A Corpus-driven Contrastive Study of the Top 100 Content Words in English and Chinese (中英最常用 100 个内容单词：一项基于语料库的对比研究)

Kang, Tingting
(康婷婷)
Lafayette College
(拉斐特学院)
kangt@lafayette.edu

Luo, Han
(骆涵)
Lafayette College
(拉斐特学院)
luoh@lafayette.edu

Abstract: This corpus-driven study examines the construction and frequency distribution of the top 100 most frequently used content words in American English, Chinese, and American and Chinese first-year students' compositions. First, this paper presents the top 100 most frequently used content words in American English and Chinese from two comparable corpora, i.e., the Corpus of Contemporary American English (COCA) and the Chinese National Corpus (CNC). Second, the top 100 most frequently used content words were drawn from two specific corpora consisting of American and Chinese freshmen's English compositions. Linguistic similarities and differences in terms of the usage of content words across the two sets of comparable corpora were identified. For example, the results showed that people from both American and Chinese cultural backgrounds relied heavily on verbs and nouns in their languages. However, Chinese people tended to prefer using direction-oriented nouns and food-related words, which were nearly absent in the COCA and American freshmen's compositions. The cultural implications associated with the linguistic similarities and differences are discussed and pedagogical implications of the findings are also offered.

提要：本文研究了当代美国英语语料库（COCA）、中国国家语料库（CNC）以及大一中美学生写作语料库中前 100 个最常用内容词的比例和分布，并详细分析了常用内容词在不同文化背景和语境中的相似和不同之处。结果表明，美国英语和汉语都大量使用动词和名词，但中国人更倾向于使用与方向相关的名词和与食物相关的内容词，而这在 COCA 和美国新生的作文中却很罕见。本文还讨论了中英内容词使用中的相似性和差异性的相关文化内涵，以及研究结果对教学实践的指导性意义。

Keywords: Chinese language, content words, contrastive analysis, corpus, culture, English language

关键词：中文、内容词、对比研究、语料库、文化、英语

1. Introduction

Anthropologist-linguist Edward Sapir (1929, 1966) argued that language is the perfect symbolic system to describe the content of every culture. Different cultures tend to conceptualize the world differently and such differences are reflected in language forms (Lakoff & Johnson, 1980; Langacker, 1987, 1990; Talmy, 2000). Not surprisingly, Benjamin Whorf (1956) proposed that studies on not only vocabulary but also grammatical structures, such as word-classes, lexical word inflection, and derivation, provide a window into the mind of people from different cultures.

Inspired by these theories on the relationships among language, thinking, and culture, also known as the Sapir-Whorf hypothesis, a large number of pioneering contrastive studies have been conducted to examine color terminologies (e.g., Berlin & Kay, 1991), contrastive rhetoric (e.g., Kaplan, 1966), space concepts (e.g., Brown, 1994), and metaphors (e.g., Lakoff & Johnson, 1980) in different languages. As a matter of fact, the Sapir-Whorf hypothesis was a well-discussed topic in the 1990s and 2000s (Regier & Xu, 2017), and the pendulum has begun to swing back in recent years (e.g., Cibelli, Xu, Austerweil, Griffiths, & Regier, 2016; Kadarisman, 2015; Neuliep, 2017; Tseng, Carstensen, Regier, & Xu, 2016; Wang, 2016). Most of these recent studies have further supported the close relationship between language and culture (i.e., the Sapir-Whorf hypothesis) by using sophisticated data analysis tools.

Moreover, with the advancement of corpus linguistic research, scholars have been able to conduct contrastive studies and compare different languages through massive linguistic data obtained from various corpora (McEnery, Xiao, & Tono, 2006). Due to the extent of similarities among alphabetic languages, existing corpus-based contrastive research, to date, has primarily focused on comparing such languages as Spanish, French, and Dutch to English (e.g., Butler, 2008; Defrancq & De Sutter, 2010; Gladkova, 2010). This is because it is easier to find formal or translational equivalents between English and other alphabetic languages, which is not the case for a non-alphabetic, character-based language like Chinese. In contrast, contrastive studies in English and Chinese are relatively more difficult to carry out, and thus corpus-driven research in this area is still rather meager despite increasing attention from a number of scholars (McEnery, Xiao, & Mo, 2003; Xiao & McEnery, 2005; Chung, 2008; Qian & Piao, 2009; Chen, 2010).

In particular, no corpus-based research has been conducted to examine the most frequently used content words in comparable English and Chinese corpora. This study attempts to fill in this research gap by analyzing the top 100 most frequently used content words as displayed in four corpora, i.e., the Corpus of Contemporary American English (COCA)[1], the Chinese National Corpus (CNC)[2], and two specific corpora of American and Chinese freshmen's English compositions.

[1] c.f. http://corpus.byu.edu/coca/.
[2] c.f. http://www.cncorpus.org/

2. Literature Review

2.1 Corpus-based Contrastive Studies

In the past few decades, corpus-based analysis has become an important method for comparing different languages by utilizing "a large and principled collection of natural texts" (Biber, Conrad, & Reppen, 1998, p. 4). With the development of English language corpora, large corpora have also become available in other languages, such as Spanish, French, German, Portuguese, Japanese, and Chinese (McEnery, Xiao, & Tono, 2006). Consequently, a good number of corpus-based contrastive studies have emerged with the aim of comparing different languages.

2.2 Corpus-based Contrastive Studies among Alphabetic Languages

In order to make linguistic features comparable across languages, words used in different languages that share similar parts of speech, meanings, and forms are most frequently examined in the field of lexical corpus-based contrastive studies. Therefore, contrastive linguistic studies among alphabetic languages are relatively easier to carry out due to the extent of similarities among these languages.

Hudson (1994) compared the percentage of nouns (i.e. common nouns, proper nouns, and pronouns) in the Brown and LOB (Lancaster-Oslo-Bergen) corpora across different genres and proposed a striking constancy that the noun ratios in written English were always between 33% and 42%. With regard to other word-classes and other languages, he concluded a trend that among written English, written Swedish, New Testament Greek, written Welsh, spoken English, and children's English, there was a negative relationship between prepositions/common nouns and verbs/pronouns.

The scope of the research subjects in contrastive corpus-based lexical studies has become narrower in recent years. For example, Butler (2008) focused on the idea, concept, and notion in English and their formal equivalents in Spanish, examining their frequencies, adjectival collocations, and idiomatic contractions within two comparable corpora, the British National Corpus (World Edition) and the Corpus del Español. The results indicated that, overall, there was a striking similarity between the use of idea, concept, and notion in English and Spanish, with some minor differences. Molina-Plaza and de Gregorio-Godeo (2010) analyzed the stretched verb collections with *give* in English and *dar* in Spanish. With regard to the different structures of verb collocations, they provided substantial pedagogical applications for the L2 learners of English and Spanish.

In addition to English and Spanish, other alphabetic languages have also been compared through corpus analysis. For instance, Gladkova (2010) explored the linguistic and cultural variations of *sympathy*, *compassion* and *empathy* in English and their translational equivalents in Russian words *socˇuvstvie*, *sostradanie*, and *soperezˇivanie*. By applying the natural semantic meta-language research method into this study, the researcher successfully explained the semantic and conceptual differences of using these emotional words in two cultures. Defrancq and De Sutter (2010) compared the contingency

hedges of English *depend*, French *dépendre* and Dutch *afhangen*, *liggen* and *zien*, and discovered some consistent linguistic features for the contingency hedges in English, French, and Dutch.

2.3 Corpus-based Contrastive Studies between English and Chinese

Despite less conceivable linguistic similarity between Chinese and alphabetic languages, corpus-based contrastive studies between English and Chinese have also started to catch up. These studies have focused on analyzing aspect markers, tenses, passive constructions, kinship terms, word metaphors, and borrowed words (e.g., Chen, 2010; Chung, 2008; Qian & Piao, 2009; Xiao & McEnery, 2005; Yu, Yu, & Lee, 2017).

In one of the first of its kind, Xiao and McEnery (2005) used a corpus of written British English, the Freiburg-Lancaster-Oslo-Bergen corpus (FLOB), and the Lancaster Corpus of Mandarin Chinese (LCMC) (i.e., a comparable corpus to FLOB), to identify some of the basic grammatical features across English and Chinese. The results showed that "English is predominantly a tense language, whereas Chinese is exclusively an aspect language" (Xiao & McEnery, 2005, p. 1). In other words, English marks tense and aspect, but there are no morphology-like devices in Chinese to mark tense, number, gender, or case but only aspect markers (e.g. *–le*, *-guo*, *-zai*, and *-zhe*) to represent differences in time and situation. By using the Chinese-English matched corpora, FLOB and LCMC, they further explored the aspect-marking differences between Chinese and British English and how British English aspect marking was translated into Chinese.

Moreover, McEnery, Xiao, and Mo (2003) demonstrated the differences and similarities of aspect markers among not only Chinese and British English but also American English by using the LCMC, Frown, and FLOB corpora. Along the same lines, Xiao, McEnery, and Qian (2006) compared the characteristics of passive constructions in British English (*be/get* + past participle) and Mandarin Chinese (*bei*/*jiao*/*rang*/*gei*) by using data of the FLOB, LCMC, and two other spoken corpora. Their findings insightfully demonstrated that passive constructions are more frequently used in English than in Chinese due to the unpleasant and undesirable semantic prosody in Chinese passives.

In addition to explorations of grammatical differences, cultural influences in word selection have also been a focus of discussion in English and Chinese corpus-based contrastive studies. Influenced by Lakoff and Johnson's (1980) conceptual metaphor theory, a number of corpus-based contrastive studies have been conducted (e.g., Chung, 2008; Chen, 2010; Qian & Piao, 2009) to examine to what degree types of metaphors and their collocations can mirror cultural similarities and differences. For example, Chen (2010) concluded that Chinese is a very typical metaphorical language that tends to link physical experience with various subjective notions due to the influence of Confucianism and Taoism. Qian and Piao (2009) compared kinship taggers in LCMC and FLOB and revealed a scheme of annotating Chinese kinship into LCMC, but due to the complex meanings of some Chinese kinship terms, tagging them all in LCMC was problematic. The two researchers explained that this is because the concept of family has always been an

important aspect of Chinese life and thus the Chinese language has a much richer cluster of words describing family relations than English does.

Finally, researchers have also used various types of learner corpora to compare the nature of English and Chinese language. For example, Tardif, Fletcher, Liang, Zhang, Kaciroti, and Marchman (2008) analyzed babies' first 10 words in their first language among English-, Mandarin- and Cantonese-speaking children. The findings showed that Chinese babies obtained more words than American babies and especially more words within the category of people terms, which echoes Qian and Piao's (2009) study. Chan (2010) looked at Chinese learners' English written errors to elaborate on the difference between English and Chinese. Based on data gathered from 387 ESL learners' free writings, Chan argued that mother tongue influence was the most important factor that leads to ESL learners' written errors.

2.4 Research Gaps & Research Questions

As discussed previously, although corpus-based contrastive studies between English and Chinese have started to emerge, the number is still rather meager partly due to the linguistic distance between Chinese and other alphabetic languages. In addition, most existing corpus-driven contrastive research comparing Chinese and English has mainly focused on specific lexical or grammatical features such as aspect markers, tenses, passive constructions, and kinship terms. Moreover, the majority of previous English and Chinese corpus-based contrastive studies tended to only address the linguistic differences surrounding certain features across the two languages, without further exploring the potential cultural connotations indicated by certain linguistic forms. As Aijmer, Altenberg, and Johansson (1996) noted, comparable corpora could possibly increase our knowledge of cultural differences in many different ways. It thus might be interesting and worthwhile to delve into the cultural explanations behind linguistic differences.

Notwithstanding many differences between English and Chinese, one linguistic similarity between the two languages is that morphologically both of them are analytical languages where lexical meanings are expressed by using separate words, so comparisons of content words across these two languages are practicable. In addition, analyzing content words expands the research from focusing on a single lexical or grammatical feature to larger numbers of lexical items. Moreover, content words, which are often used to convey intended messages, may be more appropriate for interpreting culture compared to other closed word groups, such as prepositions, pronouns, articles, and so forth. However, no research, thus far, has examined and compared the most frequently used content words in English and Chinese by looking at massive linguistic data gathered from comparable corpora.

To fill in these research gaps, this study attempts to expand the scope of earlier English and Chinese corpus-based contrastive investigations by exploring the 100 most frequent content words in American English, Chinese, and American and Chinese freshmen English compositions. It also aims to enhance knowledge of the interrelationships between language and culture. More specifically, the research questions for this study are:

(1) What are the 100 most frequent content words (i.e., noun, verb, adjective, and adverb) in American English, Chinese, and American and Chinese freshmen English compositions? (2) How are the 100 most frequent content words in American English, Chinese, and American and Chinese freshmen English compositions different or alike? The cultural implications of the research results will also be discussed whenever possible.

3. Method

3.1 Corpora Used in This Research

In order to extract the top 100 most frequently used content words in English and Chinese, this study selected four corpora: i.e., the Corpus of Contemporary American English (COCA), the Chinese National Corpus (CNC), and the two specific corpora consisting of American and Chinese freshmen's English compositions. This is because these four chosen corpora represent two main cultures (American & Chinese) and two subcultures (American first-year students & Chinese first-year students at college). In addition, including four instead of two related corpora could likely enhance the validity of this research and provide more concrete evidence to elaborate upon the relationships between language and culture. It should be noted that Chinese first-year students' compositions were written in their second language, i.e., English rather than their native language, Chinese. It is interesting to see to what degree Chinese students' first year ESL writing is influenced by their L1 and Chinese culture.

The COCA was the first large and diverse corpus of American English, and the CNC was the largest balanced corpus of Chinese. The COCA contained more than 450 million words from 1990-2012; the CNC provided a balanced collection of texts from 1919-2012. The corpus has nearly 100 million characters out of about 50 million characters are tagged. Even though the two corpora were different in size, both of them included not only written but also spoken resources in America and China. Furthermore, these two corpora contain similar text types. In the COCA, "texts are evenly divided between spoken (20%), fiction (20%), popular magazines (20%), newspapers (20%), and academic journals (20%)" (Davies, 2009). In the CNC, all the resources are in Chinese, and approximately 50% of the texts come from arts and social sciences (politics and law: philosophy, politics, religion, and law; history: history, archaeology, and nationalities; society: sociology, psychology, linguists, education, literary theory, news, and folk-customs; economy: industrial economy, agricultural economy, political economy, and economics of finance and trade; art: music, essay, biography, reportage, fiction, and spoken; military and sports; living), 30% from natural science (mathematics, biochemistry, astronomical geography, maritime meteorology, agriculture and forestry, and medical), and 20% from general fields (administrative documents, statutes, judicial documents, business proclamations, protocol speech, and expository writing). Additionally, the COCA had a function to search for the most frequently used words by part of speech, and the CNC website had a most frequently used word list annotated with part of speech.

The data of American and Chinese freshmen compositions were collected from the freshmen composition classes at an American public university. This course was a required

course for all first-year students, and it offered special sections for the international students. Therefore, the texts of American freshmen compositions were selected from 34 American students who were in the regular freshmen composition classes, and the texts of Chinese freshmen compositions were selected from the 32 Chinese students in the international students' composition sections. The minimum language requirement for them to take this course was to have internet-based TOEFL scores higher than 59 or to be currently placed in Level 5 in the intensive English program at the university. As listed in their course syllabus, all the students in the freshmen composition classes needed to complete six writing projects. The two writing texts that we chose to use in this study were a short informational argumentative essay and an extended argumentative essay. In total, this study included 68 writing texts from Americans and 64 writing texts from Chinese ESL students.

3.2 Procedures

The preliminary work in this study was to build up the top 100 most frequently used content word lists in the COCA, the CNC, and the American and Chinese freshmen compositions. For the top 100 content word list in American English, as mentioned in the previous section, the COCA website provides a search function for extracting words by part of speech and ranking them based on frequency. Therefore, the researchers searched for the top 100 most frequent nouns, verbs, adjectives, and adverbs individually to include in a master list of the 400 most frequently used content words. Then, this 400-word list was ranked by total raw frequencies (TOT). The next step was to clean data to make sure all the words that appear in this list belong to the appropriate content word categories. To define the content word categories, this research used the definition of noun, verb, adjective, and adverb in English in Longman Student Grammar of Spoken and Written English (Biber, Johansson, Leech, Conrad, & Finegan, 2002). Additionally, with the aim of including the words that carry meaningful information, the nouns in this research refer to both common nouns and proper nouns, and the verbs include lexical verbs, primary verbs (e.g. *be* and *have*), and auxiliary verbs (e.g., *can* and *will*). Also, it is worth noting that to capture the complexity of natural language and how language is used in real life, the frequency list was generated based on word tokens.

In the CNC, there was no feature as in the COCA that can search for word frequencies among different word classes, but on the CNC website, a most frequent word list, annotated with parts of speech, was available. Therefore, the researchers extracted the top 100 most frequent content words, including nouns, verbs, adjectives, and adverbs, from the master list.

To extract the top 100 content words in American and Chinese freshmen compositions, a free concordance program, AntConc 3.3.5 (Anthony, 2012), was used to count the word token frequencies among the texts from American and Chinese freshmen compositions. Each word's part of speech was labeled along with the rules that had been used in the content word list in the COCA. For the words that could have more than one part of speech, the original sentences that contained the target words were checked to mark the frequencies under appropriate content word category.

After establishing these four top 100 content word lists, their content word ratios and distribution were analyzed. First, the number of nouns, verbs, adjectives, and adverbs in each top 100 content word list was calculated. Second, the 100 ranks were further divided into 10 frequency bands to analyze their distribution. In other words, lexical items from rank 1 to 10 belong to frequency band 1, words appear from rank 11 to 20 have been grouped into frequency band 2, and the like. By doing so, the distribution of different content word classes across the four lists can be visually represented. Additionally, in order to obtain the relationships of content word distribution among the four lists, the statistical method, Spearman's rho, was utilized. Spearman's rho can range in value from -1 to +1. An absolute value of one indicates a perfect linear relationship and a value of zero indicates the absence of a linear relationship.

4. Results

4.1 Construction of the Top 100 Most Frequently Used Content Words

The top 100 most frequently used content words in American English, Chinese, and American and Chinese freshmen compositions are presented in Lists 1, 2, 3, and 4 in Appendix A. Table 1 summarizes the construction of the top 100 content words across these four corpora. Even though the noun, verb, adjective, and adverb ratios varied, some patterns and trends could be discovered.

First, the construction of the four lists was taken up mostly by nouns and verbs (total number of verb = 117; noun = 90). Second, Chinese speakers in the CNC and Chinese freshmen compositions corpora tended to use more nouns than English speakers and fewer verbs and adverbs. Third, when comparing the COCA and the CNC to the freshman compositions, there were more nouns and adjectives and fewer verbs and adverbs used in compositions than in general communication.

Table 1. Constructions of content words across American English, Chinese, and American and Chinese freshmen compositions

Corpora	Noun	Verb	Adjective	Adverb	Total
COCA	22	46	8	24	100
CNC	33	38	11	18	100
American freshmen compositions	35	33	16	16	100
Chinese freshmen compositions	47	29	13	11	100
Total	90	117	35	58	400

4.2 Frequency Distribution of the Top 100 Most Frequently Used Content Words

The results of content word distribution based on the 10 frequency bands are included in Table 2. The four tables in Appendix B demonstrated the correlation of the content words among four corpora. The absolute Spearman's rho scores greater than 0.5

were the verb distribution between American English and American freshmen writing (0.845), American English and Chinese noun distribution (0.599), and American English and Chinese freshmen compositions noun distribution (0.567). These strong correlations seem to indicate that in addition to cultural contexts, rhetorical contexts (i.e., general communication vs academic writing) also played an important role in the observed distinct content words distributions regardless of L1 or cultural backgrounds.

Table 2. Distribution of content words
(FB=frequency band; v=noun; v=verb; adj=adjective; adv=adverb)

		American English				Chinese				American freshmen compositions				Chinese freshmen compositions			
FB	Rank	n	v	adj	ad	n	v	adj	adv	n	v	adj	adv	n	v	adj	adv
1	1-10	0	10	0	0	3	3	0	4	1	8	0	1	3	6	0	1
2	11-20	2	6	0	2	3	4	1	2	1	4	2	3	5	3	1	1
3	21-30	0	5	1	4	0	7	2	1	3	4	1	2	4	2	4	0
4	31-40	2	2	1	5	6	3	1	0	5	3	1	1	4	2	0	4
5	41-50	0	9	1	0	2	6	1	1	4	4	1	1	4	3	1	2
6	51-60	3	5	0	2	1	2	2	5	5	3	1	1	6	2	2	0
7	61-70	4	2	0	4	6	1	1	2	5	1	3	1	5	4	0	1
8	71-80	5	2	1	2	4	5	0	1	5	1	2	2	5	4	1	0
9	81-90	5	3	0	2	5	3	0	2	2	2	3	3	5	2	2	1
10	91-100	1	2	4	3	3	4	3	0	4	3	2	1	6	1	2	1

5. Discussion

Unlike previous contrastive corpus-based research, which focused on one type of part of speech (e.g., Hudson, 1994) or specific lexical items (e.g., Butler, 2008), results discovered from this study provided a macro level of analysis on the use of content words in American English and Chinese.

First, the results showed that people from both American and Chinese cultures rely heavily on verbs and nouns in their languages. Specifically, American English and Chinese noun distribution based on the 10 frequency bands were quite similar, which may indicate that the two cultures conceptualize many aspects of the world in similar ways (Yu, 1995, 1998; Yu, Yu, & Lee, 2017). For example, the concept of "time" appeared among the top 100 most frequently used content words in these four corpora, which included words like *time*, *years*, *year*, and *day* in List 1, 年[year], 时[time], 现在[now], 月[month], and 时间[time] in List 2, *time*, *year*, and *years* in List 3, and *time* in List 4. Although this result may be subject to other interpretations, it seems that the value of time tends to be universal in both American and Chinese cultures (Lakoff & Johnson, 1980; Yu, 1998, 2012). The value of time in both cultures is also reflected in a number of proverbs in English and Chinese (e.g., *A stitch in time saves nine. Time and tide wait for no man. Time flies. Time will tell.* 时间就是金钱[Time is money]. 光阴似箭[Time flies like an arrow]. 岁月不待人[Time waits for no man]. 时间检验真理[Time will tell the truth]).

Substantial similarities in terms of frequently used content words were also discovered between Chinese and American freshmen's compositions. For example, the noun *parents* appeared in the top 100 content words in both of the freshmen composition

corpora. Combined with a close examination of the contents of first-year compositions, this result seems to indicate that parents still tended to play an active and important role in both groups of students' first year of college life. Other common nouns across the two corpora included *education*, *school*, and *college*, which were related to their student status. Interestingly, both groups of students frequently used words related to playing computer games, such as *games*, *video*, and *game* in List 3 for American students and *internet*, *web*, and *game* in List 4 for Chinese students, reflecting American and Chinese freshmen's common interests as peers irrespective of their different cultural backgrounds. These seemingly intuitive findings provide convincing evidence for one of the basic assumptions of the cognitive linguistic framework that language reflects human conceptualizations of world experiences (Lakoff & Johnson, 1980, 1999; Langacker, 1987, 1990, 2008).

Since language is a result of conceptualization (Lakoff, 1987; Talmy, 2000) and different cultures tend to perceive the world and human life experiences differently to various degrees (Yu, 2009), this study, not surprisingly, discovered a number of differences among the top 100 most frequently used content words between American English and Chinese. For instance, Chinese speakers in the CNC tended to use more nouns than English speakers and fewer verbs and adverbs. Interestingly, even in Chinese students' first year ESL writing, we found similar patterns when compared with American students' first-year compositions. It seems that Chinese students' first year ESL writing tended to be influenced by their L1. This difference between American and Chinese speakers in findings may be even traced back to Chinese and American people's different ways of conceptualizing nouns and verbs (Shu, Zhang, & Zhang, 2019). For example, Shen (2019) insightfully pointed out that compared to the English word class construction in which English nouns and verbs are two separate categories, nouns in Chinese constitute a superordinate category that includes the verb category, a view echoed by Wang's (2019) analysis of the conceptual spatialization of actions or activities in Chinese. In other words, according to Shen (2019) and Wang (2019), a noun-verb distinction should not be assumed in the study of Chinese grammar. Shen (2019) also provided an elaborate discussion of the cognitive and philosophical roots of this difference between English and Chinese.

In spite of some interesting similarities as discussed above, the top 100 content words used in American and Chinese freshmen's compositions also showed a number of culture-specific differences. For example, American freshmen frequently used nouns such as *age*, *alcohol*, *sex*, *drug*, *war*, *violence*, *health*, and *(stem) cell*, which were almost absent in Chinese international students' writings. The Chinese students who were studying abroad also used a large number of exclusive nouns which were less present in American students' writings. Take Chinese students' exclusive nouns that appeared in List 4 as an example. They included words such as *internet*, *money*, *food*, *guns*, *phone*, *English*, *language*, *right*, *Chinese*, *law*, *penalty*, *abortion*, *(cosmetic/plastic) surgery*, *euthanasia*, and so on. These differences are not surprising and do not seem to be too hard to explain as both sets of exclusive nouns reflect those aspects of life heavily discussed or experienced in American and Chinese students' respective cultures. Similarly, these findings corroborated nicely with the cognitive-linguistic notion of the human conceptualization of life experiences; thus, the language forms used to reflect these conceptualizations are culturally shaped (Lakoff & Johnson, 1980, 1999; Yu, 2009, 2017).

A closer examination of the results on the top 100 content words revealed some other interesting differences in the process of corpora analysis. For example, Chinese people tended to prefer using direction-related nouns than American speakers, such as words 中[middle], 上[up], 里[inside], 后[back], 下[under], and 内[inside] in List 2. Another difference revolved around the concept of "food," reflecting the significance of food in the Chinese culture. More specifically, the distinct verb, 吃[eat], appeared in the top 100 content words in the CNC; *food* and *foods* appeared in the top 100 most frequently used content words in the Chinese freshmen corpus, whereas none of the top 100 content words in the two corpora produced by Americans was related to food.

The past decade has witnessed a growing body of research on the relationship between language, culture, and cognition (Chen, 2010; Maalej & Yu, 2011; Yu, 2009, 2017). For example, Yu (2009) examined the Chinese word 心[xīn] and provided a cognitive linguistic study of the Chinese conceptualization of the heart, revealing that the word 心[xīn] covers the meanings of both "heart" and "mind" as understood in English. He further traced the roots of the conception of the heart in ancient Chinese philosophy and traditional Chinese medicine, arguing that a holistic view that sees the heart as the center of both emotions and thought lies at the core of Chinese thinking and culture. Inspired by this line of research, the authors of this study speculate that the above-mentioned differences with regard to the preference of using direction-related nouns in Chinese as well as the dense usage of words associated with food may also be explained culturally. In ancient China, people in dread of nature attempted to use certain hypotheses to explain various phenomena. One common superstition is that *center* [中], *up* [上], *north* [北], and *left* [左] symbolize unchallengeable power and nobility. For example, *China* [中国] is literally translated as "central country;" *emperor* [皇上] is literally translated as "royal up," and in Chinese architecture, the most exalted people should live in the north of an architectural complex. Traditionally, on a formal occasion, males should stand to the left of females, and the host always lets the most honorable guests sit on his/her left. For example, there is a four-character saying in Chinese, 虚左以待 (Sima, 91BC), which means "emptying my left seat to wait for my honorable guest." Chinese people's preference of using direction-related nouns may be traced back to these cultural traditions in ancient China. Similarly, the significance of food in Chinese culture also has a long history and its importance to Chinese culture is extensively manifested in the Chinese language. As the Chinese saying goes, 民以食为天, food is valued as highly as the sky in people's lives. People even ask others whether they have eaten to greet each other in Chinese daily life. For example, 你吃了吗?[Did you eat?] is equivalent to "How are you doing?" in English. All these linguistic examples show that food is an essential part of Chinese people's life and an important aspect of Chinese culture.

6. Conclusion

Supported by the Sapir-Whorf hypothesis, this study was conducted with the aim of discovering the linguistic and cultural regularities of the top 100 content words across

American English, Chinese, and American and Chinese freshmen compositions. A corpus-based method was used to analyze the relationship between language and culture. The results demonstrated that both similarities and differences in terms of frequently used content words between American English and Chinese and between American and Chinese freshmen compositions may be attributed to the cultural contexts in which speakers experience and conceptualize the world. The findings provided potential evidence for the interrelated relationship among language, culture, and cognition (Lakoff & Johnson, 1980; Yu, 2009).

This study has a number of limitations. First, the traditional word class classification (i.e., noun, verb, adjective, adverb) was adopted to define and categorize content words. However, as some linguists have pointed out, the noun-verb distinction may not apply to the Chinese language (Shen, 2019; Wang, 2019). Second, only a very limited number of freshmen compositions on two argumentative essays assignments from Chinese and American students were used for analysis. The content words used in these writings may be affected by the specific topics and thus may not be representative of word usage in freshmen's academic writing. Third, this study focused on a macro-level analysis of the construction and frequency distribution of the top 100 content words across four corpora, which made in-depth linguistic analyses of specific lexical or grammatical items impossible. Therefore, future corpus-based contrastive studies may need to reconsider the appropriateness of traditional word class labels such as nouns and verbs when it comes to languages drastically different from English (e.g., Chinese). It is also important for future studies to include a wider range of corpora to enhance representativeness. For example, it would be interesting to examine American English and Chinese corpora representing a variety of registers or subcultures, such as spoken, popular magazines, newspapers, and academic journals. Moreover, future contrastive studies may also identify specific lexical or grammatical items that are comparable in English and Chinese and conduct more in-depth linguistic and cultural analysis. Finally, this study focuses on the comparison of American English and Mandarin. Future studies may expand research along this line to other languages, such as Japanese, Korean, and Spanish, providing further evidence for the relationship between language, culture, and conceptualization.

A number of pedagogical applications can be drawn from this study on account of the linguistic and cultural similarities and differences discovered across American English and Chinese. First of all, since both American Chinese and English relied heavily on nouns and verbs, American language learners of Chinese would benefit from early instruction on frequently used nouns and verbs in Chinese. Second, since language, culture, and cognition are interrelated, Chinese instructors may encourage learners to compare and contrast Chinese and their native language and guide the students to trace the roots of the identified linguistic similarities and differences to the levels of culture and cognition. For example, explaining the cultural implications behind linguistic phenomena such as the preference of using direction-related nouns or the high frequency of food-related words in Chinese may not only help the students understand the language better, but also can potentially boost their motivation in learning Chinese language and culture. Finally, teaching culture is now considered an integral part of language instruction. Various methods and strategies have been explored to enhance students' intercultural knowledge. This study shows that teaching

culture through analyzing language can be a viable and effective channel as pervasive evidence has been established for the relationship between language and culture. It is thus important to develop students' awareness of linguistic differences between Chinese and English and seek cultural explanations for such differences.

References

Aijmer, K., Altenberg, B., & Johansson, M. (1996). *Languages in contrast*: *Papers from a symposium on text-based cross-linguistic studies*. Lund, Sweden: Lund University Press.

Anthony, L. 2012. *AntConc 3.3.5.* Retrieved from http://www.antlab.sci.waseda.ac.jp/software.html

Berlin, B., & Kay, P. (1991). *Basic color terms: Their universality and evolution*. Berkeley, CA: University of California Press.

Biber, D., Conrad, S., and Reppen, R. 1998. Co*rpus linguistics: Investigating language structure and use.* Cambridge, UK: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2002). *Longman student grammar of spoken and written English*. Harlow, England: Pearson Education Limited.

Brown, P. (1994). The INs and ONs of Tzeltal locative expressions: The semantics of static descriptions of location. *Linguistics, 32*, 743-790.

Butler, C. S. (2008). The very idea! A corpus-based comparison of idea, concept, and notion and their formal equivalents in Spanish. *Atlantis, 30*, 59-77.

Chan, A. (2010). Toward a taxonomy of written errors: Investigation into written errors of Hong Kong Cantonese ESL learners. *TESOL Quarterly, 44*, 295-319.

Chen, A. (2010). A conceptual understanding of bodily orientation in Mandarin: A quantitative corpus perspective. *Corpus Linguistics & Linguistic Theory, 6,* 1-28.

Chung, S. (2008). Cross-linguistic comparisons of the MARKET metaphors. *Corpus Linguistics & Linguistic Theory, 4*, 141-175.

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The Sapir-Whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PloS One, 11*(7), e0158725.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*, 159-190.

Defrancq, B., & De Sutter, G. (2010). Contingency hedges in Dutch, French and English: A corpus-based contrastive analysis of the language-internal and -external properties of English depend, French dépendre and Dutch afhangen, liggen and zien. *International Journal Of Corpus Linguistics, 15*, 183-213.

Gladkova, A. (2010). Sympathy, compassion, and empathy in English and Russian: A linguistic and cultural analysis. *Culture & Psychology, 16*, 267-285.

Hudson, R. (1994). About 37% of word-tokens are nouns. *Language, 70*, 331-339.

Kadarisman, A. E. (2015). Linguistic relativity, cultural relativity, and foreign language teaching. *TEFLIN Journal, 16*, 1-25.

Kaplan, R. B. (1966). Cultural thought patterns in intercultural education. *Language Learning, 16*, 1-20.
Lakoff, G. (1987). *Women, fire, and dangerous things.* Chicago, IL: University of Chicago Press.
Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York, NY: Basic Books.
Lakoff, G., & Johonson, M. (1980). *Metaphors we live by*. Chicago, IL: The University of Chicago Press.
Langacker, R. (1987). *Foundations of cognitive grammar* (Vol. 1). Stanford, CA: Stanford University Press.
Langacker, R. (1990). *Concept, image and symbol*. Berlin, Germany: Mouton de Gruyter.
Langacker, Ronald. (2008). *Cognitive grammar: A basic introduction.* Oxford, UK: Oxford University Press.
Maalej, Z. A., & Yu, N. (Eds.). (2011). *Embodiment via body parts: Studies from various languages and cultures* (Vol. 31). Amsterdam, the Netherlands and Philadelphia, PA: John Benjamins Publishing.
McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. New York, NY: Taylor & Francis.
McEnety, A., Xiao, Z., & Mo, L. (2003). Aspect marking in English and Chinese. *Literary and linguistic Computing, 18*, 361-378.
Molina-Plaza, S., & de Gregorio-Godeo, E. (2010). Stretched verb collocations with give: Their use and translation into Spanish using the BNC and CREA corpora. *Recall, 22*, 191-211.
Neuliep, J. W. (2017). Sapir-Whorf hypothesis. *The International Encyclopedia of Intercultural Communication*, 1-5.
Qian, Y., & Piao, S. (2009). The development of a semantic annotation scheme for Chinese kinship. *Corpora, 4*, 189-208.
Regier, T., & Xu, Y. (2017). The Sapir-Whorf hypothesis and inference under uncertainty. *Wiley Interdisciplinary Reviews: Cognitive Science, 8*, e1440.
Sapir, E. (1929). The study of language as a science. *Language*, *5*, 207-214.
Sapir, E. (1966). The status of linguistics and a science. In D. G. Mandelbaum (Ed.), *The selected writings of Edward Sapir in culture language and personality*. Berkeley, CA: University of California Press.
Shen, J. (2019). Types of negatives and the noun-verb distinction in English and Chinese. In D. Shu, H. Zhang, & L. Zhang (Eds.), *Cognitive linguistics and the study of Chinese* (pp. 121-155). Amsterdam, the Netherlands and Philadelphia, PA: John Benjamins Publishing.
Shu, D., Zhang, H., & Zhang, L. (Eds.) (2019*). Cognitive linguistics and the study of Chinese*. Amsterdam, the Netherlands and Philadelphia, PA: John Benjamins Publishing.
Sima, Q. (91BC). *Shi ji.* Retrieved from https://ctext.org/shiji/zhs. [司马迁. 91BC. 史记].
Talmy, L. (2000). *Toward a cognitive semantics* (Vol. II). Cambridge, MA: MIT Press.
Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology, 44*, 929–938.
Tseng, C., Carstensen, A. B., Regier, T., & Xu, Y. (2016). A computational investigation of the Sapir-Whorf hypothesis: The case of spatial relations. In A. Papafragou, D.

Grodner, & D. Mirman (Ed.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2231-2236). Austin, TX: Cognitive Science Society.
Wang, W. (2019). The conceptual spatialization of actions or activities in Chinese: The Adjective + Verb construction. In D. Shu, H. Zhang, & L. Zhang (Eds.), *Cognitive linguistics and the study of Chinese* (pp. 157-182). Amsterdam, the Netherlands and Philadelphia, PA: John Benjamins Publishing.
Wang, Y. (2016). Do we know more about Whorf? *International Journal of Applied Linguistics and English Literature, 5*, 215-223.
Whorf, B. (1956). Grammatical categories. In J. B. Carroll (Ed.), *Language, thought and reality* (pp. 87-101). Cambridge, MA: MIT Press.
Xiao, R., McEnery, T., & Qian, Y. (2006). Passive constructions in English and Chinese: A corpus-based contrastive study. *Languages in Contrast, 6*, 109-149.
Xiao, Z., & McEnery, A. M. (2005). A corpus-based approach to tense and aspect in English-Chinese translation. In W. Pan, H. Fu, X. Luo, M. Chase, & J. Walls (Eds.), T*ranslation and contrastive studies* (pp. 114-157). Shanghai, China: Shanghai Foreign Language Education Press.
Yu, N. (1995). Metaphorical expression of anger and happiness in English and Chinese. *Metaphor and Symbolic Activity, 10*, 59–92.
Yu, N. (1998). *The contemporary theory of metaphor: A perspective from Chinese* (Vol. 1). Amsterdam, the Netherlands and Philadelphia, PA: John Benjamins Publishing.
Yu, N. (2009). *The Chinese HEART in a cognitive perspective: Culture, body, and language* (Vol. 12). Berlin, Germany and New York, NY: Walter de Gruyter.
Yu, N. (2012). The metaphorical orientation of time in Chinese. *Journal of Pragmatics*, *44*, 1335-1354.
Yu, N. (2017). Life as opera: A cultural metaphor in Chinese. In Farzad Sharifian (Ed.), *Advances in cultural linguistics* (pp. 65-87). Singapore: Springer.
Yu, N., Yu, L., & Lee, Y. C. (2017). Primary metaphors: Importance as size and weight in a comparative perspective. *Metaphor and Symbol*, *32*, 231-249.

Appendix A

List 1: Top100 most frequent content words in COCA
(v=noun; v=verb; adj=adjective; adv=adverb; POS=part of speech; TOT=total frequency)

Rank	Word	POS	TOT	Rank	Word	POS	TOT
1	is	v	4210980	51	year	n	355417
2	was	v	3384970	52	should	v	354854
3	be	v	2118761	53	still	adv	341596
4	are	v	2104489	54	got	v	341321
5	have	v	2095904	55	made	v	337895
6	do	v	1520663	56	world	n	337050
7	had	v	1507568	57	take	v	332656
8	were	v	1240986	58	day	n	329131
9	has	v	1192469	59	'll	v	326576
10	said	v	1100532	60	too	adv	322122
11	would	v	1057713	61	life	n	319753
12	can	v	996271	62	come	v	311036
13	been	v	900791	63	when	adv	310792
14	so	adv	894227	64	really	adv	308855
15	will	v	862031	65	man	n	305588
16	just	adv	789921	66	never	adv	301090
17	people	n	787379	67	being	v	294906
18	did	v	772448	68	most	adv	280882
19	know	v	729773	69	school	n	277227
20	time	n	722079	70	Mr	n	276925
21	could	v	711598	71	president	n	274418
22	now	adv	695534	72	why	adv	272605
23	're	v	680528	73	right	adv	268966
24	think	v	636774	74	things	n	254785
25	how	adv	627139	75	state	n	253571
26	then	adv	623932	76	children	n	253054
27	other	adj	621657	77	house	n	252421
28	more	adv	594410	78	let	v	251330
29	get	v	585015	79	American	adj	243007
30	says	v	570281	80	might	v	239682
31	also	adv	537124	81	women	n	237129
32	going	v	535002	82	again	adv	237035
33	years	n	527554	83	percent	n	226447
34	new	adj	492596	84	where	adv	225492
35	see	v	482363	85	students	n	224843
36	here	adv	475701	86	family	n	220769
37	well	adv	472664	87	look	v	219273
38	way	n	464767	88	put	v	215548
39	very	adv	445333	89	work	n	215544

40	only	adv	429745	90	found	v	212226
41	'm	v	428957	91	thing	n	211525
42	go	v	423453	92	today	adv	210795
43	say	v	422223	93	great	adj	209705
44	make	v	410072	94	big	adj	207732
45	good	adj	409930	95	always	adv	207114
46	want	v	375134	96	old	adj	206748
47	does	v	367909	97	used	v	203493
48	've	v	366671	98	high	adj	202617
49	may	v	363636	99	came	v	202288
50	'd	v	356231	100	all	adv	201195

List 2: Top100 most frequent content words in CNC
(v=noun; v=verb; adj=adjective; adv=adverb; POS=part of speech; TOT=total frequency)

Rank	Word	Translation	POS	TOT	Rank	Word	Translation	POS	TOT
1	是	verb be	v	118382	51	研究	study/research	v	8627
2	有	have	v	53522	52	更	more	adv	8602
3	也	also/too	adv	47034	53	已	already	adv	8600
4	不	no/not	adv	46950	54	却	but	adv	8253
5	就	about/at once	adv	44145	55	再	again	adv	8199
6	中	middle	n	40105	56	最	the most	adv	7957
7	说	say	v	35047	57	主要	main	adj	7879
8	上	up	n	34850	58	不同	different	adj	7822
9	都	all	adv	34261	59	不是	verb be not	v	7765
10	人	people	n	33915	60	中国	China	n	7721
11	要	demand/want	v	27324	61	关系	relation	n	7715
12	又	again	adv	25682	62	人们	people	n	7702
13	来	come	v	25410	63	才	just	adv	7634
14	年	year	n	21818	64	作用	affect	n	7548
15	到	arrive/go to/reach	v	21665	65	现在	now	n	7527
16	还	still	adv	20735	66	已经	already	adv	7358
17	大	big	adj	20050	67	重要	important	adj	7135
18	时	time	n	17995	68	我国	our country	n	6948
19	里	inside	n	17774	69	情况	circumstance	n	6922
20	发展	develop	v	17307	70	知道	know	v	6773
21	很	very	adv	16774	71	出	out	v	6742
22	可以	can/may	v	16724	72	社会主义	socialism	n	6711
23	使	make	v	16470	73	做	do/make	v	6708
24	去	go	v	14914	74	必须	must	adv	6701
25	没有	don’t have	v	14544	75	人民	people	n	6669
26	为	become	v	14499	76	成	become	v	6592
27	能	can	v	13781	77	走	go/walk	v	6589

28	看	look	v	13755	78	月	month	n	6530
29	小	small	adj	12687	79	方面	aspect	n	6518
30	多	many	adj	12028	80	需要	need	v	6511
31	后	back/behind	n	12026	81	便	therefore	adv	6351
32	会	will/can	v	11782	82	出来	come out	v	6335
33	好	good	adj	11743	83	发生	happen	v	6315
34	社会	society	n	11461	84	水	water	n	6283
35	进行	carry on	v	11085	85	过程	process	n	6231
36	问题	question/problem	n	10899	86	只	only	adv	6142
37	下	under	n	10737	87	科学	science	n	6101
38	如	like/as	v	10312	88	方法	way	n	6098
39	国家	country	n	10138	89	叫	call/shout/name	v	6041
40	工作	job	n	9655	90	内	inside	n	6002
41	起来	rise up	v	9588	91	技术	technology	n	5978
42	生产	produce	v	9419	92	一般	common	adj	5928
43	可	can/very/but	v	9361	93	许多	many	adj	5904
44	就是	even/quite right	adv	9195	94	吃	eat	v	5893
45	新	new	adj	9157	95	具有	have	v	5870
46	用	use	v	9062	96	高	tall/high	adj	5864
47	想	think	v	9028	97	形成	form	v	5850
48	不能	can’t	v	8834	98	影响	influence	v	5786
49	生活	life	n	8694	99	时间	time	n	5736
50	经济	economy	n	8680	100	事	thing	n	5731

List 3: Top100 most frequent content words in American freshmen compositions (v=noun; v=verb; adj=adjective; adv=adverb; POS=part of speech; TOT=total frequency)

Rank	Word	POS	TOT	Rank	Word	POS	TOT
1	is	v	1995	51	take	v	122
2	are	v	1288	52	young	adj	118
3	be	v	1026	53	get	v	117
4	as	adv	758	54	sex	n	117
5	have	v	726	55	however	adv	117
6	was	v	462	56	any	adj	116
7	has	v	441	57	years	n	116
8	would	v	437	58	United States	n	111
9	can	v	417	59	cells	n	110
10	many	adj	358	60	video	n	110
11	people	n	353	61	world	n	110
12	will	v	348	62	still	adv	108
13	all	adv	291	63	parents	n	106
14	been	v	271	64	used	v	106
15	were	v	253	65	different	adj	105

16	do	v	250	66	violent	adj	105
17	other	adj	247	67	believe	v	103
18	water	n	245	68	same	adj	103
19	also	adv	242	69	school	n	103
20	only	adv	237	70	where	adv	103
21	should	v	222	71	animals	n	101
22	being	v	219	72	dental	adj	99
23	when	adv	217	73	drug	n	99
24	women	n	210	74	education	n	99
25	how	adv	201	75	stem	n	99
26	research	n	193	76	then	adv	99
27	age	n	190	77	American	adj	98
28	drinking	v	182	78	very	adv	95
29	may	v	181	79	why	adv	95
30	could	v	172	80	new	adj	93
31	time	n	165	81	become	v	92
32	just	adv	163	82	health	n	92
33	life	n	157	83	men	n	92
34	make	v	152	84	fact	n	91
35	children	n	146	85	use	n	90
36	government	n	145	86	game	n	89
37	some	adj	145	87	now	adv	89
38	games	n	143	88	war	n	89
39	help	v	142	89	good	adj	88
40	had	v	141	90	made	v	88
41	way	n	137	91	go	v	87
42	well	adv	136	92	high	adj	87
43	violence	n	133	93	did	v	86
44	students	n	132	94	example	n	84
45	cell	n	130	95	money	n	84
46	does	v	130	96	society	n	83
47	need	v	130	97	laws	n	82
48	year	n	126	98	often	adv	82
49	alcohol	n	123	99	person	n	82
50	child	n	122	100	able	adj	81

List 4: Top100 most frequent content words in Chinese freshmen compositions
(v=noun; v=verb; adj=adjective; adv=adverb; POS=part of speech; TOT=total frequency)

Rank	Words	POS	TOT	Rank	Words	POS	TOT
1	is	v	1855	51	information	n	132
2	people	n	1192	52	new	adj	132
3	can	v	943	53	According	v	131
4	are	v	895	54	been	v	131
5	have	v	718	55	technology	n	131

6	be	v	557	56	school	n	129
7	will	v	519	57	human	n	128
8	children	n	495	58	cannot	v	127
9	death	n	446	59	learn	v	127
10	some	adj	437	60	food	n	127
11	time	n	425	61	English	n	126
12	students	n	404	62	need	v	125
13	penalty	n	383	63	development	n	125
14	also	adv	341	64	find	v	124
15	has	v	333	65	Web	n	124
16	should	v	309	66	women	n	124
17	life	n	299	67	how	adv	124
18	do	v	284	68	language	n	123
19	parents	n	275	69	jobs	n	122
20	many	adj	275	70	game	n	120
21	use	n	266	71	college	n	118
22	other	adj	254	72	important	adj	118
23	online	adj	232	73	believe	v	114
24	think	v	215	74	get	v	114
25	make	v	210	75	problems	n	111
26	education	n	209	76	could	v	110
27	China	n	202	77	lot	n	110
28	world	n	187	78	environment	n	109
29	good	adj	186	79	right	n	107
30	when	adv	182	80	cell	n	106
31	part	n	180	81	become	v	105
32	all	adv	174	82	better	adj	105
33	Internet	n	167	83	want	v	103
34	was	v	163	84	just	adv	102
35	abortion	n	163	85	some	adj	102
36	very	adv	162	86	foods	n	100
37	countries	n	156	87	Chinese	adj	100
38	way	n	147	88	euthanasia	n	99
39	different	adj	147	89	young	adj	99
40	society	n	146	90	guns	n	98
41	abroad	adv	144	91	phone	n	98
42	know	v	143	92	shows	v	97
43	may	v	142	93	law	n	97
44	surgery	n	142	94	oil	n	96
45	only	adv	142	95	why	adv	95
46	however	adv	138	96	study	v	93
47	government	n	137	97	global	adj	93
48	public	adj	136	98	country	n	92
49	help	v	134	99	crime	n	92
50	money	n	134	100	years	n	92

Appendix B

Table 1: Spearman's rho scores of noun distribution across the top 100 content word lists in English, Chinese, and American and Chinese freshmen compositions (n1=noun distribution in List 1; n2=noun distribution in List 2; n3=noun distribution in List 3; n4=noun distribution in List 4)

	n1	n2	n3	n4
n1		0.599	0.426	0.567
n2			0.288	0.016
n3				0.340

Table 2: Spearman's rho scores of verb distribution across the top 100 content word lists in English, Chinese, and American and Chinese freshmen compositions (v1=verb distribution in List 1; v2=verb distribution in List 2; v3=verb distribution in List 3; v4=verb distribution in List 4)

	v1	v2	v3	v4
v1		0.220	0.845	0.322
v2			0.373	-0.112
v3				0.071

Table 3: Spearman's rho scores of adjective distribution across the top 100 content word lists in English, Chinese, and American and Chinese freshmen compositions (adj1=adjective distribution in List 1; adj2=adjective distribution in List 2; adj3=adjective distribution in List 3; adj4=adjective distribution in List 4)

	adj1	adj2	adj3	adj4
adj1		0.411	-0.137	0.244
adj2			-0.117	0.495
adj3				0.106

Table 4: Spearman's rho scores of adverb distribution across the top 100 content word lists in English, Chinese, and American and Chinese freshmen compositions (adv1=adverb distribution in List 1; adv2=adverb distribution in List 2; adv3=adverb distribution in List 3; adv4=adverb distribution in List 4)

	adv1	adv2	adv3	adv4
adv1		-0.492	-0.057	0.027
adv2			0.085	-0.357
adv3				-0.328

