency for content words, mean |
| **93** | WRDFRQa | CELEX Log frequency for all words, mean |
| **94** | WRDFRQmc | CELEX Log minimum frequency for content words, mean |
| **95** | WRDAOAc | Age of acquisition for content words, mean |
| **96** | WRDFAMc | Familiarity for content words, mean |
| **97** | WRDCNCc | Concreteness for content words, mean |
| **98** | WRDIMGc | Imageability for content words, mean |
| **99** | WRDMEAc | Meaningfulness, Colorado norms, content words, mean |
| **100** | WRDPOLc | Polysemy for content words, mean |
| **101** | WRDHYPn | Hypernymy for nouns, mean |
| **102** | WRDHYPv | Hypernymy for verbs, mean |
| **103** | WRDHYPnv | Hypernymy for nouns and verbs, mean |
| **Readability** | | |
| **104** | RDFRE | Flesch Reading Ease |
| **105** | RDFKGL | Flesch-Kincaid Grade Level |
| **106** | RDL2 | Coh-Metrix L2 Readability |

## Appendix B

### Sample Nationality Confirmation Email in English

Huazhong Agricultural University
1st Shizishan Street
Hongshan District
Wuhan 430070
Hubei
People’s R China
January 10 2019

Dear Sir or Madam,

I am a student from Huazhong Agricultural University, Wuhan, China, expecting to graduate with a Bachelor’s degree this summer. And now I am designing a research for my graduation paper entitled “A Comparative Analysis of Linguistic Features of English Agricultural Journal Abstracts Written by American and Chinese Scientists”. This paper intends to makes a comparative analysis of the linguistic features of 600 English agricultural journal abstracts written by American and Chinese scientists with the purpose of offering guidance for agricultural scientists to write in a style closer to the agricultural research field and the native speakers so as to get their manuscripts better reviewed and more easily published.

What is now at the top of the agenda is the classification of cultural backgrounds. So we intend to employ and adjust the two criteria proposed by Wood (2001) which is the common practice of many studies. First, the first author is required to be affiliated with an institute within the country of classification. Second, the author’s name is required to be “typical” of the country of classification. While such a technique is by no means perfect. So, some nationality confirmation emails are sent to random authors in order to figure out whether the majority of abstracts will be appropriately categorized using these guidelines. All in all, I am making an impassioned plea for your kind help.

I’m trying to use in my research the abstract of the following article as part of my corpus:

***“Comparative Immunology of Allergic Responses”***

I’ve learnt that the first author of this article is Gershwin, Laurel. I am now writing to confirm whether the author’s nationality is American, in order to make sure the accuracy of American scientist corpus. I pledge that the information will be only used in this research.

I’d be very appreciated if you could reply soon. I am looking forward to hearing from you.

Thank you!
Sincerely yours,
Wenli Xu

## Appendix C

### Sample Nationality Confirmation Email in Chinese

尊敬的老师您好！

我是华中农业大学的一名大四学生，我正在为我的毕业论文做调查。论文题目是“中美学者农科英语学术期刊论文摘要的语言特征对比研究”。本研究收集了 600 篇分别来自中国作者和美国作者发表在农科英语学术期刊上的摘要，自建 2 个语料库，通过对比中美两国农科学者摘要的语言特征，旨在提高农科学者英语学术期刊论文摘要的写作水平，使之更容易获得国际学术界的认可。

本研究的一个亟待解决的问题是，如何确定中美作者的身份。由 Wood（2001）提出的作者身份确认标准是学术研究最常用的方法，本研究拟在此标准上做一些修改。具体来说，第一，第一作者来自中国大陆/美国科研院所；第二，第一作者的姓和名是中国大陆/美国普遍人名。第一作者同时满足以上两个条件的文章摘要即可收录进语料库。但这种方法也存在一定的不准确性，例如如何判断人名是否属于中国/美国普遍人名。因此，需要随机选择一些摘要发送邮件给通讯作者以确认第一作者国籍身份，由此判断此标准的准确率。在此，我诚恳地请求您的帮助。

我拟将 Hong Yang 老师的题为“Growth, digestive and absorptive capacity and antioxidant status in intestine and hepatopancreas of sub-adult grass carp Ctenopharyngodonidella fed graded levels of dietary threonine”的文章的摘要收录进中国作者摘要语料库中。从 Web of Science 上我了解到，Hong Yang 老师的通讯地址是中国，我想向您确认一下 Hong Yang 老师是否是中国国籍。我向您承诺，老师的国籍信息只会用于确保本次研究中国作者摘要语料库的准确性，绝不他用。

冒昧通信，期待您的回复！

祝：工作顺利！

学生: 徐雯丽
湖北省武汉市洪山区狮子山街道 1 号
华中农业大学

## Copyrights