A Corpus-Based Quantitative Study of Nominalizations across Chinese and British Media English

Ying Liu[1], Alex Chengyu Fang[1], and Naixing Wei[2]

[1]Department of Linguistics and Translation, City University of Hong Kong, Hong Kong
yliu227-c@my.cityu.edu.hk, acfang@cityu.edu.hk
[2]School of Foreign Languages, Beihang University, Beijing, China
nxwei@buaa.edu.cn

Abstract

This paper reports on a corpus-based quantitative study of the use of nominalizations across China English and British English in two comparable media corpora. In contrast to previous corpus-based studies of nominalizations, we start by using a syntactic approach and proceed with some methodological innovations incorporating large lexical databases and syntactically annotated corpora. The data show that there are significant differences in the use of nominalizations across these two English varieties. It is hoped that this research will offer useful insights on variations in nominalization across different English varieties and also on the understanding of the two English varieties in question.

1 Introduction

Nominalization can refer to "the process of forming a noun from some other word-class (e.g. *red* + *ness*) or the derivation of a noun phrase from an underlying clause (e.g. *Her answering of the letter*… from *She answered the letter*)" (Crystal, 1997: 260). It has been approached by scholars from various perspectives, covering aspects of its form, meaning and use, as in the traditional grammar (e.g. Quirk et al., 1985), generative grammar (e.g. Lees, 1960; Chomsky, 1970), functional grammar (e.g. Halliday, 1994; Eggins, 2004) and cognitive grammar (e.g. Langacker, 1991). Among other things, nominalization is of close relevance to language variation studies due to its function to distinguish a nominal and compressed style from a colloquial one (e.g. Biber, 1986; Greenbaum, 1988, etc). However, in spite of numerous theoretical discussions, nominalization has only been touched upon sporadically in corpus-based studies, with notable exceptions of Biber (1986), Biber et al. (1998, 1999) and Leech et al. (2009). Due to an overwhelming word-based approach and a reliance on suffixes for identification, only a limited scope of nominalizations has been included in previous corpus-based studies. In addition, although these studies have revealed the discriminatory power of nominalization in language use with regard to spoken and written registers and genres, there are few attempts to investigate the use of nominalization across different language varieties except Leech et al. (2009).

The research to be reported on in this paper attempts to bridge the afore-said gaps. It is exploratory and descriptive in nature and attempts to examine the cross-variety quantitative differences in the use of nominalizations across China English and British English in two comparable media corpora. Our study is different from previous corpus-based studies in several important respects. First, our study is not about variations of nominalization across registers and genres, but will explore variations across different English varieties, a different level of linguistic variation. The reason why we chose China English is that previous studies (Xu, 2008, 2010) have shown that there are frequent uses of nominalization in China English. British English is chosen as the base for comparison. Second, the present study will adopt a syntactic approach to nominalizations, an approach that has not been undertaken in previous corpus-based studies. We will explain this further in Section 3. Third, there are some methodological innovations in the identification and retrieval of nominalizations in this study. We will not rely on the suffix-based

28th Pacific Asia Conference on Language, Information and Computation pages 101–110

method. Instead, we will show how large lexical databases and syntactically annotated corpora can fruitfully complement each other in research into a syntactic feature which is not easily extracted from corpora.

The research questions that we intend to address are the following: (1) Are there any significant quantitative differences in the use of nominalizations across Chinese and British Media English? (2) In what way, if any, does Chinese Media English differ from British Media English in terms of the quantitative use of nominalizations? It is hoped that the current study will not only show whether or not these two varieties demonstrate any significant quantitative differences regarding this particular linguistic construction but also be able to suggest reasons for the differences we found.

This paper is organized as follows. We will review related work concerning corpus-based studies of nominalization in Section 2. Section 3 will describe our approach to nominalization in the present study. In Section 4, we will introduce the methodology including the corpora used and the procedures to retrieve nominalizations. Section 5 will present the quantitative findings, followed by some discussions in Section 6. Section 7 concludes this research with prospects for future work.

2 Related Work: Corpus-Based Studies of Nominalization

Most previous research of nominalization is theoretical in nature. Nominalization has so far not attracted wide-spread interests among corpus linguists. For the few previous corpus-based studies, focus has been on how its uses vary in different registers.

Chafe (1982) investigated the use of nominalizations in 9,911 words of informal spoken language (from dinner table conversations) and 12,368 words of formal written language (from academic papers). He has shown that nominalizations occur about 11 times more in the written language than in the spoken language. He further explained that such difference is due to the function of nominalization to integrate more information into fewer words which contributes to the integration and detachment of the written language in contrast to the fragmentation and involvement of the spoken language.

Biber (1986) investigated nominalizations (i.e. words ending in *-tion*, *-ment*, *-ness*, and *-ity*) in the *LOB Corpus* and the *London-Lund Corpus*. Nominalization is interpreted as having the function which "marks a highly abstract, nominal content and a highly learned style" (Biber, 1986: 395). It is found that nominalizations occur more often in written texts (e.g. official documents, academic prose, and editorial letters) but less in spoken texts (e.g. telephone and face-to-face conversations). Biber et al. (1998) have shown that the academic prose has a frequency of nominalizations (i.e. words ending in *-tion*/*-sion*, *-ment*, *-ness*, and *-ity*) almost four times larger than fiction and speech based on findings from the *Longman-Lancaster Corpus* and the *London-Lund Corpus* and concluded that nominalizations tend to occur more in more formal texts. Biber et al. (1999) investigated nominalizations (i.e. words ending in *-tion*, *-ity*, *-ism*, and *-ness*) in four registers (i.e. conversation, fiction, newspaper, and academic prose) in the *Longman Spoken and Written English Corpus*. They found that the frequency of nominalization grows sharply from conversation to fiction, newspaper language, and academic prose. They concluded that nominalization is a reliable indicator for register distinction.

Moreover, Leech et al. (2009) have examined the frequency of nominalizations ending in 12 suffixes in two different English varieties in four corpora (i.e. *LOB*, *Brown*, *FLOB* and *Frown*). They found that American English consistently uses more nominalizations across all four registers (i.e. press, general prose, learned and fiction) than British English. They therefore concluded that American English displays a more compressed style and a higher level of density of content than British English.

Despite many findings mentioned above, there are several areas where further research is necessary. First, previous empirical studies of nominalizations are overwhelmingly word-based. Yet it is clear that "nominalization is no mere substitute for a verb or an adjective. Instead, the use of a nominalized expression requires an entirely different organization of the whole sentence" (Downing and Locke, 2006: 461). This is exactly how nominalization can pack much information into a single noun phrase and contribute to the compressed and nominal style.

Second, they have been fairly limited in the

scope of nominalizations included due to the current practice of identifying nominalizations by searching suffixes. This suffix-based method seems rather random since there are usually no explanations why certain suffixes are chosen over others. Another important drawback of the suffix-based method is that nominalizations derived from verbs through conversion are left out. For example, deverbal nouns such as *increase* derived from the verb *increase* cannot be retrieved by the suffix-based method. Therefore, the existing corpus-based studies have so far only focused on nominalizations derived through suffixation although researchers are aware that nominalizations include those derived by means of both suffixation and conversion (e.g. Tyrkkö and Hiltunen, 2009; Biber and Gray, 2013).

Finally, till now, generalizations about how the uses of nominalizations vary across linguistic contexts have mostly based on their occurrences in registers and genres in British and/or American English. It is rare to find corpus-based studies of nominalizations across different English varieties.

Therefore, in the present study, we will adopt a syntactic approach and a different methodology to identify and retrieve nominalizations, and extend the scope of previous studies well beyond registers and genres to different English varieties. This will be discussed further in the following sections.

3 Our Approach: A Syntactic Approach to Nominalization

As already mentioned above, our concern will be with nominalization defined as a syntactic feature. Nominalization in this study refers to "a noun phrase such as *the quarrel over pay* which has a systematic correspondence with a clause structure and the noun head of such a phrase is normally related morphologically to a verb (i.e. a deverbal noun)" (Quirk et al., 1985:1288).

To be more specific, deverbal nouns refer to nouns that are produced by combining suffixes with verb bases (Quirk et al., 1985:1550) and nouns that are produced through the process of conversion (Quirk et al., 1985:1558). Thus, unlike previous corpus-based studies which only include nominalizations derived through suffixation, nominalizations in our study include both suffixed nominalizations (e.g. *his refusal to help*) and converted nominalizations (e.g. *the quarrel over pay*). As for the correspondence between a nominalization and a clause structure, it is stated that "the relation between a nominalization and a corresponding clause can be more or less explicit, according to how far the nominalization specifies, through modifiers and determinatives, the nominal or adverbial elements of a corresponding clause" (Quirk et al., 1985:1289). For example, sentence [1] can have the following nominalizations:

[1] *The reviewers criticized his play in a hostile manner*.
[1a] *the reviewers' hostile criticizing of his play*
[1b] *the reviewers' hostile criticism of his play*
[1c] *the reviewers' criticism of his play*
[1d] *the reviewers' criticism*
[1e] *their criticism*
[1f] *the criticism*
[1g] *criticism*

(Quirk et al., 1985:1289)

According to Quirk et al. (1985:1289), the above noun phrases are "ordered from the most explicit [1a] to the extreme of inexplicitness [1g] but each of them could occupy the function of a nominalization". We therefore will consider the correspondence between a nominalization and a clause structure as on a continuum, being explicit or implicit, and all the above constructions from [1a] to [1g] will be taken into account in this study.

With nominalizations defined as syntactic structures, we will then turn to the methodology to retrieve them from corpora in the following section.

4 Methodology

4.1 Corpora

The data for our study were drawn from two comparable corpora, namely, the *Chinese Media English Corpus* (Henceforth CMEC) and the *British Media English Corpus* (Henceforth BMEC) (Fang et al., 2012). The two media corpora, with about one million words each, are of the same design and structure and consist of about 2,000 texts sampled from three mediums, namely, newspaper, magazine and the Internet. The texts of various topics are sampled from specially allotted separate sections in the three mediums. The five categories in CMEC and BMEC are: arts and culture, business, editorial, news report, and social

life. Arts and culture is largely concerned with topics of fine arts and cultural heritage. Business is about commerce, finance or economics. Editorial is "a lengthy opinion piece that provides the official view of the newspaper on particular issues" (Semino, 2009: 442) while news report is "a relatively short piece which consists of a 'factual' account of events that have occurred since the last edition of the newspaper" (Semino, 2009: 441). Social life is primarily associated with the topics of lifestyle and leisure. As can be seen, the five categories differ in various topics and so we would predict that there will be systematic differences in the uses of nominalizations.

Although the overall size of CMEC and BMEC is only about one million word tokens, the major advantage of the two corpora is the fact that they are comparable in design which allows for direct comparison between the two. The summary statistics of the two corpora is shown in Table 1.

	CMEC		*BMEC*	
Category	*Texts*	*Tokens*	*Texts*	*Tokens*
Arts&culture	451	200,464	430	205,353
Business	434	200,110	366	193,162
Editorial	371	200,456	314	196,910
News report	457	203,449	374	198,834
Social life	513	199,144	395	196,053
Total	2,226	1,003,623	1,879	990,312

Table 1. Basic Statistics of CMEC and BMEC

4.2 Retrieval of Nominalizations

In line with the definition of nominalization mentioned in Section 3, the extraction of nominalizations is operationalized in three steps as shown in Figure 1: (1) to parse the raw CMEC and BMEC; (2) to generate a list of deverbal nouns that function as the noun head of nominalizations; (3) to extract all noun phrases headed by these deverbal nouns from the parsed CMEC and BMEC.

For the first step, CMEC and BMEC have been parsed by the Stanford Parser [1] (Version 3.2.0; Klein and Manning, 2003). The Stanford parser is trained on the *Penn Treebank Corpus* (Marcus et al., 1993) and uses the Penn Treebank POS tagset (Santorini, 1990) and syntactic tagset (Santorini et al., 1991). Its parsing accuracy in terms of F1 score is reported to have reached 90.4% (Socher et al., 2013).

[1] See http://nlp.stanford.edu/software/lex-parser.shtml.

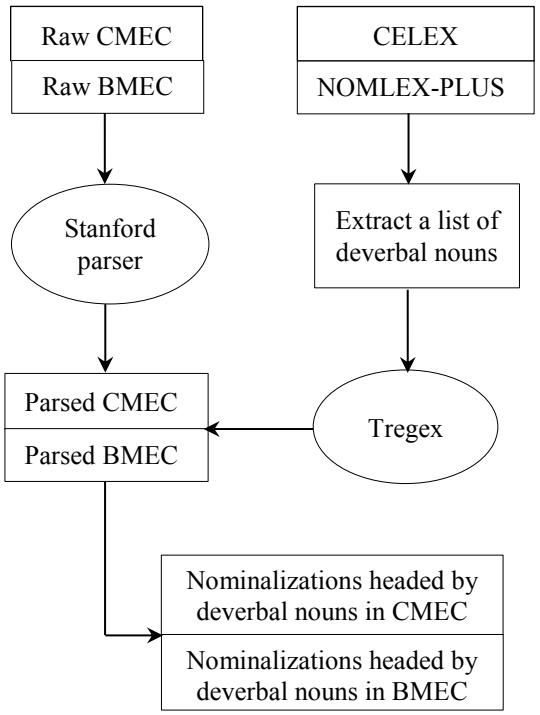

Figure 1. Flow Chart to Retrieve Nominalizations

As has been discussed in Section 2, the suffix-based method to identify nominalizations has certain drawbacks. Thus, in the second step, we adopted a wordlist method which uses lexical databases to extract deverbal nouns. Based on a survey of existing large lexical databases, CELEX and NOMLEX-PLUS which have derivational morphology information were used in this study. *CELEX English Lexical Database* (Baayen et al., 1995) consists of 52,447 lemmas [2] (or 160,595 word forms) which are extracted from *Oxford Advanced Learner's Dictionary* (1974) and *Longman Dictionary of Contemporary English* (1978). NOMLEX-PLUS (Meyers, 2007) is an extension of NOMLEX (Macleod et al., 1998), a dictionary of English deverbal nouns. In addition to NOMLEX, another source for deverbal nouns in NOMLEX-PLUS is COMLEX Syntax (Grishman et al., 1994), a dictionary annotated with rich syntactic information for nouns, adjectives and verbs. Deverbal nouns ending in *ing* in NOMLEX-PLUS were excluded in this study because their POS tagging as nouns is based on their usage in a specific corpus and their noun status is subject to

[2] Although CELEX-lemmas are not extracted from corpus, they cover 92% of the 17.9-million-word COBUILD corpus.

change elsewhere. In total, we extracted 5,538 deverbal noun lemmas derived by means of both suffixation and conversion from CELEX and NOMLEX-PLUS, which account for 27.64% of all noun tokens in CMEC and 29.15% in BMEC [3].

The last step was facilitated with Tregex [4] (version 3.2.0; Levy and Andrew, 2006), which is a tree query tool for matching patterns in trees. Tregex contains the main functionality of TGrep2 (Rohde, 2005) and adds a few more relations for syntactic trees such as dominance, precedence, and headship which are perfectly useful for our research purpose. We successively went through the syntactically parsed CMEC and BMEC and retrieved those nominalized structures headed by the deverbal nouns in our list.

Following the method outlined above, 66,850 nominalizations from CMEC and 65,104 nominalizations from BMEC were retrieved. The summary statistics is presented in Table 2.

Category	*#CMEC*	*#BMEC*
Arts & culture	10,938	11,295
Business	16,061	15,126
Editorial	15,220	14,061
News report	13,862	13,580
Social life	10,769	11,042
Total	66,850	65,104

Table 2. Summary Statistics of Retrieved Nominalizations from CMEC and BMEC

An example of the extracted nominalization headed by *development* from CMEC is shown in Figure 2.

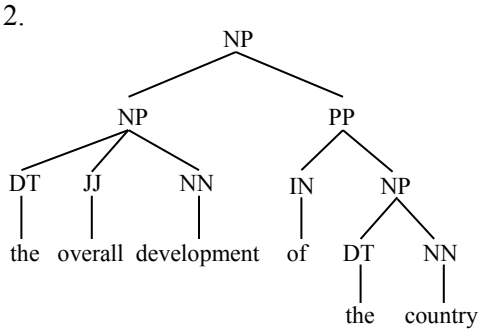

Figure 2. An Example of Retrieved Nominalizations (from c_m_ed_bjr_021.txt.prd)

[3] We admit that our deverbal noun wordlist is not a complete one. In fact, no such a complete list exists. But the deverbal noun is only one kind of nouns. Considering its coverage, we claim that nominalizations extracted in terms of our list are sufficient for our research purpose.

[4] See http://nlp.stanford.edu/software/tregex.shtml.

5 Results

5.1 Frequency and Distribution of All Nominalizations across CME and BME

Figure 3 gives a barplot representation of the mean frequencies of nominalizations across CME and BME and the five categories. Relative frequencies of nominalizations were calculated per 1,000 words in order to make comparisons of texts of diverse lengths possible. For statistical testing, we computed the relative frequency of nominalizations per 1,000 words for each text in CMEC and BMEC. Then an independent sample *t*-test was run to determine whether significant differences exist in the mean nominalization frequencies. The *t*-test results are presented in detail in Table 3.

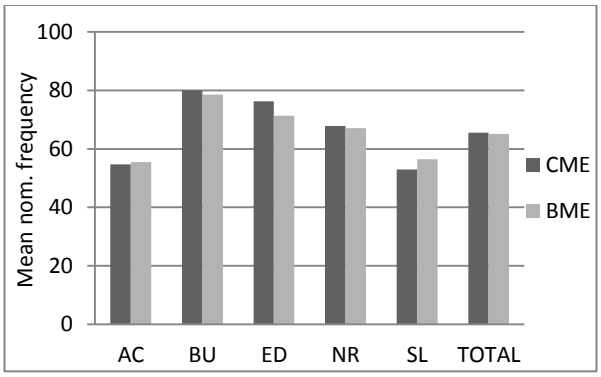

Figure 3. Barplots of Mean Nominalization Frequencies across CME and BME

As can be seen in Figure 3, the mean nominalization frequency for the overall CME (M=65.52) is a little higher than that for BME (M=65.14). But the *t*-test result shows that there is no significant difference in the uses of nominalizations in the overall CME and BME (t=0.578, p=0.563). With regard to the five categories, we can see from Figure 3 that the mean values in business (M=80.02), editorial (M=76.23), and news report (M=67.83) in CME are also higher than those in BME. However, the *t*-test result shows that only the difference in editorial is statistically significant (t=3.668, p=0.000), indicating that there are more uses of nominalizations in editorial in CME than BME. As for arts and culture and social life, the mean frequencies for BME look higher than those for

CME, but we only find statistically significant difference in social life (t=-2.964, p=0.003).

Category	*Variety*	*N. of Texts*	*Mean*	*Std. D*	*T*	*df*	*p-value*
Arts & culture	CME	451	54.71	18.39	-.634	870.583	.526
	BME	430	55.44	15.88			
Business	CME	434	80.02	23.17	1.006	780.582	.315
	BME	366	78.60	16.77			
Editorial*	CME	371	76.23	18.94	3.668	682.965	.000*
	BME	314	71.34	15.91			
News report	CME	457	67.83	22.45	.538	826.300	.591
	BME	374	67.05	19.45			
Social life*	CME	513	52.98	18.30	-2.964	906	.003*
	BME	395	56.51	17.10			
Overall corpus	CME	2226	65.52	23.11	.578	4102.285	.563
	BME	1879	65.14	19.25			

Note: * indicates a statistically significant difference ($p < 0.05$).

Table 3. Results of t-test of Mean Nominalization Frequency across CME and BME

To sum up, in terms of the mean frequencies, there are significantly more uses of nominalizations in CME in editorials, whilst BME uses significantly more nominalizations in social life than CME. Before we draw tentative conclusions, we will investigate the frequencies and distributions of suffixed nominalizations and converted nominalizations respectively.

5.2 Frequency and Distribution of Suffixed Nominalizations across CME and BME

This section shows the frequency and distribution of suffixed nominalizations (e.g. *his refusal to help*). The barplots are shown in Figure 4, and t-test results are presented in Table 4.

In terms of the overall corpus, it is observed from Figure 4 that the mean suffixed nominalization frequency for CME (M=32.58) is higher than that for BME (M=29.03). The barplot representation indicates that there are more uses of suffixed nominalizations in the overall CME than BME, and this is confirmed by the t-test result (see Table 4) which suggests that the difference in CME and BME is statistically significant (t=8.121, p=0.000). Interestingly, a higher mean score for CME can also be consistently seen in all the five categories although it is not so evident in social life. The t-test results show that there are significantly more uses of suffixed nominalizations in CME in arts and culture (t=2.903, p=0.004), business (t=5.625, p=0.000), editorial (t=7.167, p=0.000), and news report (t=3.393, p=0.001).

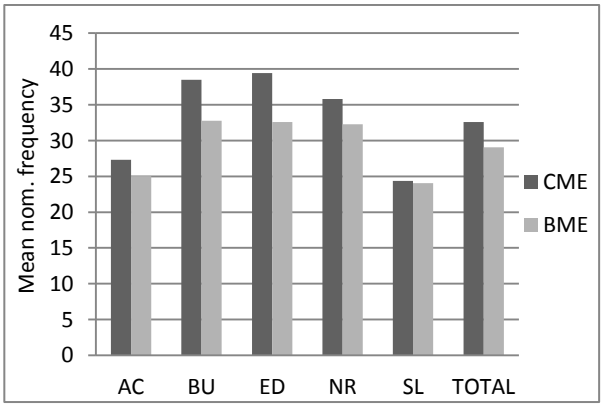

Figure 4. Barplots of Mean Suffixed Nominalization Frequencies across CME and BME

However, nominalizations are slightly more frequent but not significantly so in social life in CME (t=0.431, p=0.666). One possible interpretation is that there are few nominalizations used in social life in both CME and BME because as we previously mentioned in Section 4.1, texts in social life in CMEC and BMEC are often concerned with more casual topics such as lifestyles and leisure. We can see from Table 4 that the mean suffixed nominalization frequency in social life in CME (M=24.38) is the lowest among all the five categories (27.33 for arts and culture, 38.48 for business, 39.43 for editorial, and 35.81 for news report). The same is true for the mean frequency of suffixed nominalization in social life

in BME. Therefore, the lowest frequency of suffixed nominalization in social life might have resulted in the insignificant difference in the two English varieties.

Category	*Variety*	*N. of Texts*	*Mean*	*Std. D*	*T*	*df*	*p-value*
Arts & culture*	CME	451	27.33	13.11	2.903	841.441	.004*
	BME	430	25.05	10.07			
Business*	CME	434	38.48	16.82	5.625	775.448	.000*
	BME	366	32.75	11.90			
Editorial*	CME	371	39.43	14.60	7.167	663.180	.000*
	BME	314	32.57	10.34			
News report*	CME	457	35.81	16.64	3.393	828.903	.001*
	BME	374	32.27	13.47			
Social life	CME	513	24.38	11.41	.431	906	.666
	BME	395	24.04	11.48			
Overall corpus*	CME	2226	32.58	15.81	8.121	4069.139	.000*
	BME	1879	29.03	12.16			

Note: * indicates a statistically significant difference ($p < 0.05$).

Table 4. Results of *t*-test of Mean Suffixed Nominalization Frequency across CME and BME

Previous studies (e.g. Biber, 1986; Biber et al., 1998, 1999) have shown that suffixed nominalizations tend to occur in texts which convey highly abstract information and mark a formal and nominal style. The findings that there are significantly more uses of suffixed nominalizations in CME and also in its categories (except social life) suggest that CME adopts a more formal style in media English writing than BME.

A closer observation of the data reveals that this nominal style in CME has been in use to differing extents in various categories and it is particularly more prominent in business and editorial. When we look at the distribution of suffixed nominalizations across the five categories in CME, we can see a clear descending order for their mean frequencies: editorial (M=39.43) > business (M=38.48) > news report (M=35.81) > arts and culture (M=27.33) > social life (M=24.38), but the categories in BME are not so sharply differentiated since the mean suffixed nominalization frequencies are similar for business (M=32.75), editorial (M=32.57), and news report (M=32.27). In addition, it can be seen that the more suffixed nominalizations a category in CME uses, the larger difference in the uses of nominalizations across CME and BME is. We can also find a descending order for the mean differences in CME and BME: $D_{editorial}$ (6.86 per 1,000 words) > $D_{business}$ (5.73 per 1,000 words) > $D_{news\ report}$ (3.54 per 1,000 words) > $D_{arts\ and\ culture}$ (2.28 per 1,000 words) > $D_{social\ life}$ (0.34 per 1,000 words). This suggests that differences in the two English varieties are the most salient in categories dealing with more serious topics, but smaller in those concerned with casual topics.

5.3 Frequency and Distribution of Converted Nominalizations across CME and BME

In this section, we look at the frequency and distribution of converted nominalizations (e.g. *the quarrel over pay)*. Barplots representation and *t*-test results are presented in Figure 5 and Table 5 respectively.

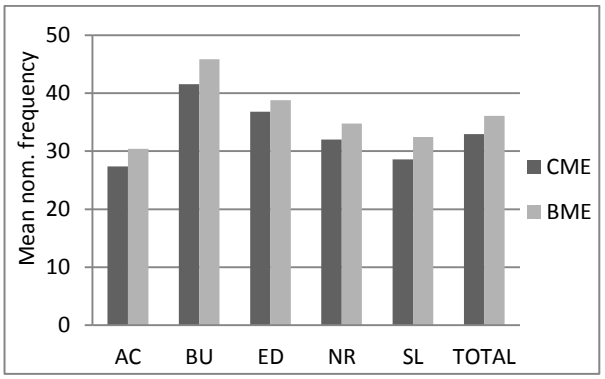

Figure 5. Barplots of Mean Converted Nominalization Frequencies across CME and BME

Figure 5 shows that the mean frequency for the overall BME (M=36.11) is higher than that for CME (M=32.95), indicating that BME has more

uses of converted nominalizations. The *t*-test result confirms that this difference is statistically significant (t=-7.431, p=0.000). Furthermore, the mean values for all the five categories in BME are consistently higher than those of CME as shown in Figure 5. The *t*-test results in Table 5 confirm that all the five categories in BME use significantly more converted nominalizations than those in CME.

We have previously shown that for suffixed nominalizations, differences across CME and BME are sharper in categories dealing with more serious topics but smaller in those concerned with casual topics. However, this does not hold true for converted nominalizations. As can be seen from Table 5, the mean frequency differences in the five categories across CME and BME show a similar tendency. For example, BME has 3.01 more occurrences of nominalization per 1,000 words in arts and culture than in the case of arts and culture in CME, and it has 1.97 more occurrences of nominalization per 1,000 words in editorial than in the case of editorial in CME. Also, the mean difference in news report is 2.76 per 1,000 words, which is similar to that for arts and culture.

Category	*Variety*	*N. of Texts*	*Mean*	*Std. D*	*T*	*df*	*p-value*
Arts & culture*	CME	451	27.38	12.18	-3.716	879	.000*
	BME	430	30.39	11.87			
Business*	CME	434	41.54	17.09	-4.111	782.119	.000*
	BME	366	45.85	12.46			
Editorial*	CME	371	36.80	11.34	-2.291	683	.022*
	BME	314	38.77	11.04			
News report*	CME	457	32.02	12.84	-3.273	822.781	.001*
	BME	374	34.78	11.46			
Social life*	CME	513	28.60	12.78	-4.806	887.657	.000*
	BME	395	32.46	11.36			
Overall corpus*	CME	2226	32.95	14.40	-7.431	4088.963	.000*
	BME	1879	36.11	12.89			

Note: * indicates a statistically significant difference ($p < 0.05$).

Table 5. Results of *t*-test of Mean Converted Nominalization Frequency across CME and BME

6 Discussion

Our data provide a clear indication that there are significant quantitative differences in the uses of nominalizations across CME and BME, and such differences across the two varieties are far sharper in terms of the suffixed and converted nominalizations than in terms of nominalizations as an overall group.

With regard to suffixed nominalizations, CME uses significantly more nominalizations than BME overall and also across the five categories (except social life), indicating that CME tends to be more nominal and formal compared to BME. We also have found that this nominal style has been in use to differing extents in various categories and becomes even more evident in those dealing with more serious topics such as business. But there is no such sharp differentiation across categories in BME. Moreover, differences across CME and BME are sharper in categories dealing with more serious topics than those concerned with casual topics. According to Collins and Yao (2013), a number of quantitative differences across English varieties have a stylistic basis. We may reason that variations in the uses of nominalizations found in this study may be ascribed to the English users' particular consciousness of stylistic formality in Media English writing in China. This consciousness can be tentatively attributed to certain social factors. Unlike the status of institutionalized varieties such as Indian English, English is not an official or second language in China and not widely used in intra-national communication. CME, as an edited register of China English, is specialized in its target readership. Texts in CMEC are all sampled from the leading media in China such as *China Daily* and *Beijing Review* (Fang et al., 2012) which serve as a key source for information concerning China

for overseas media as well as well-educated people at home and abroad. Moreover, it also provides learning materials for English learners in China as "many schools subscribe to *China Daily* and *Beijing Review* for their students and teaching staff" (Zhao and Campbell, 1995). Text writers in CME, mostly non-native users of English, who are aware of such informational purpose and the specialization of readership, are particularly careful with a formal style of Media English, especially in categories concerned with more serious topics.

However, British English is found to be influenced by the process of colloquialization, a stylistic shift which has brought many grammatical changes in English (Biber, 2003; Leech et al., 2009). The trend of colloquialization has made written genres more like spoken ones, and this has also manifested itself in the fewer uses of suffixed nominalizations in BME, as shown in this study.

As for the fewer uses of converted nominalizations in CME, one possible interpretation might be that English users in China are not so familiar with the usage of converted nominalizations as native speakers of English. Converted nominalization is not derived through a productive derivational rule that can be easily generalized to other word by the adding of suffixes to word bases. Instead, conversion requires a fairly large amount of lexical knowledge which non-native users of English may not have possessed, compared to native speakers of English. Another possible reason is that converted nominalizations might be associated with informality of writing and may occur more often in informal texts and less in formal texts. This is why there are fewer uses of them in the more nominal and formal CME, but more in the less formal BME.

The factors which may account for the variations in using nominalizations are of different types. What we have sketched above is only tentative and warrants further investigation.

7 Conclusions and Future Work

This paper has presented a corpus-based quantitative account of nominalizations across CME and BME. It has been observed that, for nominalization as a group, there is no significant difference in the overall CME and BME and significant differences have been found only in categories of editorial and social life. With regard to suffixed nominalizations, we have found that CME uses significantly more nominalizations than BME overall and also across the five categories (except social life), indicating a more nominal and formal style in CME. In terms of converted nominalizations, BME has significantly more uses of nominalizations overall and in all the five categories, which might have something to do with Chinese English users' ability of using converted nominalizations and the possible association between converted nominalizations and informality of writing.

Needless to say, quantitative evidence in the present study is not sufficient to describe the differences in the uses of nominalizations between China English and British English, but it nevertheless forms a practical starting-point for further research. The initial quantitative findings merit a more in-depth exploration into the uses of nominalizations in terms of the lexical patterns and syntactic structures in the future, which might offer more useful insights on variations in nominalization across different English varieties and also on the understanding of the two English varieties in question.

Acknowledgements

Research described in this paper was supported in part by grant received from the General Research Fund of the University Grant Council of the Hong Kong Special Administrative Region, China (Project no. CityU 142711) and City University of Hong Kong (Project nos. 6354005, 7004091, 9610283, 7002793, 9610226, 9041694, and 9610188).

References

Baayen, R. H, Piepenbrock, R. and Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Biber, D. (1986). Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. *Language*, 62(2): 384-414.

Biber, D. (2003). Compressed Noun Phrases in Newspaper Discourse: The Competing Demands of Popularization vs. Economy. In J. Aitchison and D. Lewis (Eds.). *New Media Language*. London: Routledge, pp. 169-181.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Language Use*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

Biber, D., and Gray, B. (2013). Nominalizing the Verb Phrase in Academic Science Writing. In B. Aarts, J. Close, G. Leech and S. Wallis (Eds.). *The Verb Phrase in English: Investigating Recent Language Change with Corpora*. Cambridge: Cambridge University Press, pp. 99-132,

Chafe, W. L. (1982). Integration and Involvement in Speaking, Writing, and Oral Literature. In D. Tannen (Ed.). *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, N. J.: Ablex, pp. 35-54.

Chomsky, N. (1970). Remarks on Nominalization. In R. A. Jacobs and P. S. Rosenbaum (Eds.). *Readings in English Transformational Grammar*. Waltham, Massachusetts: Ginn, pp. 184-221.

Collins, P. and Yao, X. Y. (2013). Colloquial Features in World Englishes. *International Journal of Corpus Linguistics*, 18 (4): 479-505.

Crystal, David. (1997). *A Dictionary of Linguistics and Phonetics (Fourth Edition)*. Oxford: Blackwell Publishers Ltd.

Downing, A. and Locke, P. (2006). *English Grammar: A University Course (Second Edition)*. London: Routledge.

Eggins, Suzanne. (2004). *An Introduction to Systemic Functional Linguistics (Second Edition)*. New York/ London: Continuum.

Fang, A. C., Le, F. and Cao, J. (2012). The Design, Establish and Primary Study of a Comparable Corpus of China English. *Studies in Language and Linguistics*, 32(2): 113-127.

Greenbaum, S. (1988). Syntactic Devices for Compression in English. In J. Klegraf and D. Nehls (Eds.). *Essays on the English Language and Applied Linguistics on the Occasion of Gerhard Nickel's 60th Birthday*. Heidelberg: Julius Groos Verlag, pp. 3-10.

Grishman, R., Macleod, C., and Meyers, A. (1994). *COMLEX Syntax: Building a Computational Lexicon*. Presented at Coling 1994, Kyoto.

Halliday, M.A.K. (1994). *An Introduction to Functional Grammar* (*Second Edition*). London: Edward Arnold.

Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Langacker, R. W. (1991). *Foundations of Cognitive Grammar*, *Vol*. *2*. Stanford: Stanford University Press.

Leech, G., Hundt, M., Mair, C., and Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Lees, R. B. (1960). *The Grammar of English Nominalizations*. The Hague: Mouton de Gruyter.

Levy, Roger and Andrew, Galen. (2006). *Tregex and Tsurgeon: Tools for querying and manipulating tree data structures*. The 5th International Conference on Language Resources and Evaluation (LREC-2006).

Macleod, C., Grishman, R., Meyers, A., Barrett, L. and Reeves, R. (1998). NOMLEX: A Lexicon of Nominalizations. In *Proceedings of Euralex 98*.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19 (2): 313-220.

Meyers, A. (2007). *Those Other NomBank Dictionaries – Manual for Dictionaries that Come with NomBank*. Technical report, New York University.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Rohde, D. L. T. (2005). *TGrep2 User Manual*. Version 1.15 edition.

Santorini, B. (1990). *Part-of-speech Tagging Guidelines for the Penn Treebank Projec*t. Technical report MS-CIC-90-47, Department of Computer and Information Science, University of Pennsylvania.

Santorini, B., and Marcinkiewicz, A. (1991). *Bracketing Guidelines for the Penn Treebank Project*. Unpublished manuscript, Department of Computer and Information Science, University of Pennsylvania.

Semino, Elena. (2009). Language in Newspaper. In J. Culpeper et al. (Eds.). *English Language: Description, Variation and Context*. Basingstoke: Palgrave Macmillan, pp. 439-453.

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013). Parsing with Compositional Vector Grammars. In *Proceedings of ACL*, pp. 455-465.

Tyrkkö, J. and Hiltunen, T. (2009). Frequency of Nominalization in Early Modern English Medical Writing. In A. H. Jucker, D. Schreier, and M. Hundt (Eds.) *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi, pp. 297-320.

Xu, Z. C. (2008). Analysis of Syntactic Features of Chinese English. *Asian Englishes*, 11(2): 4-31.

Xu, Z. C. (2010). *Chinese English: Features and Implications*. Hong Kong: Open University of Hong Kong Press.

Zhang, Y and Campbell, K. P. (1995). English in China. *World Englishes*, 14(3), 377-390.